

Scene and Graph Domain Adaptation to Tackle Domain Shifts in Real-world Image Semantic Segmentation and Human Activity Recognition

Haitao Tian

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Electrical Engineering and Computer Science
Faculty of Engineering, University of Ottawa

Abstract

Deep learning has transformed computer vision by automating feature extraction and representation learning across multiple data modalities, surpassing traditional paradigms that relied on manual feature engineering. However, deploying deep learning models in real-world applications often encounters significant challenges in generalization due to domain shifts between the training dataset (source domain) and the deployment environment (target domain). When the source domain distribution is misaligned or represents only a subset of the target domain, the deep learning model's performance tends to degrade in the target domain, limiting its scalability in many real-world applications.

The research in this thesis addresses the domain shifts challenge with the development of novel Domain Adaptation (DA) frameworks. Particularly, DA formulates adaptive models that can be trained on the source domain and adapted to the target domain where data is scarce. Moreover, unlike previous DA research that focuses on a single type of computer vision application, this thesis expands the research on individual DA frameworks for two different applications, that is, RGB image semantic segmentation and skeleton video human action analysis respectively, which can be detailed as follows.

- 1) The first proposed DA framework, named Enhanced Scene Domain Adaptation (**ESDA**), focuses on mitigating scene-based domain shifts for the task of semantic segmentation on street images. Street scene understanding is critical for autonomous driving and robotic navigation, but outdoor scenes often exhibit scene-based visual discrepancies caused by variations in lighting, weather, and regional characteristics. Therefore, **ESDA** introduces three innovative DA mechanisms: *pixel-wise adversarial adaptation*, *prototypical knowledge adaptation*, and *a target-specific*

adaptation classifier, each learning a model trained on image samples from one region (e.g., a city) to achieve robust performance on samples from other regions (e.g., alternative cities) for semantic segmentation.

- 2) The second proposed framework, named *Enhanced Graph Domain Adaptation (EGDA)*, addresses graph domain shifts in skeleton-video based human activity analysis. The latter has been widely utilized to address a variety of real-world video surveillance tasks, where human actions are represented by the trajectories of skeletal joints captured by acquisition systems. By proposing four novel methods, *cross-view adaptation*, *cross-sensor adaptation*, *cross-sequence adaptation*, and *cross-permutation adaptation*, **EGDA** effectively encourages the deep learning model to aggregate domain-invariant action dynamics from skeletal joint trajectories while dealing with the domain shift arising from varying camera perspectives, sensor configurations, and sequence lengths or activity ordering.

While evaluating the **ESDA** and **EGDA** frameworks, the research utilizes existing large-scale benchmarks to mimic various domain shift situations while creating pairs of a source domain and a target domain. For instance, in semantic segmentation, **ESDA** utilizes the dataset composed of synthetic street images as the source domain and evaluates the model on a real-world street image dataset. In human action analysis, **EGDA** leverages several large-scale skeleton datasets that are collected in different environments to mimic the domain shifts in skeleton data. Experimental results demonstrate that the proposed frameworks are effective for alleviating both scene and graph domain shifts from different data modalities and successfully improve the adaptability of the baseline deep learning models for image semantic segmentation and skeleton action analysis tasks.

The original contributions of this thesis include a comprehensive investigation of two types of domain shifts commonly encountered in computer vision: scene domain shifts in RGB image semantic segmentation and graph domain shifts in skeleton-based video human action recognition. Novel methodologies are introduced to enhance domain adaptation performance in these computer vision applications. The proposed frameworks enable adaptive deep learning models to be trained on large-scale benchmarks while easily adapting to real-world target domains where labeled data is scarce or unavailable. The original domain adaptation approaches are designed to enhance the base networks without additional computational complexity, ensuring efficient operation on the real-world target domain.

Acknowledgements

I am heartfelt grateful to all those who have supported and guided me along the way of my Ph.D. journey.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Pierre Payeur. The journey of completing my Ph.D. has been an incredible learning experience under his unwavering support and invaluable guidance. I am truly fortunate to conduct research on my own interests under Dr. Pierre Payeur's continuous encouragement. I also extend my sincere appreciation to Dr. Shiru Qu, for her constructive feedback and valuable suggestions on my research, which have significantly shaped my research and provided key insights that have enriched this thesis.

I am grateful to my colleagues, James McCulloch Dickens, Shengsong Yang, and Ghazal Rouhafzay, at the University of Ottawa, who have been a constant source of support, collaboration, and intellectual exchange. The stimulating research environment and countless discussions have enriched my learning experience and made this journey more enjoyable. A special thank to the team at Spectronix Inc. for their mentorship during my internship and for providing valuable resources and real-world perspectives that strengthened the impact of this research. I would like to acknowledge the financial support that I received from Mitacs which has enabled me to dedicate myself fully to my studies.

To my family and friends, your support and patience have been invaluable. You have been my greatest source of strength throughout this journey. I could not have reached this milestone without you.

Thank you all.

Table of Contents

Abstract	ii
Acknowledgements	v
List of Figures	ix
List of Tables.....	xiii
List of Acronyms.....	xvii
Chapter 1. Introduction	1
1.1 Motivation	1
1.2 Problem Statement and Objectives	4
1.3 Proposed frameworks and thesis structure	5
1.4 Original Contributions.....	7
Chapter 2. Background.....	11
2.1 Convolutional Neural Networks.....	11
2.2 Image Semantic Segmentation	12
2.2.1 Deep learning models	12
2.2.2 Semantic segmentation on street scenes	12
2.3 Human Activity Analysis.....	13
2.3.1 Human action recognition	13
2.3.2 Human action segmentation	14
2.4 Benchmarks, Datasets, and Acquisition Technologies	14
2.4.1 Image data for semantic segmentation on street scenes	15
2.4.2 Skeleton data in human action analysis	15
2.5 Domain Shift	16
2.5.1 Domain shift in street images	17
2.5.2 Domain shift in skeleton videos	18
2.6 Domain Adaptation	21
2.6.1 Data-based adaptation methods	22
2.6.2 Model-based adaptation methods	23
2.7 Summary	24
Chapter 3. Enhanced Scene Domain Adaptation for RGB Images	25
3.1 Pixel-wise adaptation	25
3.1.1 Summary.....	25

3.1.2 Original contributions	26
3.1.3 Article	26
3.2 Prototypical knowledge adaptation	43
3.2.1 Summary	43
3.2.2 Original contributions	43
3.2.3 Article	43
3.3 Target-dependent classifier adaptation	60
3.3.1 Summary	60
3.3.2 Original contributions	60
3.3.3 Article	61
3.4 Supplementary Information	77
3.4.1 Additional implementation details	77
3.4.2 Computational cost	77
3.4.3 Additional experiments	78
Chapter 4. Enhanced Graph Domain Adaptation for Skeleton Videos	81
4.1 Cross-View Adaptation	82
4.1.1 Summary	82
4.1.2 Original contributions	82
4.1.3 Article	83
4.2 Cross-Sensor Adaptation	98
4.2.1 Summary	98
4.2.2 Original contributions	98
4.2.3 Article	99
4.3 Cross-Sequence Adaptation	107
4.3.1 Summary	107
4.3.2 Original contributions	107
4.3.3 Article	108
4.4 Cross-Permutation Adaptation	118
4.4.1 Summary	118
4.4.2 Original contributions	118
4.4.3 Article	119
4.5 Supplementary Information	131

4.5.1 Additional implementation details.....	131
4.5.2 Computational cost.....	131
Chapter 5. Conclusion.....	134
5.1 Summary of Contributions.....	134
5.1.1 Contributions from ESDA.....	135
5.1.2 Contributions from EGDA.....	136
5.2 Discussion.....	137
5.2.1 Common characteristics of ESDA and EGDA.....	137
5.2.2 Limitations.....	139
5.3 Future work.....	140
Additional References.....	143

List of Figures

Figure 2.5.1. Illustration of scene domain shifts in street image semantic segmentation.	18
Figure 2.5.2. Illustration of graph domain shifts in human activity recognition. (a): domains shifts across camera configurations; (b): domain shifts across skeleton representations that are extracted by different skeleton acquisition sensors....	19
Figure 2.5.3. Conceptual illustration of domain shift.....	21
Figure 2.5.4. Conceptual illustration of domain adaptation.....	22
Figure 3.1.1. Illustration of proposed pixel-wise weighting adaptation.....	30
Figure 3.1.2. Illustration of proposed weighting maps.	34
Figure 3.1.3. A conceptual overview of the proposed weighted adaptation network.....	35
Figure 3.1.4. Qualitative results obtained for semantic segmentation without and with domain adaptation.....	37
Figure 3.1.5. Feature clusters visualization by t-SNE. Each color represents a different class cluster.	37
Figure 3.2.1. Illustration of prototypical knowledge oriented adaptation.....	46
Figure 3.2.2. Visualization of the prototypical knowledge adaptation (PKA) process.....	48
Figure 3.2.3. A conceptual overview of the proposed prototypical knowledge oriented adaptation (PKA) framework.....	50
Figure 3.2.4. Illustration of iterative updates of the proposed network.	50
Figure 3.2.5. Qualitative results obtained for semantic segmentation without and with domain adaptation.	54
Figure 3.2.6. Feature clusters visualization in the embedding space by t-SNE.	54

Figure 3.2.7. Performance (on mIoU) of PKA+ while altering the hyperparameter T.	54
Figure 3.2.8. Discrimination confidence visualization at different iteration steps (Unilateral discriminator (first row) vs. Conventional adversarial discriminator (second row)).....	55
Figure 3.2.9. PKA in real-world domain shifts.	56
Figure 3.3.1. Illustration of conventional domain adaptation and the proposed method.....	63
Figure 3.3.2. Cluster assumption violation in the target domain.....	64
Figure 3.3.3. Overview of the proposed two-phase strategy of learning a target-dependent classifier.	65
Figure 3.3.4. A conceptual overview of the PCEM fine-tuning phase.....	65
Figure 3.3.5. Illustration of meta-learning based target-dependent classifier learning strategy.	66
Figure 3.3.6. Left: unsupervised fine-tuning proposes a sequential learning process upon vanilla domain adaptation and conditional entropy minimization.....	67
Figure 3.3.7. Prediction results obtained for semantic segmentation on the task “GTA5 to Cityscapes” without and with domain adaptation.....	70
Figure 3.3.8. Input perturbations with regard to different types of noise on the input image samples of the target domain (Cityscapes).	71
Figure 3.3.9. Class clusters visualization using t-SNE for two classifiers under the baseline ADVENT on the embedding space of the L4 layer of ResNet101.....	74
Figure 3.4.1. Comparison of the computational costs of different implementations.....	77
Figure 4.1.1. Unsupervised domain adaptation (UDA) framework combining the proposed CPL and TAM strategies.....	87

Figure 4.1.2. Tensor flow of the training pipeline. Green symbols relate to the source domain, while red ones relate to the target domain.	89
Figure 4.1.3. T-SNE visualization on action prototypes on the feature space of ST-GCN	94
Figure 4.1.4. Performance (Top1 Acc) of TAM with varying hyperparameter $\mathcal{T}$	95
Figure 4.1.5. Performance (Top1 Acc) of TAM with varying hyperparameter λ	95
Figure 4.2.1. Skeletal joints detected by different generations of Microsoft Kinect.....	101
Figure 4.2.2. A conceptual overview of the proposed adaptation network for human action recognition.	102
Figure 4.2.3. Illustration of network architectures for the proposed adaptation framework.....	103
Figure 4.2.4. Joint remapping from Microsoft Kinect V2 to Microsoft Kinect V1.....	104
Figure 4.3.1. (a): Two video sequences from PKU MMD v2 exhibiting multiple actions that can be segmented into several class-specific actions.....	109
Figure 4.3.2. A concatenation of skeleton sequences resulting from stitching along the temporal axis.....	110
Figure 4.3.3. Conceptual overview of the proposed framework.	113
Figure 4.3.4. Effectiveness (Acc%) of Skeleton-Stitch.....	115
Figure 4.4.1. Trend of contrastive learning performance in skeleton-based action segmentation by publication year.	120
Figure 4.4.2. Workflow for learning action representations from trimmed skeleton sequences.....	121
Figure 4.4.3. Computation workflow of the proposed DuoCLR approach.....	123

Figure 4.4.4. Transfer learning in skeleton-based human action segmentation.....	125
Figure 4.4.5. Feature clusters visualization in the embedding space by t-SNE.	126
Figure 4.4.6. The confusion matrices obtained by ROR on different action granularities.....	127
Figure 4.5.1. Comparison of the computational costs for different components implemented as part of the proposed EGDA framework.....	132

List of Tables

Table 1.3.1. Table 1.3.1 Comparison of the proposed frameworks.....	5
Table 3.1.1. Experimental results of the adaptation task “GTA5 to Cityscapes” under the metric of mean IoU.....	39
Table 3.1.2. Experimental results of the adaptation task “SYNTHIA to Cityscapes” under the metric of mean IoU.....	40
Table 3.1.3. Ablation studies in the task “GTA5 to Cityscapes”.....	40
Table 3.2.1. Experimental results of the adaptation task “GTA5 to Cityscapes” under the metrics of IoU on each class and mean IoU on overall performance.....	53
Table 3.2.2. Experimental results of the adaptation task “SYNTHIA to Cityscapes” under the metrics of IoU on each class and mean IoU on overall performance.....	55
Table 3.2.3. Ablation studies. FDA refers to feature space based adversarial adaptation.....	55
Table 3.2.4. Experimental comparison between Entropy minimization loss ($H(x)$) and Maximum Squares loss ($MS(x)$).....	56
Table 3.2.5. Adaptation results over the <i>MSeg</i> dataset under the metrics mean IoU on each dataset and the harmonic mean over all datasets (the testing split).....	57
Table 3.2.6. Experimental results of the adaptation task “Cityscapes to Oxford RobotCar” under the metrics of IoU on each class and mean IoU on overall performance.....	57
Table 3.2.7. Experimental results of the adaptation task on video semantic segmentation.....	57
Table 3.3.1. Experimental results for the first “synthetic to real-city” adaptation task “GTA5 to Cityscapes” in terms of mean IoU.	68

Table 3.3.2. Experimental results for the second “synthetic to real-city” adaptation task “SYNTHIA to Cityscapes”	69
Table 3.3.3. Experimental results of the adaptation task “Cityscapes to Oxford RobotCar” under the metrics of IoU on each class and mean IoU on overall performance.	70
Table 3.3.4. Experimental ablation study results for the “GTA5 to Cityscapes” task.	71
Table 3.3.5. Tuning of the hyperparameter α	72
Table 3.3.6. Tuning of the hyperparameter β	72
Table 3.3.7. Comparison on the variants of the locally-Lipschitz constraint.....	73
Table 3.3.8. Tuning of the hyperparameter γ under the meta-learning (GTA5 to cityscapes) evaluation for ADVENT.....	73
Table 3.3.9. Comparison on the computational costs of different implementations of meta- learning.....	73
Table 3.4.1. Experimental results for the task “Cityscapes to Real Cities”.....	80
Table 4.1.1. Experimental results on unsupervised domain adaptation (test case 1).....	91
Table 4.1.2. Experimental results on unsupervised domain adaptation (test case 2).....	91
Table 4.1.3. Experimental results on unsupervised domain adaptation (test case 3).....	91
Table 4.1.4. Experimental results on unsupervised domain adaptation (test case 4).....	91
Table 4.1.5. Performance (Top 1 accuracy) for different proportions of the target data samples in classifier C refinement training in contrastive prototype learning.....	92
Table 4.1.6. Comparison to state-of-the-art works.	93

Table 4.1.7. Performance of “vanilla” fully supervised learning vs. semi-supervised learning. “NTU” stands for “NTU RGB+D”, and “PKU” stands for “PKU MMD Part I”.	93
Table 4.1.8. The effectiveness of using pseudo-labels with TAM.	94
Table 4.2.1. Comparative performance of the proposed method under the metric of Top-1 Accuracy.	104
Table 4.2.2. Experimental results of the adaptation task “NTU RGB+D to Northwestern UCLA”.	105
Table 4.2.3. Performance variation related to different proportions of the target data samples used for domain adaptation.....	105
Table 4.2.4. Experimental comparison of the discriminator implemented with different network structures.....	105
Table 4.2.5. Performance variation related to hyperparameter λ	105
Table 4.3.1. Experimental results with supervised temporal adaptation from “NTU RGB+D to PKUMMD 1”.	114
Table 4.3.2. Experimental results with supervised temporal adaptation from “NTU RGB+D to PKUMMD 2”.	114
Table 4.3.3. Experimental results with zero-shot adaptation.....	114
Table 4.3.4. Action segmentation results with temporal domain adaptation.	115
Table 4.4.1. Experimental comparison in the multi-class skeleton-based action segmentation task.....	125
Table 4.4.2. Experimental comparison in the composite skeleton-based action segmentation task.....	125

Table 4.4.3. Semi-supervised evaluation ($mAP_{0.1}$) with a low proportion of samples from the target domain.....	126
Table 4.4.4. The ratio of labeled data is saved while reaching the same performance with the vanilla model that was trained on 100% data.....	126
Table 4.4.5. Ablation study of two surrogate tasks: CPC (Cross Permutation Contrasting), ROR (Relative Order Reasoning), and combined DuoCLR.....	126
Table 4.4.6. Experimental results (Acc) of pre-training using untrimmed datasets.....	126
Table 4.4.7. Experimental comparison of DuoCLR and its variants using different skeleton data augmentation strategies	127
Table 4.4.8. Experimental comparison between two types of classifiers.....	127

List of Acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CPC	Cross Permutation Contrasting
CPL	Contrastive Prototypes Learning
CS	Cross-Subject
CV	Cross-View
DA	Domain Adaptation
DuoCLR	Dual-Surrogate Contrastive Learning
ESDA	Enhanced Scene Domain Adaptation
EGDA	Enhanced Graph Domain Adaptation
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
GCL	Granular Contrastive Learning
GCN	Graph Convolutional Network
IoU	Intersection over Union
MAML	Model-Agnostic Meta-Learning
mIoU	Mean Intersection over Union
MLDG	Meta-Learning Domain Generalization
MoCap	Motion Capture
MoCo	Momentum Contrastive

NN	Neural Network
PCEM	Pixel-wise Conditional Entropy Minimization
PKA	Prototypical Knowledge Adaptation
PKR	Prototypical Knowledge Refinement
TAM	Temporal Adaptation Modeling
TCN	Temporal Convolutional Network
TOR	Temporal Order Reasoning
T-SNE	t-Distributed Stochastic Neighbor Embedding
UDA	Unsupervised Domain Adaptation

Chapter 1. Introduction

1.1 Motivation

Computer vision has revolutionized various fields of engineering, enabling automatic vision systems to process, extract, and analyze meaningful information from images or videos in the real world. Early computer vision systems rely heavily on handcrafted feature engineering where features are manually designed using rule-based algorithms to interpret visual information [1–7]. However, these methods often struggle with complex real-world scenarios, which limit their generalizability and scalability in practice. Over the last decade, computer vision has transitioned from handcrafted feature engineering methods to deep learning approaches applying deeply structured Neural Networks (NNs) into end-to-end supervised optimization to learn models that can perform various tasks [8–12]. For instance, Convolutional Neural Networks (CNNs) [13] enable automatic feature extraction and representation learning for image classification [11]. Fully Convolutional Networks [14] are invented for end-to-end structure prediction tasks such as image semantic segmentation [8]. In human activity analysis [15], Graph Convolutional Networks (GCNs) [16] are leveraged to capture topological relations from abstract human body joints (skeleton data [17]), leading to computationally efficient models compared to traditional video based models [18], [19].

Despite significant advancements, the high dimensionality in the architecture of deep learning models spirals the number of network parameters, e.g., network weights, making deep learning a data-hungry paradigm that relies on large-scale training samples for effective training. To tackle the limitation, various benchmarks [20–25] have been

proposed. The latter offer two key benefits. First, benchmarks provide a supervised learning pipeline for training deep learning models, reducing the effort and cost for data collection and annotation. Second, benchmarks standardize model training processes, enabling consistent evaluations and facilitating the development of state-of-the-art models in deep learning.

However, the use of benchmarks introduces challenges, particularly the risk of overfitting. First, data samples from benchmarks often exhibit biases due to their controlled collection settings or localized environments, leading to a limited range of variations. Second, real-world data often exhibits significant variability, coming from diverse sources and dynamic environments. For example, an autonomous driving system [26] needs to operate in outdoor scenarios and must deal with large appearance variations in street scenes caused by differences in lighting, weather, and regional characteristics. As a result, deep learning models trained on benchmarks may perform well on the same benchmark but struggle to generalize to new benchmarks or real-world datasets not encountered during training.

Fine-tuning deep learning models with new data from the target environment is a common approach to eliminating the overfitting risk. But creating a representative dataset with accurate annotations for real-world scenarios that encompass a wide range of variability can be extremely time-consuming and expensive. For instance, capturing a representative dataset for street scenes requires accounting for variations in weather conditions, times of day, types of vehicles, pedestrians, buildings, and road types, making it both technically challenging and labor-intensive. Moreover, data labeling and annotation

add to the complexity and cost, potentially introducing human biases during the annotation process.

The phenomenon of data bias in deep learning has been extensively studied in research on **domain shift** [27], [28], [29]. In detail, **domain shift** refers to the mismatch on the marginal distributions, i.e., covariate shift, between **a source domain** (training data) and **a target domain** (deployment data). Domain shift occurs when a model trained on a bias source domain performs poorly when applied to a new, unseen target domain. Importantly, performance degradation due to domain shift cannot be resolved simply by adding more training samples from the source domain.

In computer vision applications, domain shift manifests in various forms:

- **Scene domain shift** arises from discrepancies in visual appearance, such as color, texture, and shape, caused by environmental and imaging inconsistencies. Scene domain shift occurs in Euclidean space and is commonly encountered in image data-based models, e.g., image classification [11], object detection [10], and semantic segmentation [9], which specializes in encoding scene semantic.
- **Graph domain shift** pertains to models [16], [30] that focus on graph semantic knowledge where learning instances in a scene are treated as nodes connected by edges representing interactions and correlations. Examples include interactions among traffic participants in street scenes, relationships between furniture in indoor scenes, or human-object interactions in video analysis. Unlike scene domain shift, graph domain shift relates to the topology space, interpreting relationships between graph nodes.

Human and deep learning show distinct capabilities in dealing with domain shift. Humans, through a lifetime of learning and experience, are capable of not only precisely

understanding and interpreting diverse visual scenes while recognizing objects, but also transferring the learnt visual knowledge and rich contextual understanding to an unseen scene by relying on their vision and cognition systems. However, for deep learning models, cross-domain knowledge transfer remains challenging. Recent research on explainable artificial intelligence (AI) [31] has demonstrated that deep learning lacks of the interpretability in learning human-understandable patterns from a large volume of training data, or even being entangled to data biases or peculiarities, which suggests the vulnerability of deep learning models dealing with domain shifts.

1.2 Problem Statement and Objectives

In many engineering applications, deep learning models are formulated from benchmarks or laboratory environments with the expectation of deployment into a target application environment without further model retraining or data collection. However, once the source domain is biased, with its dataset distribution lying in a subset of or being misaligned with the target domain, the model inference tends to degrade when transposed into the application due to domain shift, resulting in suboptimal performance.

This research proposes Domain Adaptation (DA) frameworks to improve on the generalization capability of deep learning models when deployed in real-world application environments that differ from their training conditions. DA formulates adaptive deep learning models that are designed to be trained from a controlled environment, e.g., a laboratory dataset, to deploy into an uncontrolled domain with minimum effort on data collection and annotation (i.e., the target domain is unlabeled or partially labeled) while maintaining performance comparable to that in the source domain.

1.3 Proposed frameworks and thesis structure

This thesis introduces two innovative DA frameworks respectively named Enhanced Scene Domain Adaptation (**ESDA**) and Enhanced Graph Domain Adaptation (**EGDA**). These DA frameworks are designed for two distinct computer vision tasks: image-based semantic segmentation and skeleton-based human activity recognition, respectively. The motivation stems from the observation that these two computer vision tasks have seen significant transpositions into diverse real-world applications, such as autonomous driving [26], [32], [33] and video surveillance [34], [35], but their development heavily depends on labeled data, and their adaptability continues to face considerable challenges due to respective domain shifts. As summarized in Table 1.3.1, **ESDA** focuses on RGB image data to address scene domain shifts, while **EGDA** operates on skeleton data to handle graph domain shifts. The commonality of the two frameworks lies in that they share the same implementation of feature space adaptation where both **ESDA** and **EGDA** learn domain-invariant feature representations that can adapt between the source and target domains while addressing domain shift challenges. Moreover, since **ESDA** and **EGDA** act as external

Name	Individual Characteristics				Common Characteristics		
	Task	Data Modality	Domain shifts	Backbone	Data availability	Goal	Implement.
ESDA	Street image semantic segmentation	RGB Images	Scene Shift (lighting, textures, and scene composition)	Fully Convolutional Networks (FCN)	Two domains (Source (labeled) + Target(labelled/unlabeled))	Learn Domain-Invariant Feature Representation	External Adaptation Modules
EGDA	Skeleton-based human activity recognition	Skeleton Videos	Graph Shift (camera viewpoints, sensor types, temporal variations, and ordering variations)	Graph Convolutional Networks (GCN)			

Table 1.3.1 Comparison of the proposed frameworks.

adaptation components to the base network during training, they do not introduce additional computational complexity during inference. Once trained, the models retain their learned adaptability and can be deployed efficiently without requiring extra adaptation steps at test time.

Enhanced Scene Domain Adaptation (ESDA). Street image semantic segmentation [36] aims to assign object category labels to each pixel in an input image, providing both high-level semantic information and low-level scene details about the street scene, which plays a pivotal role in outdoor computer vision tasks such as autonomous driving [26], [32], [33] and robotic navigation [23]. One major challenge of semantic segmentation in street scenes lies in that outdoor scenes are highly dynamic and diverse, the discrepancy on street appearance (i.e., scene domain shift) can be manifested due to a wide range of variability in regions, weathers, and seasons. The Fully Convolutional Network (FCN) model trained on a local region (e.g., a given city) will achieve limited inference results on new regions (e.g., alternative cities). To tackle the difficulty, **Chapter 3** focuses on **ESDA** frameworks for cross-domain semantic segmentation and is organized as follows: **Section 3.1** introduces *a pixel-wise weighted adaptation framework* that investigates effective interdomain discrepancy regularization for semantic segmentation. **Section 3.2** presents *a prototypical knowledge-oriented adaptation framework* that incorporates pixel-wise adaptation into a soft adversarial adaptation loss. This approach enables semantic segmentation models to handle target domain pixels more effectively and precisely. **Section 3.3** proposes a novel hypothesis for cross-domain semantic segmentation, *the target-dependent classifier*, which effectively adapts to target domain clusters even in cases of partial domain alignment failure.

Enhanced Graph Domain Adaptation (EGDA). 3D skeleton videos have been widely utilized in human action analysis tasks such as action recognition [15] and action segmentation [37], interactive gaming [38], video surveillance [35], etc. Recently proposed GCN models [39–41] are able to aggregate and propagate action dynamics from the sequences of skeleton joint trajectories in topological space. However, a GCN-based human action analysis system [12] could be sensitive to the variations in sensory settings, e.g., environmental configurations, camera setups, and sensing modalities, where models trained on local, controlled environment may struggle to generalize to open, uncontrolled environments, leading to reduced generalization ability. **EGDA** (in Chapter 4) addresses these challenges by leveraging graph visual knowledge in topological space. First, **Section 4.1** introduces a contrastive learning-based adaptation framework for *cross-view* skeleton-based human action recognition. Second, **Section 4.2** presents an adversarial adaptation framework designed for *cross-sensor* skeleton-based human action recognition, addressing the challenges posed by varying sensor configurations. Furthermore, **Section 4.3** formulates a novel *cross-sequence* adaptation method, enabling the learning of human action segmentation models from short, coarsely labeled sequences for target applications with limited unlabeled long sequences available for refinement. Last, **Section 4.4** formulates a novel *cross-permutation* framework that improves skeleton-based action segmentation via pre-training on trimmed skeleton videos.

1.4 Original Contributions

In a nutshell, the proposed **ESDA** framework introduces three original adaptation approaches in the scope of semantic segmentation, while **EGDA** introduces four innovative

approaches dedicated to human action recognition and segmentation. The related original contributions can be detailed as follows.

- 1) **Pixel-Wise Adversarial Adaptation in ESDA** introduces a novel inter-domain discrepancy regularization scheme for cross-domain semantic segmentation. **ESDA** is capable of interpreting and measuring pixel-wise domain shift information on network encoding spaces, to deploy weighted adaptation regularization that pays particular attention to the target domain. This work has been published in the paper “Unsupervised Pixel-wise Weighted Adversarial Domain Adaptation” (H. Tian, S. Qu, P. Payeur), published at the *16th International Symposium on Visual Computing (ISVC), 2021*.
- 2) **Prototypical Knowledge Adaptation in ESDA** integrates pixel-wise adaptation into a soft adversarial adaptation loss which allows semantic segmentation models to treat target domain pixels particularly and precisely. It leverages prototypical knowledge on the target domain to relax its hard domain label to a continuous domain space, which in turn enables elaborating specific adaptation strategies on respectively under-aligned regions and well-aligned regions of the target domain. The introduction of prototypical knowledge and soft adversarial adaptation loss are both new to the domain adaptation research community. This work has been published in the paper “A Prototypical Knowledge Oriented Adaptation Framework for Semantic Segmentation” (H. Tian, S. Qu, P. Payeur), *IEEE Transactions on Image Processing*, 2021.
- 3) **Target Domain Dependent Classifier in ESDA** is proposed as a novel hypothesis for cross-domain semantic segmentation. By observing the fact that traditional joint-domain adaptation could be biased to the source domain given a certain degree of failure on domain alignment, which violates the cluster assumption in the target domain, the

development of the target-dependent classifier enables effective adaptation to the target domain clusters. The model can readily be applied to existing adaptation models and leads to state-of-the-art adaptation performance. This work has been published in the paper “Learning a Target-Dependent Classifier for Cross-domain Semantic Segmentation: Fine-tuning versus Meta-learning” (H. Tian, S. Qu, and P. Payeur), *Pattern Recognition*, 2024.

- 4) **Cross-View Adaptation in EGDA** is proposed to approach the cross-view shift involved by extrinsic environment-specific characteristics. It presents a novel linkage of contrastive learning to domain adaptation in human action recognition that allows for learning critical skeletal knowledge from extrinsic environment variations, thereby leading to a view-robust model for practical deployment. In practice, the contrastive learning model learns invariant skeleton representations from the skeletal sequences collected in controlled laboratory conditions to formulate a deployable action recognition model to operate in uncontrolled settings. This work has been published in the paper “Unsupervised Temporal Adaptation in Skeleton-Based Human Action Recognition” (H. Tian and P. Payeur), *Algorithms* 2024.
- 5) **Cross-Sensor Adaptation in EGDA** presents the first research on learning an adaptive skeletal human action recognition model that effectively functions on heterogeneous skeleton graph configurations related to sensory particularities. A cross-sensor adaptation framework is proposed to learn critical skeleton correlations supported by heterogeneous skeletal graphs and generate transferable skeleton knowledge across sensory systems. This research has been published in the paper “Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition” (H. Tian, J. Dickens, P.

Payeur), *21st Conference on Robots and Vision (CRV)*, 2024.

- 6) **Cross-Sequence Adaptation in EGDA** presents an original formulation of short- to-long-sequence adaptation that allows the network to model action knowledge from short, coarsely labeled sequences for a target application domain where only a small number of unlabeled long sequences are available for model refinement. The research has been published in the paper “Stitch, Contrast, and Segment: Learning a Human Action Segmentation Model Using Trimmed Skeleton Videos” (H. Tian and P. Payeur), the *39th Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2025.
- 7) **Cross-Permutation Adaptation in EGDA**. This work aims at improving skeleton-based action segmentation via pre-training while action permutations are varying within a sequence of trimmed skeleton videos. By shuffling and concatenating segments of actions from different sequences, multi-action permutations provide varying contexts in which the model is encouraged to focus on invariant cross-permutation action features rather than being influenced by its specific temporal or contextual neighbors. The research is presented in the paper “DuoCLR: Dual-Surrogate Contrastive Learning in Skeleton Data Based Human Action Segmentation” (H. Tian, P. Payeur), which was accepted to the *International Conference on Computer Vision (ICCV 2025)* at the time this thesis is finalized.

Chapter 2. Background

This chapter reviews the background in deep learning, image semantic segmentation, human activity analysis, and related public datasets. It also covers the concepts of domain shift and domain adaptation, highlighting the gaps in existing methodologies. Given that this thesis is based on articles, the included publications in subsequent chapters feature their own specific literature review, which provide in-depth details tailored to each topic. The background review presented here aims to complement the literature reviews included in publications compiled in Chapter 3 and Chapter 4 and the related additional references are listed at the end of this thesis.

2.1 Convolutional Neural Networks

Deep learning has revolutionized several fields in computer vision, such as image classification [11], object detection [10], and semantic segmentation [9], with the concept of Convolutional Neural Networks (CNNs) [42], a deeply structured model evolved from LeNet-5 [13] which was initially designed for learning discriminative semantics for handwritten digit recognition. Over the last decade, the architecture of CNN has undergone continuous development with the introduction of AlexNet [43], VGGNet [44], and GoogLeNet [32], which progressively achieved state-of-the-art performance on a variety of computer vision tasks. More recently, ResNet [45] showcased an effective strategy on extremely deep CNNs, by introducing the concept of residual connections, making itself one of the most attractive network backbones downstream a variety of computer vision models. In general, the advancements in CNN have largely superseded traditional machine learning methods, offering more efficient, accurate, and scalable solutions for a wide range of real-world computer vision applications.

2.2 Image Semantic Segmentation

2.2.1 Deep learning models

Image semantic segmentation [9], a key step for image analysis in computer vision, aims to provide dense predictions over an input image by assigning each pixel a category label. As a result, it is capable of providing both high-level semantic information and low-level pixel position information. Compared to traditional methods for semantic segmentation that are generally developed upon hand-crafted features [46][7][4], deep learning-based semantic segmentation, such as Fully Convolutional Neural Networks (*FCNs*) [14], enable end-to-end training with supervised learning over pixel-wise labeled image samples. More recently, DeepLab [8] connects Conditional Random Fields with dilated convolutions while addressing the issues related to object boundaries and discerning context information, demonstrating an effective integration of the traditional method into *FCNs*. Furthermore, Pyramid Scene Parsing Network (PSPNet) [47] utilizes a pyramid pooling module to capture multi-scale contextual information.

2.2.2 Semantic segmentation on street scenes

Semantic segmentation of street scenes [7] aims to assign a ground truth label, such as road, trees, sky, etc., to the corresponding pixels on a street scene image, providing a detailed understanding of outdoor scenes for various applications, such as traffic scene understanding [26], robotic navigation [23], and autonomous driving [26], [32]. The difficulty of street scene segmentation lies in the high visual variability across street scenes. For instance, the appearance discrepancy between entities forming the traffic, street backgrounds, and traffic signs will be significant in different cities, countries, weathers, and terrains, leading to challenge the model's adaptability to different street scenes [36].

2.3 Human Activity Analysis

2.3.1 Human action recognition

Human action recognition [15] addresses a variety of human-action-related computer vision-based tasks, such as human activity analysis, human-computer interaction, and video surveillance, playing a key role in computer vision. Early approaches focus on the development of handcrafted features, such as Histogram of Oriented Gradients [48] and Histogram of Optical Flow [18], to represent local spatial and temporal patterns of human body motion from RGB sequences prerecorded on indoor scenes. However, these methods were limited in capturing complex temporal dynamics [15]. With the advent of 3D convolution operations [49], a RGB sequence is considered as an information volume of spatiotemporal data, allowing 3D convolutions to effectively capture both spatial and temporal features on action recognition.

Recent research on human action recognition is largely facilitated by the utilization of human skeletons, a compact spatial-temporal sequence extracted by processing raw data from 3D human sensing devices, such as depth sensors [50][51], point cloud sensors [52], [53], and inertial sensors [54], [55]. Skeleton representations offer the merit of being robust to variations in the environment and preserving privacy by preventing the leak of visual data on human subjects. Recent skeletal data-based human action recognition research relies on Graph Convolutional Networks [37] (*GCNs*), which excel at capturing human action dynamics within graph-structured data. The difficulty of skeletal action recognition lies in that the ambiguities in skeletal data representing human posture and motion can be more challenging to resolve without the additional visual cues that RGB data provides. Actional-Structural Graph Convolutional Networks (AS-GCN) [40] introduce directed graph neural networks to model the temporal dependencies among skeletal joints. Spatial-Temporal

Graph Convolutional Networks (ST-GCN) [41] employ spatial-temporal graph convolutional networks to capture both spatial and temporal information in skeletal data, achieving competitive results for action recognition. Channel-Wise Topology Refinement Graph Convolutional Networks (CTR-GCN) [39] propose a relations-modeling framework of GCN with the goal of interpreting relationships between joints.

2.3.2 Human action segmentation

Despite substantial progress, current human action recognition models are still limited for practical deployment. Specifically, GCN-based models are elaborated for sequence-wise classification such that the model input will be considered as a short sequence containing frames representing a unique action performed by a subject. However, in practice the input data could be of variable length or contain frames of a series of different actions. For practical adaptation, recent research focused on human action segmentation [37][56–62], which involves the process of frame-wise action recognition, providing both the action category information and timewise localization information (i.e., when specific actions start and end). However, powerful action segmentation models need to train over large-scale frame-by-frame annotated datasets, which is time-consuming, labor-intensive and prone to annotation biases.

2.4 Benchmarks, Datasets, and Acquisition Technologies

Compared to traditional computer vision tasks such as image classification, the development of *image semantic segmentation* and *human action analysis* is more reliant on large-scale benchmarks and datasets. This is because semantic segmentation requires pixel-wise labeled annotations, while human action analysis depends on the collection of extensive skeleton videos.

2.4.1 Image data for semantic segmentation on street scenes

For semantic segmentation on street scenes, a variety of datasets have been proposed, e.g., Cityscapes [21], KITTI [23], Mapillary Vistas [25], NTHU [63], and Oxford RobotCar [64], in which high-resolution images of streets are collected under the perspective of driving vehicles, capturing various urban scenes, weather conditions, and terrains. These datasets are also made available with pixel-level annotation for frequent object classes. Researchers are also interested in synthetic images as an appealing workaround to drive the deep learning model. For instance, GTA [65], Synscapes [66], and SYNTHIA [67], which are photo-realistic synthetic images rendered by graphic engines, are used to simulate street scenarios that cover a wide range of appearance diversity. Since synthetic images can be annotated programmatically, it eliminates the need for manual annotation, which is labor-intensive, expensive, and prone to human error.

2.4.2 Skeleton data in human action analysis

Various techniques have been employed to extract 3D skeleton representations from raw sensory data, each one presenting distinct acquisition pipelines to parse a skeleton model and identify joint distributions.

Microsoft Kinect. Microsoft Corporation has released three generations of Microsoft Kinect that are equipped with infrared sensors that realize markerless 3D human skeleton capture in real time [38], [68], [69]. Specifically, infrared sensors provide IR image and 3D depth data. Through the Microsoft SDK, the former is used to estimate human body segmentation over which to regress the initial joints position in 2D space for anatomy modeling, while the latter is used to map the 2D joint positions into a 3D skeleton output. The major discrepancy between different Kinect generations lies in the number of skeleton joints tracked. For instance, Kinect V1 provides 20 skeleton joints while Kinect V2 and

Azure Kinect support 25 and 32 skeleton joints tracking, respectively. Besides, the first two generations, Kinect V1 and Kinect V2, utilize traditional machine learning algorithms, e.g., decision forest, to detect body joints whereas Microsoft Azure Kinect improved detection robustness by utilizing deep learning models.

Existing 3D skeleton datasets are commonly recorded by using Kinect V1 (e.g., UT-Kinect [70], SBU [71], CAD 120 [72], and UCLA [73]) and Kinect V2 (e.g., PKU-MMD [74], NTU RGB+D [17], and NTU RGB+D 120 [75]).

Motion capture systems realize 3D skeleton acquisition by utilizing retroreflective markers, presenting a distinct acquisition pipeline compared to Microsoft Kinect devices. First, markers are placed on the subject's body which impose precise localization of key body joints. Second, triangulation on markers allows for the generation of 3D coordinates of human body joints around which 3D skeleton models are constructed. Two motion capture systems, Vicon [76] and OptiTrack [77] are widely used in skeleton data acquisition that relies on passive markers to transmit joint position information by reflecting light emitted by pre-calibrated cameras. Alternatively, some systems, such as OptiTrack Flex, use active markers which automatically emit infrared pulses, allowing for more precise reconstruction of skeleton models in the case of poor lighting conditions and occlusions.

2.5 Domain Shift

In many engineering applications, collecting vision data can be time-consuming and prohibitive. Leveraging publicly available data or laboratory datasets to train a model from a well-labeled source domain and then transfer to a label-scarce target domain is an attractive solution as it reduces dependency on labeled data while achieving target-adaptive models. However, when the source training data is biased or underrepresents the target

application domain, the resulting model risks overfitting to the source training data and could fail to generalize to the target domain.

Definition: Domain shift [28] refers to the difference in data distributions between the source domain (where a model is trained) and the target domain (where the model is applied), but the conditional distribution of the labels given the inputs remains unchanged. In practice, deep learning models trained on the source domain may perform well during evaluation on the same domain but struggle to generalize to unseen domains. Studies[78] confirm that supervised learning models can learn shortcuts or overfit to peculiarities in the data rather than capturing meaningful semantic patterns if the training data is noisy and contains outliers [78], [79].

2.5.1 Domain shift in street images

In street scene understanding, a computer vision system operating in outdoor scenarios may need to deal with large appearance gaps in street images where domain shift is formed due to the variations within a same semantic class between a source domain and a target domain (as illustrated in Figure 2.5.1). Common factors contributing to the domain shift include:

Appearance variation occurs when the visual information of objects or backgrounds changes across scenes. For instance, street scenes are highly dynamic and diverse. The visual texture of a same object, such as streets, buildings, vehicles, traffic signs, and other objects, can be relatively consistent in local regions, but tend to vary in different cities, regions, or terrains.

Lighting conditions may impact the process of the reflection and diffusion of light onto an object surface, thereby resulting in discrepancies during imaging. For instance, the time



Figure 2.5.1. Illustration of scene domain shifts in street image semantic segmentation.

of day (daytime or nighttime), weather (rain, snow, or fog), and seasons, can significantly affect the illumination on street objects. Models trained over data collected under consistent illumination and good lighting conditions may not perform as well in poor lighting conditions, and vice versa.

Synthetic images generated using advanced graphic engines are increasingly used to augment datasets for supervised network training [65]. Nevertheless, a deep learning model pre-trained on synthetic datasets will generally perform poorly on real-world datasets. Compared to real-world images, synthetic images rendered by computer graphics tend to preserve less details on textures, backgrounds, and contextual information.

2.5.2 Domain shift in skeleton videos

Skeletal data encodes the dynamics of human movements by depicting the trajectories of the skeletal joints over time. Skeleton data is often collected in indoor scenes using 3D human imaging sensors which generate an abstract data format compared to raw RGB data. However, variations in data collection environments can lead to domain shifts, causing

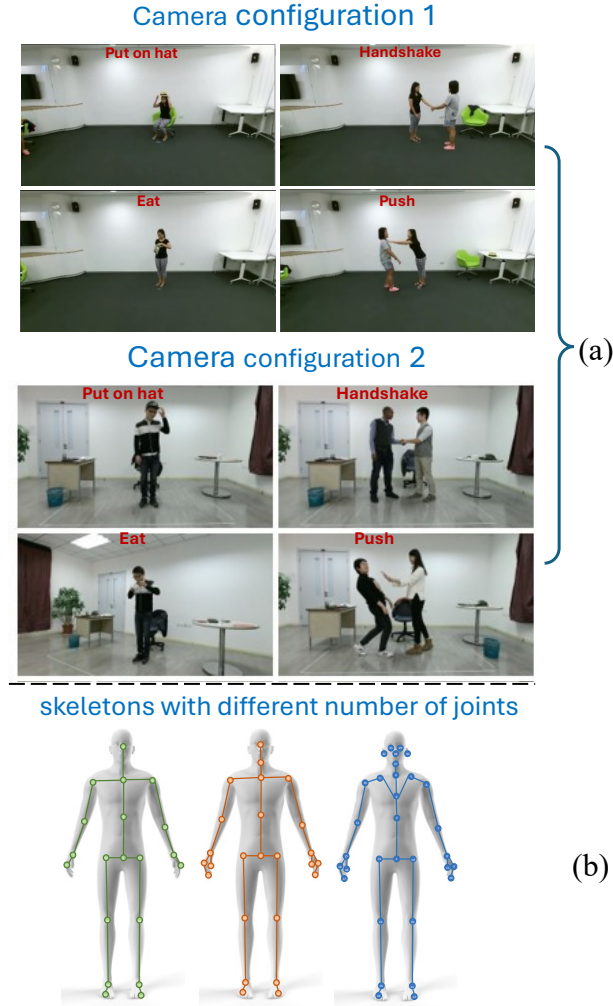


Figure 2.5.2. Illustration of graph domain shifts in human activity recognition. (a): domains shifts across camera configurations; (b): domain shifts across skeleton representations that are extracted by different skeleton acquisition algorithms.

models trained in controlled environments to struggle in uncontrolled, real-world settings. Key sources of domain shift in skeleton videos are illustrated in Figure 2.5.2, which include:

Configuration shift. Unlike street scenes, which typically exhibit well-structured and consistent layouts, human action recognition scenarios are often less controlled and more variable. Although 3D skeletons are represented by spatial coordinates, making them

inherently more robust to changes in imaging viewpoints, the process of extracting skeleton representation can still introduce significant variability. For example, when a body part is occluded or partially visible from a specific view (e.g., side view), the estimated joint positions can be highly unstable or inaccurate. This leads to inconsistent estimation while using the same sensor under a different view (e.g., front view). This contributes to configuration shift, where the representation of the same action can differ due to variations in sensor configurations scene[17][73].

Skeleton encoding shift. Skeleton representations vary upon different skeleton acquisition techniques, such as skeleton resolution (i.e., number of joints) and skeleton distribution (i.e., joint positions). For instance, the Microsoft Azure Kinect software [68] depicts 32 skeleton joints for the entire human body versus Microsoft Kinect V2 [69] which supports only 25 joints. Such a discrepancy on resolution complicates the transferability of models between systems. More critically, since skeletal data is not directly captured but rather extracted from raw depth or RGB-D data using pose estimation algorithms, the extraction process can introduce significant noise due to several factors such as sensor types, sensing modality, and algorithms used for skeleton extraction, which can degrade the quality of joint localization. Therefore, the skeleton representation of the same action can differ due to the use of different skeleton acquisition algorithms.

Sequence length shift. While coming from a real-time action acquisition system, skeleton data sequences can be generated with different durations, covering mixed action segments [62]. For example, greeting a person would involve a sequence of individual actions from different label categories (e.g., “standing”, “shaking hands”, “talking”, and “sitting”). In this context, the temporal length shift is reflected by the discrepancy on the

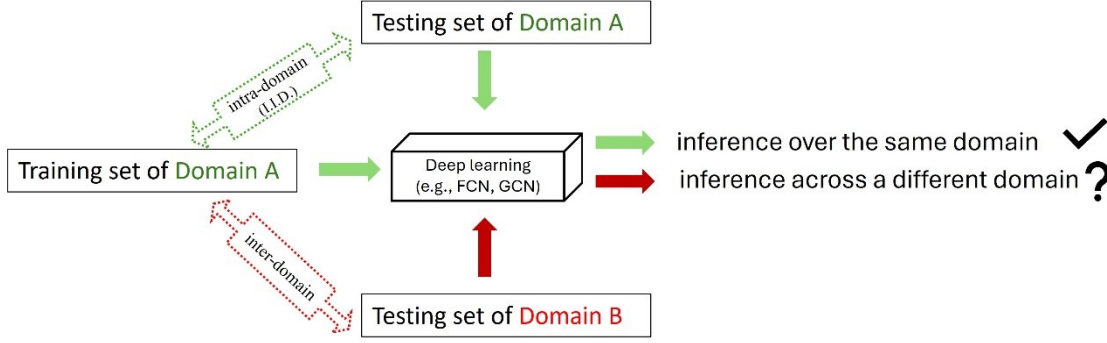


Figure 2.5.3 Conceptual illustration of domain shift.

duration of skeletal sequences where a source data domain contains short well-trimmed and uniquely labeled action, and a target data domain is composed of longer untrimmed skeleton sequences concatenating mixed action segments. Although such a type of domain shift has been investigated in RGB action video analysis [80], where pretraining paradigms (e.g., I3D [81]) are used for pre-learning on short RGB videos [82] before training on untrimmed videos, the problem remains unexplored on skeleton videos.

2.6 Domain Adaptation

Successful implementations of supervised learning with deep learning models generally involve that the training and testing data exhibit an independently identical distribution (I.I.D.), which guarantees low testing errors for a model that achieved low training errors on the training data (as illustrated in Figure 2.5.3). In this work, a more challenging scenario is considered, that is the training (source domain) and the testing data (target domain) are not I.I.D., given the presence of a domain shift. In this context, the capability of a model to reach low testing error on the target domain becomes crucial, and such a capability is commonly referred to as *generalization* in the deep learning community.

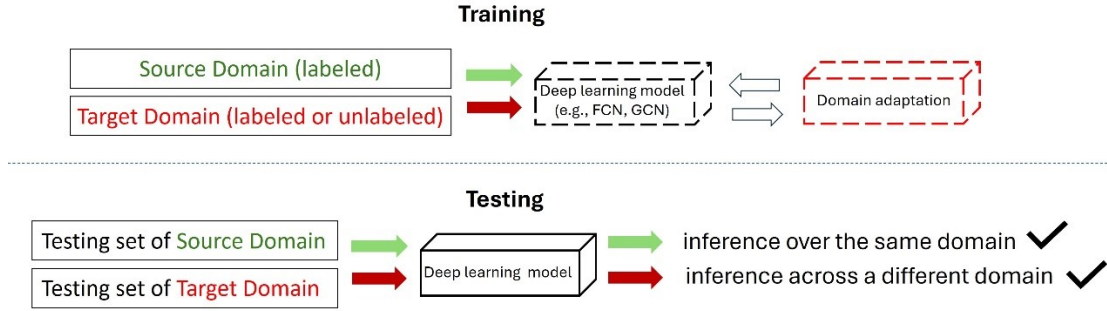


Figure 2.5.4. Conceptual illustration of domain adaptation.

Definition: Domain adaptation refers to a machine learning framework that adapts a computer vision model trained on one data distribution (the source domain) to perform well on a different but related data distribution (the target domain), for which labeled data is limited or unavailable (as conceptually illustrated in Figure 2.5.4). Existing paradigms have proposed model adaptation methods to improve the model’s generalization ability, which can be regrouped into two main categories: data-based (supervised) adaptation methods and model-based (unsupervised) adaptation methods.

2.6.1 Data-based adaptation methods

Model fine-tuning. An intuitive solution to eliminate the domain shift is to fine-tune the source-trained model upon a sufficient number of data collected in the target domain until the convergence at a low generalization error condition is met on the test dataset of the target domain. However, such a supervised learning paradigm imposes the collection and annotation of a large volume of data samples, which is extremely time-consuming in practice.

Data Augmentation. Effective data generation or augmentation techniques help enrich the source data distribution by simulating various scenarios and covering possible data

variations in the target domain. For instance, image-to-image translation [83] can be employed to capture the visual diversity of different domains to aid in training models that generalize well across varied environments. In human action recognition, configuring multiple camera setups and sensory configurations is a common practice to simulate the viewpoint diversity in the real world, e.g., NTU RGB+D 120 [75] involves data variations by configuring 3 camera angles, 16 variations in height and distance, as well as involving 40 actors into actions performance.

Multi-Domain Datasets. Another approach [84–86] includes the integration of data from diverse sources into a composite training dataset to yield a more robust model in comparison to training on individual datasets. However, label space clashes and inconsistent annotation practices across datasets substantially reduce the accuracy of models trained on multiple datasets.

2.6.2 Model-based adaptation methods

Generative adversarial learning. Generative Adversarial Networks (*GANs*) [87], [88] have been widely used for representation learning, with the pivotal competency for eliminating domain discrepancy among distribution encodings and learning domain-agnostic models. Domain-Adversarial Neural Networks (DANN) [89] introduce a domain-adversarial training procedure, which aligns feature representations across domains by using a domain classifier to enforce domain invariance.

Contrastive learning extracts data representations by contrasting positive pairs against negative pairs from a pre-defined dynamic dictionary [90], where it clusters samples from positive pairs while discriminating the samples from negative pairs. For instance, in skeletal action recognition, the contrastive learning scheme regards each skeletal sequence as a

unique class represented by a *GCN* encoder and exploits the skeletal invariances and similarities from the dynamic dictionary formulated by skeleton sequences without using data annotations [91–95].

Meta-learning is well known as the competency of “learning to learn” on reliable network initialization [96]. For instance, MAML [97] and Reptile [98] are devoted to learning good weights initialization for few-shot learning network that can lead the network to adapt to new tasks in a few gradient steps. MLDG [99] then extends this strategy for adapting an image classification model from multi-source domains to an unseen domain, referred to as domain generalization. It develops a meta-learning procedure to put regularization on gradients optimization such that the network is optimized in a consistent direction amongst source domains. Li and Hospedales [100] propose an online meta-learning strategy to combine naïve meta-learning algorithms in several general domain adaptation problems. Guo et al. [101] propose a domain-aware meta-learning strategy for domain adaptation to sample reliable source pixels to facilitate self-supervised adaptation progress with pseudo labels.

2.7 Summary

This chapter summarized key concepts and background that support the development of various methodologies for domain adaptation to be detailed through a series of publications presented in the following chapters. More extensive details about the state-of-the-art for elements of the methodology are included in the respective publications.

Chapter 3. Enhanced Scene Domain Adaptation for RGB Images

This chapter introduces three original formulations for the proposed Enhanced Scene Domain Adaptation (**ESDA**). It details the methodology and results of the three different frameworks. All three approaches were already published in peer-reviewed venues, as follows:

Section 3.1 on **Pixel-wise adaptation** is a conference paper published at *ISVC 2021* [102]; Section 3.2 on **Prototypical knowledge adaptation** is a journal paper published in *IEEE Transactions on Image Processing* in 2021 [103]; and Sections 3.3 on **Target domain dependent classifier** is a journal paper published in *Elsevier Pattern Recognition* in 2023 [104]. For the sake of clarity, the contents of each paper are briefly summarized in the following sections, emphasizing the corresponding original contributions. The publications are then presented as published. Finally, Section 3.4 presents additional experimental results that are not covered in the papers.

3.1 Pixel-wise adaptation

3.1.1 Summary

ESDA first proposes a pixel-wise weighted DA scheme for cross-domain semantic segmentation. It draws on the fundamental observation that objects in street scenes exhibit different degrees of visual appearance discrepancy across different regions, e.g., vegetation on the sides of street may be more visually consistent than dynamic objects, such as cars and pedestrians. The pixel-wise weighted scheme builds on the intuition for semantic segmentation from pixel-wise classification problems where **ESDA** is allowed to interpret and measure pixel-wise domain shift information on network encoding spaces.

In implementation, it first aligns the cross-domain distributions by interpreting pixel-wise inter-domain discrepancy amongst the target domain pixels to deploy weighted adaptation regularization that allows domain adaptation to pay more attention to deeply shifted regions than to lightly shifted ones. As such, each pixel contributes differently and separately to the adaptation loss. Second, given the uncertainty associated with the generation confidence on discrepancy indication, it introduces discrimination confidence from an auxiliary discriminator to refine the generation confidence with the goal of achieving further improvements. The formulation of generation and discrimination confidence does not introduce additional computations over the fundamental DA framework and effectively regularizes the output-space adaptation process at a finer level of detail.

3.1.2 *Original contributions*

The original contributions of **ESDA** in this section are:

- A pixel-wise weighted domain adaptation approach, which leverages a synthetic-data-trained segmentation network more effectively when applied on a real-world dataset;
- An auxiliary discriminator on the feature space, which is trained to measure the inter-domain discrepancy in a different perspective from what is found in the literature.

3.1.3 *Article*

This research has been reported in the paper “Unsupervised Pixel-wise Weighted Adversarial Domain Adaptation” (H. Tian, S. Qu, P. Payeur), published at the *16th International Symposium on Visual Computing (ISVC 2021)*, LNCS vol. 13017, 2021 [102]. This paper was awarded the “*Springer Best Paper Award*” at the conference. The online

version can be found at https://link.springer.com/chapter/10.1007/978-3-030-90439-5_46,
and the original paper is provided below as published. Permission to reprint the paper in this
thesis was obtained from the publisher Springer Nature Switzerland AG.



Unsupervised Pixel-Wise Weighted Adversarial Domain Adaptation

Haitao Tian^{1,2(✉)}, Shiru Qu¹, and Pierre Payeur²

¹ Northwestern Polytechnical University, Xi'an, China
htian026@uottawa.ca

² University of Ottawa, Ottawa, Canada

Abstract. Fully convolutional networks have been leveraged extensively in semantic segmentation tasks. While possessing demonstrated competency for dense prediction, such supervised learning networks are restricted to labeled data during training and hence show poor generalization while confronting unseen domains. As a pivotal transfer learning technique, domain adaptation aims to alleviate discrepancies between distinct domain distributions to improve the performance of generalization in unsupervised manners. Although a family of domain adaptation methods have demonstrated significant effectiveness on cross-domain semantic segmentation tasks, the overlook of pixel-wise domain divergences leads to over-adaptation. To deal with this problem, we investigate effective pixel-wise inter-domain discrepancy metrics to regularize the training of adaptation networks at a pixel-wise level. We first leverage generation confidence encoded from the output space as a weighting map to impose more adaptation emphasis on deeply shifted regions. Furthermore, we employ discrimination confidence on the feature space to refine generation confidence into a more reliable weighting map. The formulation of generation and discrimination confidence does not introduce additional computations over the fundamental DA framework. In our experiments, the proposed pixel-wise weighted adaptation approach outperforms state-of-the-art methods on two cross-domain segmentation tasks and demonstrates effective alleviation of over-adaptation.

Keywords: Domain adaptation · Semantic segmentation · Pixel-wise adaptation · Generation confidence · Discrimination confidence

1 Introduction

Even though fully convolutional networks (FCNs) recently dominated the field of semantic segmentation [1–3], such networks are generally trained using a huge number of pixel-wise labeled data, which precludes its application in practical scenarios, such as autonomous driving and robotic navigation, where collecting labeled data in changing scenarios with large appearance gaps is extremely expensive. Recently developed photo-realistic synthetic street-scene datasets [4] offer an appealing workaround by simulating various scenarios for supervised network training. However, an FCN pre-trained on synthetic datasets will generally fail on real-world inference, which is referred to as

the cross-domain shift [5, 6]. This domain shift cannot be eliminated even with a large number of synthetic data.

The domain adaptation (DA) technique is recognized for its capability of transferring learnt semantic patterns from a source domain to a target domain without using labeled data from that target domain. Therefore, DA has been well leveraged into cross-domain semantic segmentation tasks between synthetic datasets (source domain) and real-world datasets (target domain). Early research [7–9] developed a discrimination network (discriminator) on the feature space to align domain distributions. However, feature-level adaptation is limited when decoding high dimensional visual cues that have been suppressed in the feature space [10]. To overcome this issue, alternative approaches [10–12] turned to utilize structural information on the output space and hence obtained promising domain adaptation performance.

Despite such advances, pixel-wise divergence on inter-domain discrepancy is generally overlooked in conventional DA works. For instance, as shown in Fig. 1, while clear domain shift exists in between two domains, the source-trained network still preserves a capability to correctly classify parts of the pixels that are originally exhibiting light domain discrepancy in the target domain. However, conventional domain adaptation methods would have negative effectiveness on such pixels while treating all pixels under a same degree of domain shift, thereby segmenting parts of lightly shifted pixels incorrectly.

In this paper, we propose a pixel-wise weighted DA scheme for semantic segmentation. We align the cross-domain distributions by interpreting pixel-wise inter-domain discrepancy amongst the target domain pixels, so as to deploy weighted adaptation regularization that allows domain adaptation to pay more attention to deeply shifted regions than to lightly shifted ones. As such, each pixel contributes differently and separately to the adaptation loss. Moreover, given the uncertainty associated with generation confidence on discrepancy indication, we introduce discrimination confidence from an auxiliary discriminator to refine the generation confidence with the goal of further improvements.

The main contributions of this work are: 1) a pixel-wise weighted domain adaptation approach, which leverages a synthetic-data-trained segmentation network more effectively when applied on a real-world dataset; and 2) an auxiliary discriminator on the feature space, which is trained to measure the inter-domain discrepancy in a different perspective.

2 Related Work

2.1 Semantic Segmentation

Fully convolutional network [3] based models have outperformed traditional non-CNN models [13, 14] in the field of semantic segmentation. For instance, FCN-based Deeplab v2 [1] and PSPNet [2] show state-of-the-art performance on semantic segmentation tasks. Meanwhile, the development of pixel-wise labeled datasets (e.g., PASCAL VOC 2012 [15] and Cityscapes [16]) made FCNs usage possible in supervised training setting. But when confronting cross-domain segmentation tasks, the domain shift across

datasets degrades the performance of many state-of-the-art FCNs. Expanding the volume of datasets is the first attempt to deal with this problem. But collecting datasets with widespread variability that cover various testing scenarios is extremely time-consuming in the real world. As such, provided that the domain shift problem can be properly handled, FCN-based segmentation networks can alternatively be trained on synthetic datasets [4, 17], which are easily generated by graphics engines, and transposed onto real-world datasets and scenarios.

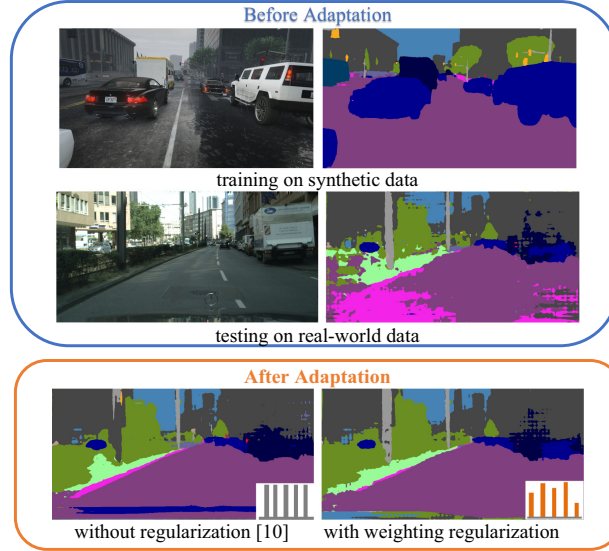


Fig. 1. Illustration of proposed pixel-wise weighting adaptation. **Top 2 rows:** samples from source and target domain (lft), a source-domain pretrained model for semantic segmentation inferred on these samples without adaptation (right). **Bottom row:** prediction results on target domain are improved by output-space domain adaptation method [10], while visible over-adaptation results (negative segmentations) exist (left). The proposed pixel-wise weighting adaptation method obtains comparable improvement while it alleviates over-adaptation (right).

2.2 Adversarial Domain Adaptation

Adversarial domain adaptation [6] has been successfully leveraged in cross-domain image classification [18, 19] and object detection tasks [20, 21]. When applied to semantic segmentation [7–12, 22–25], it embeds a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. By imposing the DA operation, the segmentation network not only learns discriminative representations but also invariant encodings from different domains. Specifically, the DA model consists of a generator and a discriminator. During training, as a generator, the segmentation network interacts with the discriminator in an adversarial manner [26] by which the discriminator is trained to upper bound the source and target domains distributions. The generator then minimizes this bound to eliminate the discrepancy between two distinct distributions. With this adversarial training process, the generator manages to produce

invariant-discriminative features, such that the segmentation network can be applied on both domains.

3 Preliminaries

In this section, we introduce mathematical preliminaries on supervised segmentation model settings and unsupervised domain adaptation settings, upon which we introduce the pixel-wise weighted adaptation approach in the next section.

The goal of domain adaptation for semantic segmentation is to train an FCN-based segmentation network on the source domain and transpose it to the target domain without supervision. To this end, we set a source domain $\mathcal{D}_S = \{(X_s^{[i]}, Y_s^{[i]})\}_i^{n_S}$, where $X_s^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is the i th of n_S synthetic dataset images with a size of $W \times H \times 3$ in $\mathcal{D}_S$, and $Y_s^{[i]} \in \mathbb{R}^{W \times H \times L}$ is a corresponding pixel-wise annotation label with L categories. We also set a target domain in the same way, i.e., $\mathcal{D}_T = \{(X_t^{[i]})\}_i^{n_T}$, where $X_t^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is a real-world dataset image. Note that there are no available labels in $\mathcal{D}_T$. For clarity, (X_s, Y_s) and X_t correspond to a random sample from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively in the rest of this paper.

3.1 Supervised Semantic Segmentation

For training a FCN-based segmentation network, a source domain sample (X_s, Y_s) is fed into a feature encoder F , then a dense classifier C takes high-dimensional feature encodings from F and produces a final soft-max probability distribution $P_s = C(F(X_s))$, where the structural prediction likelihood is $P_s^* = \max_l(P_s)$. The objective loss function is formed by the multi-class cross entropy with P_s and Y_s , formulated as:

$$\mathcal{L}_{\text{seg}}(F, C) = -\mathbb{E}_{(X_s, Y_s) \in \mathcal{D}_S} \left[\sum_{(w, h)} \sum_l I_{[l=y_s^{(w, h)}]} \log C(F(x_s^{(w, h)}))^{(l)} \right] \quad (1)$$

where $\mathbb{E}[\cdot]$ is the statistical expectation and ground truth label Y_s is correspondingly encoded into one-hot vector.

As for a target sample, $X_t \in \mathcal{D}_T$, the source-trained segmentation network under the parameter distribution of $\mathcal{D}_S$ can also generate a direct probability distribution P_t , while it would hardly reflect the genuine label distribution of $\mathcal{D}_T$ because of the presence of domain shift. Meanwhile the absence of pixel annotation in $\mathcal{D}_T$ does not allow for fine-tuning on the trained segmentation network.

3.2 Unsupervised Output-Space Based Domain Adaptation

Different from feature-space based DA models that embed a discriminative network on the feature space, the output-space adaptation scheme considers the entire FCN-based segmentation model as a generator ($G = C[F(\cdot)]$), and embeds a discriminator D at the end of G in order to take advantage of structural information available on the output space.

During training, the G and D are updated alternatively by adversarial optimization on binary cross entropy loss (2). In detail, the discriminator D is firstly trained to classify

each element on the soft-max outputs $\mathbf{P}_s$ and $\mathbf{P}_t$ into its original domain label (we denote source domain label as 0, and target domain as 1), aiming to best represent inter-domain distribution discrepancy. $\mathbf{D}$ is updated with back propagation by minimizing (2), formulated as:

$$\begin{aligned} \mathcal{L}_{adv}(\mathbf{G}, \mathbf{D}) = & -\mathbb{E}_{\mathbf{X}_s \in \mathcal{D}_S} [\sum_{(w,h)} \log(1 - \mathbf{D}(\mathbf{G}(x_s^{(w,h)})))] \\ & -\mathbb{E}_{\mathbf{X}_t \in \mathcal{D}_T} [\sum_{(w,h)} \log(\mathbf{D}(\mathbf{G}(x_t^{(w,h)})))] \end{aligned} \quad (2)$$

During the discriminative network training, $\mathbf{G}$ is fixed for just taking part in the forward propagation.

Second, with the supervision from $\mathbf{D}$, the generator $\mathbf{G}$ is trained to produce “source-style” soft-max output $\mathbf{P}_t$ to eliminate the inter-domain discrepancy. $\mathbf{G}$ is trained by maximizing (2), while $\mathbf{D}$ is only allowed for forward propagation in this stage.

Combining with supervised training using (1), the segmentation network $\mathbf{G}$ manages to generate discriminative and invariant distributions among pixels within the source and target domains.

4 Pixel-Wise Weighted Adaptation Method

In this section, building upon the conventional output-space DA model, the proposed pixel-wise weighted adaptation approach is introduced in two stages. It first investigates the use of the generation confidence to develop a pixel-wise weighting method for output-space based adaptation. Second, it employs an auxiliary discriminator on the feature space to measure the inter-domain discrepancy from a different perspective, which contributes to further improvements.

4.1 Generation Confidence Weighted Adaptation

In Tsai *et al.* [10], although the pixel-wise adaptation process is carried out by an output-space based DA network, each element on the output space is considered under a same inter-domain discrepancy. In other words, such an approach implements a global adaptation for each target domain sample without any consideration on the pixel-wise inter-domain discrepancy. To deal with the problem, the proposed pixel-wise weighted adaptation focuses on emphasizing precise regularization on the adaptation process by using pixel-wise inter-domain discrepancy.

From the perspective of self-supervised adaptation [27, 28], the structural prediction likelihood $\mathbf{P}_s^*$ for a source domain sample shows the segmentation confidence of an FCN when classifying pixels into correct categories. It further infers that, for a target domain sample, $\mathbf{P}_t^*$ indicates the generation confidence in the context of DA, which tells the confidence of a FCN generating distributions on the target domain that are consistent with the source domain. In other words, a large value in $\mathbf{P}_t^*$ corresponds to small inter-domain discrepancy as the generator can segment the target-domain pixel with high generation confidence. As such, pixel-wise inter-domain discrepancy could effectively be estimated by the generation confidence.

Based on the above observation, we introduce a pixel-wise weighted adversarial adaptation framework that minimizes distribution distance across domains on the output space according to the pixel-wise inter-domain discrepancy. We consider the pixel-wise inter-domain discrepancy as a weighting map to drive the adaptation model to pay more attention on deeply shifted pixels. To further protect well aligned regions from over-adaptation, we set the pixel-wise discrepancy to a small value, here empirically set at 0.01, on well aligned regions (where the generation confidence is higher than a threshold, T). As such, the generation-confidence weighting map $d^{(w,h)}$ is formulated as:

$$d^{(w,h)} = \begin{cases} 0.01 & , \text{ if } p_t^{(w,h)*} > T, \\ e^{-p_t^{(w,h)*}} & , \text{ otherwise.} \end{cases} \quad (3)$$

The generation-confidence weighting map (shown on the upper-left of Fig. 2) allows domain adaptation to pay more attention to deeply shifted regions than to lightly shifted ones. Specifically, the weighting map $d^{(w,h)}$ is utilized to pixel-wise weight conventional adversarial adaptation loss (2). As such, each pixel contributes differently and separately to the adaptation loss. Following [11], we also employ a small adaptive weight (see details in Section V.B) to stabilize the adversarial training process. Hence, the proposed pixel-wise weighted adaptation approach can be formulated as:

$$\begin{aligned} \mathcal{L}_{advI}(\mathbf{G}, \mathbf{D}) = & -\mathbb{E}_{X_s \in \mathcal{D}_S} \left[\sum_{(w,h)} \log(1 - \mathbf{D}(\mathbf{G}(x_s^{(w,h)}))) \right] \\ & -\mathbb{E}_{X_t \in \mathcal{D}_T} \left[\sum_{(w,h)} (d^{(w,h)} + \mu) \cdot \log(\mathbf{D}(\mathbf{G}(x_t^{(w,h)}))) \right] \end{aligned} \quad (4)$$

4.2 Discrimination Confidence Weighted Aadaptation

In this section, we introduce discrimination confidence to integrate with generation confidence for a more reliable adaptation weighting map. It is based on the observation that target domain pixels that are far away from source domain decision boundaries would be hardly interpreted by generation confidence due to the unreliable value of $p_t^{(w,h)*}$ generated by the source FCN on the output space. In this way, a part of the discrepancy-agnostic pixels would be wrongly weighted according to the second condition of (3). This factor influences the generation confidence in the adaptation procedure, which eventually is prone to over-adaptation on the target domain.

To handle the above problem, it is proposed to rather use an auxiliary discriminator $\mathbf{D}^{\text{aux}}$ on the feature space to estimate domain divergence from a different perspective. As such, a complementary regularization is introduced that aims to further refine the generation-confidence weighting map. Specifically, we first introduce a discriminator on the feature space of the FCN for distinguishing the difference grid of the feature encodings between source and target domain images. Second, we involve the discriminator $\mathbf{D}^{\text{aux}}$ into an auxiliary adversarial training with feature encoder $\mathbf{F}$ using the binary cross entropy loss, as formulated in (5).

$$\begin{aligned} \mathcal{L}_{\text{aux}}(\mathbf{F}, \mathbf{D}^{\text{aux}}) = & -\mathbb{E}_{X_s \in \mathcal{D}_S} \left[\sum_{(w,h)} \log(1 - \mathbf{D}^{\text{aux}}(\mathbf{F}(x_s^{(w,h)}))) \right] \\ & -\mathbb{E}_{X_t \in \mathcal{D}_T} \left[\sum_{(w,h)} \log(\mathbf{D}^{\text{aux}}(\mathbf{F}(x_t^{(w,h)}))) \right] \end{aligned} \quad (5)$$

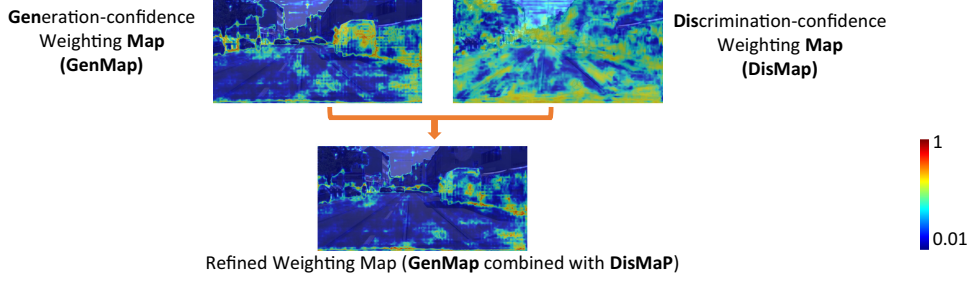


Fig. 2. Illustration of proposed weighting maps. The generation-confidence weighting map is encoded by the segmentation network that is trained on the source domain and tested on a target domain image. The discrimination-confidence weighting map is encoded by the auxiliary discrimination network that is trained on both domains. The two weighting maps interpret the inter-domain discrepancy from different perspectives. The combination of the two maps, Eq. (6), provides a refined weighting map for pixel-wise adaptation regularization.

In the auxiliary adversarial training, $\mathbf{D}^{\text{aux}}$ allocates a membership of the target domain, $\mathbf{D}^{\text{aux}}(\mathbf{F}(x^{(h,w)})) \in [0, 1]$, to each feature grid. It further infers that the encodings of $\mathbf{D}^{\text{aux}}$ represent the confidence of a source-trained feature encoder when generating consistent-with-target distributions. As such, pixel-wise domain discrepancy in the target domain is estimated by the discrimination confidence as $\mathbf{D}^{\text{aux}}(\mathbf{F}(x_t^{(h,w)}))$, from a different perspective. We also visualize the discrimination-confidence weighting map on the upper-right of Fig. 2. Lastly, the generation-confidence weighting map is refined by the discrimination-confidence weighting map, which is illustrated on the bottom of Fig. 2. The weighted adversarial adaptation loss is therefore updated as in (6):

$$\begin{aligned} \mathcal{L}_{\text{advII}}(\mathbf{G}, \mathbf{D}) = & -\mathbb{E}_{\mathbf{X}_s \in \mathcal{D}_S} [\sum_{(w,h)} \log(1 - \mathbf{D}(\mathbf{G}(x_s^{(w,h)})))] \\ & -\mathbb{E}_{\mathbf{X}_t \in \mathcal{D}_T} [\sum_{(w,h)} (d^{(w,h)} + \mu) \cdot \mathbf{D}^{\text{aux}}(\mathbf{F}(x_t^{(w,h)})) \cdot \log(\mathbf{D}(\mathbf{G}(x_t^{(w,h)})))] \end{aligned} \quad (6)$$

4.3 Network Overview and Optimization

The proposed pixel-wise weighted adaptation network is illustrated in Fig. 3, which is formed of three components: a generator $\mathbf{G}$, a discriminator $\mathbf{D}$ (embedded on the softmax output space) and an auxiliary discriminator $\mathbf{D}^{\text{aux}}$ (embedded on the final feature layer). The network training is carried out by the interaction between (1), (5) and (6), which are alternatively optimized according to the stages below:

- **Segmenter Updating.** The segmentation network is initially trained in a supervised way utilizing labeled source domain data $(\mathbf{X}_s, \mathbf{Y}_s) \in \mathcal{D}_S$. Parameters in $\mathbf{F}$ and $\mathbf{C}$ are updated by minimizing the loss function (1) as follows:

$$\min_{\mathbf{F}, \mathbf{C}} [\mathcal{L}_{\text{seg}}(\mathbf{F}, \mathbf{C})] \quad (7)$$

- **Generator updating.** At this stage, unlabeled target domain data $X_t \in \mathcal{D}_T$ are used to optimize the generator G and feature encoder F . Note that, unlike the previous stage, G and F are updated to confuse the D and D^{aux} by which the target-generated features and soft-max output will be considered as derived more likely from the source domain. The D and D^{aux} are fixed at this step, and G and F are updated by maximizing the loss functions (5) and (6) as in (8), where λ_{adv} is the trade-off weight used to balance the supervised training in (7) and adversarial training in (8).

$$\max_{G, F} \lambda_{\text{adv}} [\mathcal{L}_{\text{advII}}(G, D) + \mathcal{L}_{\text{aux}}(F, D^{\text{aux}})] \quad (8)$$

- **Discriminator updating.** Data and domain labels in $\mathcal{D}_S$ and $\mathcal{D}_T$ are used to update D and D^{aux} simultaneously to ensure the capability of distinguishing soft-max output and feature encodings respectively. At this stage, the F and C are fixed. The loss functions (5) and (6) are minimized as in (9).

$$\min_{D, D^{\text{aux}}} [\mathcal{L}_{\text{advII}}(G, D) + \mathcal{L}_{\text{aux}}(F, D^{\text{aux}})] \quad (9)$$

5 Experiments

In this section, the proposed method is evaluated on two classical synthetic-to-real semantic segmentation tasks. Meanwhile we analyze the results with qualitative and quantitative comparisons to the state-of-the-art.

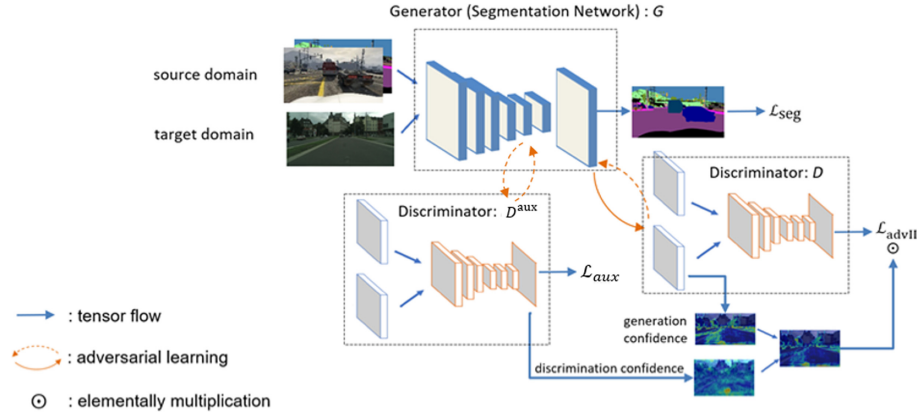


Fig. 3. A conceptual overview of the proposed weighted adaptation network. Entire network is composed of 3 different fully convolutional networks: segmentation (generation) network, G , and 2 discrimination networks, D and D^{aux} . The weighted adaptation is deployed through an adversarial learning with the pixel-wise weighting map.

5.1 Datasets

Cityscapes [16] is a real-world street scene dataset collected by dash cameras mounted on a moving car wandering in European cities. It contains 2,975 training images and 500 validation images, with high resolution (2048×1024), and pixel-wise labels in 34 categories of street objects. The training set (without labels) is considered as the target domain. **GTA5** [17] is a synthetic street scene dataset extracted from a realistically rendered computer game: Grand Theft Auto V. As rendered and annotated by a graphics engine, it forms a large dataset with 24,966 images with high resolution (1914×1052), and pixel-wise labels in 19 of the 34 categories of Cityscape. The entire image set with ground truth labels are used as the source domain in the task “GTA5 to Cityscapes”. **SYNTHIA** [4] contains 9,400 synthetic images with 16 of the 19 categories of GTA5. The resolution of each image is 760×1280 . While this dataset is also rendered by a graphics engine, it is less realistic than GTA5 and uses different viewing angles. That presents more severe domain shift that will challenge the adaptation capability of the proposed model. This dataset is used as the source domain in the task “SYNTHIA to Cityscapes”.

5.2 Implementation Details

The proposed network is deployed using PyTorch on a NVIDIA GTX 1080Ti GPU. Following the settings in MRNet [12], we use ResNet-101 with memory module as the FCN backbone for generator G (the segmentation network). Discriminator D comprises four convolutional layers and a classifier layer with stride 2 and kernel size 4 followed by a leaky ReLU. The auxiliary discriminator D^{aux} comprises three convolution layers with stride 1 and kernel size 1 followed by a leaky ReLU, and a classifier layer with kernel size 1. Before feeding into D and D^{aux} , the encodings from output space and feature space are up-sampled into the input image size of $W \times H$.

The performance of the proposed method is evaluated under the metric of mean Intersection over Union (mIoU) proposed in [15]. During hyper-parameter learning, we first consider threshold T that is used to protect well aligned (high generation confidence) pixels from over-adaptation. Following self-training strategies proposed in [25, 27], we set T to 0.9 to involve sufficient well aligned pixels. In our best model, the adaptive weight μ is set to 0.1. We then consider hyper-parameter λ_{adv} as a trade-off for the supervised training in (7) and the adversarial training in (8). We follow [10] to set it as 0.0005.

5.3 Adaptation from GTA5 to Cityscapes

The overall quantitative experimental results over 19 classes are detailed in Table 1. The “Source-only” model (trained on data from the source domain (GTA5) only and inferred on the target domain (Cityscapes) reaches 36.6% on mean IoU. The proposed method outperforms the “source-only” model by 10.6% on mean IoU after imposing domain adaptation training. We also compare against the method proposed by Tsai *et al.* [10] considered here as a vanilla method for output-space based DA network. The proposed method brings a 5.8% improvement over [10] on mean IoU and demonstrates efficient

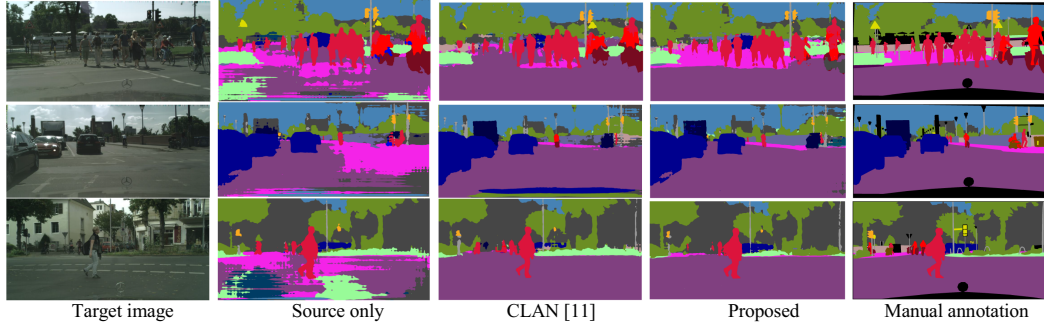


Fig. 4. Qualitative results obtained for semantic segmentation without and with domain adaptation. The third column is the adaptation results shown in CLAN [11]. The fourth column is the proposed pixel-wise weighted adaptation results, and the fifth column is the ground truth manual annotation.

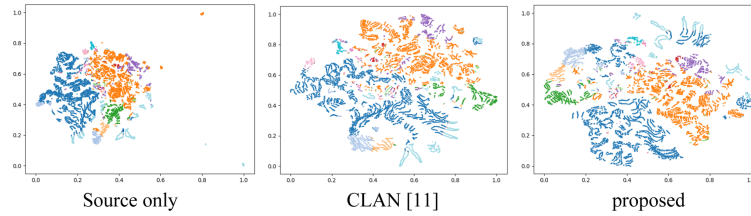


Fig. 5. Feature clusters visualization by t-SNE [34]. Each color represents a different class cluster.

alleviation of over-adaptation on several classes (e.g., “pole” and “bike”) as a result of the proposed pixel-wise weighted adaptation regularization.

Besides, we also present experimental comparisons with three state-of-the-art DA models [11, 12], and [29] in Table 1. Luo *et al.* [11] introduce a weighted adaptation method based on a different discrepancy measurement. The proposed method outperforms [11] on mean IoU by a margin of 4%. It shows a more efficient adaptation operation under similar complexity of the model structure (i.e., two discriminators vs. two generators). Zheng and Yang [12] recently proposed an output-space adaptation method that introduces a memory module for semantic segmentation. Conversely, Pan *et al.* [29] utilize entropy information in the intra-domain space for adaptation operation. As shown in Table 1, the proposed method leads to further adaptation improvements for [12] and [29] in general mean IoU, demonstrating the complementary effectiveness of the proposed adaptation method on the conventional methods that were elaborated on different DA strategies. Nevertheless, the proposed method happens to underperform on some classes (e.g., “train” and “truck”). We infer that a possible unbalance may develop in the pixel-wise weighted adaptation when several categories are involved simultaneously, which deserves further investigation.

At last, we illustrate qualitative experimental evaluation of the proposed method. First, in Fig. 4, it can be observed that the “Source only” model experiences a severe drop on segmentation performance when applied on the target domain. Compared to the “Source only” model, the proposed method provides a clear improvement in the segmentation results. Second, in comparison to [11], there are visible improvements on specific classes in the segmentation results, demonstrating the adaptation capability of the proposed model on those classes. Lastly, we visualize the feature space encodings by using t-SNE [30] in Fig. 5. It reveals how the proposed method provides more separable feature clusters for the target domain, which facilitates the domain adaptation process.

5.4 Adaptation from SYNTHIA to Cityscapes

We next demonstrate the efficacy of the proposed method on the task “SYNTHIA to Cityscapes”. As detailed in Table 2, the proposed method brings a 12.3% improvement over the “Source only” model compared to a 7.3% improvement obtained in [10]. It is also worth noting that the leading performance of [29] on mean IoU in the task “*GTA5 to Cityscapes*” is here surpassed by the proposed method and by [12], which indicates that, even though all three methods leverage different underlying principles, the proposed method could obtain a more reliable performance in both cross-domain adaptation tasks. Regarding specific classes, the proposed method outperforms [12] on seven classes, such as “sky”, “person” and “moto.”, and outperforms [29] on nine classes, such as “side.”, “person” and “bike”.

5.5 Ablation Studies

Three experimental ablation studies are conducted with respect to the two proposed weighting maps (illustrated in Fig. 2), to evaluate the efficacy of each component introduced in this work. First, we examine the generation-confidence weighting map when disregarding the protection of well aligned pixels by setting $T = 1$ (**GenMap** ^{$T=1$}). As shown in Table 3, **GenMap** ^{$T=1$} provides an improvement of 1.0% on mean IoU compared to the DA method developed upon the baseline [12] only. It indicates that while the generation-confidence weighting map might be limited by the implicit uncertainty mentioned in section IV.B, it still remains capable to implement pixel-wise adaptation regularization. The ablation study also conducts domain adaptation based on the generation confidence weighting map with a lower threshold at $T = 0.9$ (**GenMap** ^{$T=0.9$}), and on combined generation and discrimination confidence weighting maps (**GenMap** ^{$T=0.9$} + **DisMap**), under the experimental settings detailed in section V.B. As shown in Table 3, the adaptation model with **GenMap** ^{$T=0.9$} reaches mean IoU of 46.8%, which indicates that an improvement can be achieved through applying thresholding ($T = 0.9$). Furthermore, the combined adaptation model **GenMap** ^{$T=0.9$} + **DisMap** reaches 47.2%, demonstrating that **DisMap** is a valid complementary strategy to further refine the generation confidence to reach higher performance.

Table 1. Experimental results of the adaptation task “GTA5 to Cityscapes” under the metric of mean IoU.

-GTA5 to Cityscapes-	Road	Side	Build	Wall	Fence	Pole	Light	Sign	Vege	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Moto	Bike	mIoU
Source only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdaptSegNet [10]	86.5	25.9	79.8	22.1	20.0	23.0	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
CLAN [11]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
MRNet [12]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.7	85.3	33.7	76.4	60.2	33.7	86.0	36.1	43.3	5.9	22.8	30.8	45.5
IntraDA [29]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
Proposed	91.4	55.1	80.3	30.6	18.1	35.0	35.0	27.9	84.3	29.8	74.6	60.9	28.0	83.0	28.2	42.5	0.1	35.8	48.9	47.2

Table 2. Experimental results of the adaptation task “SYNTHIA to Cityscapes” under the metric of mean IoU.

-SYNTHIA to Cityscapes-	road	side	build	light	sign	vege	sky	person	rider	car	bus	moto	bike	mIoU
Source only	55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
AdaptSegNet [10]	79.2	37.2	78.8	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.1	45.9
CLAN [11]	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.1	73.0	32.9	22.6	30.7	47.8
MRNet [12]	82.0	36.5	80.4	18.0	13.4	81.1	80.8	61.3	21.7	84.4	32.4	14.8	45.7	50.2
IntraDA [29]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9
Proposed	85.1	41.2	79.2	10.1	13.1	79.0	85.6	61.7	26.6	77.4	36.4	23.4	42.6	50.9

Table 3. Ablation studies in the task “GTA5 to Cityscapes”.

Methods	MRNet [12]	GenMap ^{T=1}	GenMap ^{T=0.9}	DisMap	mIoU
DA without pixel-wise weighting	✓				45.5
Pixel-wise weighted DA ($T = 1$)	✓	✓			46.5
Pixel-wise weighted DA ($T = 0.9$)	✓		✓		46.8
Pixel-wise weighted DA (refined w/ DisMap)	✓		✓	✓	47.2

6 Conclusion

In this work, we propose a pixel-wise weighted adversarial adaptation framework. We first utilize generation confidence to regularize the output-space adaptation process at a finer level of details. Second, we introduce an auxiliary feature space based discriminator as a complementary discrepancy indicator to refine generation confidence, which contributes additional improvements on adaptation. The experimental results demonstrate that the proposed method is effective for over-adaptation alleviation and outperforms leading state-of-the-art methods for semantic segmentation. Future work will involve designing more efficient domain discrepancy metrics and combining DA schemes to improve the performance of the proposed method for application on cross-domain semantic segmentation.

Acknowledgments. This work was supported by the China Scholarship Council (CSC) under Grant: 201906290105, and by Natural Sciences and Engineering Research Council of Canada under the Discovery grants program.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Trans. Patt. Recog. Mach. Intell. (PAMI) **40**(4), 834–848 (2018)

2. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890 (2017)
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
4. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016)
5. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias. CoRR, vol. [arXiv:1505.01257](https://arxiv.org/abs/1505.01257) (2015)
6. Gong, B., Sha, F., Grauman, K.: Overcoming dataset bias: an unsupervised domain adaptation approach. In: NIPS Workshop on Large Scale Visual Recognition and Retrieval (LSVRR) (2012)
7. Hoffman, J., Wang, D., Yu, F., Darrell, T.: FCNS in the wild: Pixel-level adversarial and constraint-based adaptation. CoRR, abs/1612.02649 (2016)
8. Hoffman, J., et al.: Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML) (2018)
9. Chen, Y. H., et al.: No more discrimination: cross city adaptation of road scene segmenters. In: IEEE International Conference on Computer Vision (ICCV) (2017)
10. Tsai, Y.H., Hung, W.C., Schuster, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7472–7481 (2018)
11. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2507–2516 (2019)
12. Zheng, Z., Yang, Y.: Unsupervised scene adaptation with memory regularization in vivo. In: International Joint Conference on Artificial Intelligence (IJCAI) (2020)
13. Zhang, C., Wang, L., Yang, R.: Semantic segmentation of urban scenes using dense depth maps. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 708–721. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_51
14. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
15. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. (IJCV) **111**(1), 98–136 (2015)
16. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016)
17. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 102–118. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_7
18. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML) (2015)
19. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_16
20. Sun, B., Saenko, K.: From virtual to reality: fast adaptation of virtual object detectors to real domains. In: BMVA British Machine Vision Conference (BMVC) (2014)
21. Vazquez, D., Lopez, A.M., Marin, J., Ponsa, D., Geronimo, D.: Virtual and real world adaptation for pedestrian detection. Trans. Pattern Recog. Mach. Intell. (PAMI) **36**(4), 797–809 (2014)

22. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: addressing domain shift for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3752–3761 (2018)
23. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3723–3732 (2018)
24. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional Learning for Domain Adaptation of Semantic Segmentation,” in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6936–6945 (2019)
25. Biasetton, M., Michieli, U., Agresti, G., Zanuttigh, P.: Unsupervised domain adaptation for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
26. Goodfellow, J. et al.: Generative adversarial nets. In: NIPS (2014)
27. Zou, Y., Yu, Z., Vijaya Kumar, B.V.K., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 297–313. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_18
28. Zou, Y., Yu, Z., Liu, X., Kumar, B.V., Wang, J.: Confidence regularized self-training. In: IEEE International Conference on Computer Vision (ICCV), pp. 5982–5991 (2019)
29. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Maaten, L.V., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11), 2579–2605 (2008)

3.2 Prototypical knowledge adaptation

3.2.1 Summary

To support further improvement in cross-domain semantic segmentation, **ESDA** also introduces a new domain adaptation loss, the soft adversarial adaptation loss, which allows domain adaptation models to treat target domain pixels particularly and precisely. In this way, the development of prototypical knowledge allows elaborating specific adaptation strategies on under-aligned regions and well-aligned regions of the target domain.

In its implementation, it first leverages prototypical knowledge on the target domain to relax its hard domain label to a continuous domain space, where pixel-wise domain adaptation is developed upon a soft adversarial loss. Second, it employs a unilateral discriminator to alleviate implicit uncertainty on prototypical knowledge. The proposed method does not introduce additional parameters beyond that of the conventional methods while demonstrating competitive performance with state-of-the-art methods.

3.2.2 Original contributions

The main contributions of **ESDA** in this section are threefold:

- The introduction of prototypical knowledge for output space-based DA models which allows the latter to treat target domain pixels particularly and precisely;
- An unsupervised output-space based adversarial domain adaptation approach that formulates a soft adversarial adaptation loss without introducing additional parameters beyond that of conventional adversarial loss;
- A unilateral discriminator on the feature space that facilitates the interpretation of inter-domain discrepancy.

3.2.3 Article

This research has been presented in “A Prototypical Knowledge Oriented Adaptation

Framework for Semantic Segmentation” (H. Tian, S. Qu, P. Payeur), which is published at *IEEE Transactions on Image Processing*, 2021 [103]. Its online version is available at <https://ieeexplore.ieee.org/abstract/document/9623375>. The original paper is provided below as published. The paper is reprinted in this thesis with permission of the IEEE.

A Prototypical Knowledge Oriented Adaptation Framework for Semantic Segmentation

Haitao Tian^{ID}, Shiru Qu, and Pierre Payeur^{ID}

Abstract—A prevalent family of fully convolutional networks are capable of learning discriminative representations and producing structural prediction in semantic segmentation tasks. However, such supervised learning methods require a large amount of labeled data and show inability of learning cross-domain invariant representations, giving rise to overfitting performance on the source dataset. Domain adaptation, a transfer learning technique that demonstrates strength on aligning feature distributions, can improve the performance of learning methods by providing inter-domain discrepancy alleviation. Recently introduced output-space based adaptation methods provide significant advances on cross-domain semantic segmentation tasks, however, a lack of consideration for intra-domain divergence of domain discrepancy remains prone to over-adaptation results on the target domain. To address the problem, we first leverage prototypical knowledge on the target domain to relax its hard domain label to a continuous domain space, where pixel-wise domain adaptation is developed upon a soft adversarial loss. The development of prototypical knowledge allows to elaborate specific adaptation strategies on under-aligned regions and well-aligned regions of the target domain. Furthermore, aiming to achieve better adaptation performance, we employ a unilateral discriminator to alleviate implicit uncertainty on prototypical knowledge. At last, we theoretically and experimentally demonstrate that the proposed prototypical knowledge oriented adaptation approach provides effective guidance on distribution alignment and alleviation on over-adaptation. The proposed approach shows competitive performance with state-of-the-art methods on two cross-domain segmentation tasks.

Index Terms—Domain adaptation, semantic segmentation, prototypical knowledge, unsupervised learning, transfer learning.

I. INTRODUCTION

THE vision and cognition systems endow humans the ability to not only precisely segment objects from familiar scenes, but also transfer learnt visual knowledge to an

unknown scene. However, for computer vision models, such as semantic segmentation, cross-domain knowledge transfer remains challenging. With the goal of further boosting state-of-the-art performance [1], [2] for structural prediction, fully convolutional networks (FCNs) [3] have recently dominated the field of semantic segmentation, while involving training on huge pixel-wise labeled datasets in a supervised, end-to-end way. In spite of such massive training, FCN-based segmentation networks yet exhibit inherent limitations for practical use. In applications such as autonomous driving and robotic navigation that need to operate in changing scenarios, trained networks must deal with large appearance gaps, and thus may need to be retrained using a large number of pixel-wise labeled data in the new scenarios to transfer visual knowledge. However, it is difficult to collect and annotate such amounts of data in practice. A recently introduced and particularly appealing workaround [4] consists of utilizing photo-realistic synthetic street-scene images rendered by graphic engines which can simulate various scenarios for supervised network training. Nevertheless, an FCN pre-trained on synthetic datasets will generally perform poorly on real-world datasets (as shown in Fig. 1), which is mainly caused by a domain shift [5] in the data associated with different scenarios. The domain shift cannot be eliminated even with a large number of synthetic data.

Domain adaptation (DA), a transfer learning method [6] that is effective for handling label distribution mismatch between different domains, has been employed for computer vision tasks such as image classification to transfer learnt knowledge from a source domain to a target domain without using any labeled data from that target domain. In semantic segmentation tasks, this idea has been well leveraged to address domain-shift problems between synthetic datasets (source domain) and real-world datasets (target domain). Early researches [7]–[9] looked at the feature-level alignment that aims to align high-dimensional and suppressed features by employing a discrimination network (discriminator) on the feature space. More recently, researchers [10]–[13] paid closer attention to the use of structural information on the output space of FCN and hence obtained better adaptation performance than feature-space based methods. It is also revealed that feature-level adaptation provides limited improvement because DA suffers from encoding various complex visual cues that have been suppressed in the feature space. Given this, feature-based methods tend to resort to more extensive prior statistical knowledge from the target domain such as object proportion,

Manuscript received January 27, 2021; revised July 21, 2021 and September 14, 2021; accepted October 27, 2021. Date of publication November 22, 2021; date of current version November 30, 2021. This work was supported in part by the China Scholarship Council (CSC) under Grant 201906290105 and in part by the Natural Sciences and Engineering Research Council of Canada under the Discovery Grants Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. (Corresponding author: Haitao Tian.)

Haitao Tian is with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China, and also with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: htian026@uottawa.ca).

Shiru Qu is with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: qushiru@nwpu.edu.cn).

Pierre Payeur is with the Sensing and Machine Vision for Automation and Robotic Intelligence Research Laboratory, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: ppayeur@uottawa.ca).

Digital Object Identifier 10.1109/TIP.2021.3128311

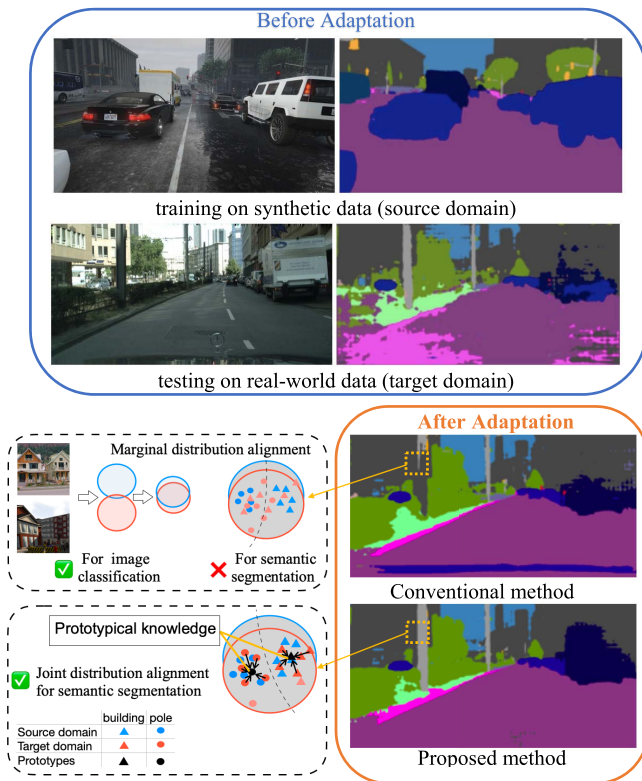


Fig. 1. Illustration of prototypical knowledge oriented adaptation. Top 2 rows: samples from source and target domain (left), a source-domain pretrained model for semantic segmentation inferred on samples without adaptation (right). An obvious performance drop can be observed on the prediction result of the target domain. **Bottom first row:** marginal distribution alignment is used for domain adaptation for image classification (left), where image labels are adapted by the alignment over global domain features ($P(F[\text{red}]) = P(F[\text{blue}])$). However, marginal distribution alignment would have negative effectiveness on output-space domain adaptation (right) because the overlook on local appearance variance causes visible over-adaptation results on the target domain, i.e., the segmentations are corrupted after domain adaptation. **Bottom second row:** The proposed adaptation method (left) utilizes prototypical knowledge to develop a joint distribution alignment (e.g., category features from the prototype “building” are aligned by: $P(\text{red}[\text{triangle}]) = P(\text{blue}[\text{triangle}])$) and obtains better improvement while it alleviates over-adaptation (right).

street scene layout, static object priors, etc., to reduce the inter-domain discrepancy.

Both DA models for image classification and semantic segmentation tasks assign hard domain labels to samples (images or pixels) and operate an adaptation process by optimizing a binary cross-entropy loss (further detailed in Section III.B.) However, we argue that the hard domain label assignment hinders the DA performance for semantic segmentation tasks and gives rise to over-adaptation performance on the target domain. First of all, different from cross-domain image classification tasks that operate marginal distribution alignment while handling an entire image as an instance, cross-domain semantic segmentation tasks need to align joint distributions by handling each pixel as an instance, to encode local appearance variance in a single image sample (see Fig. 1). Accordingly, inter-domain discrepancy of instances varies widely and causes intra-domain divergence among pixels in a single image sample. However, hard domain label assignment lacks consideration for this observation. Second, given the existence

of the possible prior domain semantic consistency between the source and the target domain, a source-trained FCN is able to correctly segment parts of aligned target regions without any adaptation. But binary cross-entropy adaptation loss cannot exempt those regions from over-adaptation. One simple solution to the hard domain label assignment is to deploy specific discriminators for each class on the feature space [9], [14], which aims to disentangle feature encodings for independent adaptation. Nevertheless, such an approach not only introduces a huge number of parameters to the DA network which gives rise to a computational burden during training, but also is limited by adaptation performance on the feature space.

In this paper, we introduce a new domain adaptation loss for cross-domain semantic segmentation tasks, named the soft adversarial adaptation loss, to tackle the over-adaptation problem existing in conventional adaptation methods and improve the adaptation performance. First, prototypical knowledge is used to relax hard domain labels to a continuous domain space where constraints from binary adversarial loss are disentangled and a soft domain label based adversarial adaptation loss is developed. Second, with the soft domain label, we treat local regions on the target domain separately and particularly, thereby proposing distinct adaptation strategies for under-aligned (domain-specific) and well-aligned (domain-invariant) regions, in order to conduct pixel-specific adaptation on the target domain. Lastly, given the uncertainty associated with prototypical knowledge on the output space, we introduce discrimination confidence yielded from a unilateral discriminator to refine the prototypical knowledge. In this way, the adaptation performance is further improved.

The proposed prototypical knowledge oriented adaptation framework draws on two key observations: (1) visual appearance changes, such as illumination and object context differences between the source and target datasets, may give rise to different degrees of inter-domain discrepancy among domain pixels; and (2) there exist embedding spaces in which this variance may be interpreted and measured in a DA network. The main contributions of this work are threefold:

- The introduction of prototypical knowledge for output-space based DA models, which allows DA models to treat target domain pixels particularly and precisely.
- The introduction of an unsupervised output-space based adversarial domain adaptation approach that formulates a soft adversarial adaptation loss without introducing additional parameters beyond that of conventional adversarial loss;
- The introduction of a unilateral discriminator on the feature space, which is trained in a unilateral way instead of adversarial one. It preserves a persistent capability on interpreting inter-domain discrepancy.

II. RELATED WORK

A. Supervised Pixel-Wise Structural Prediction

Semantic image segmentation is important for vision-based applications, e.g., scene understanding, autonomous driving,

medical image processing and robotic navigation. Fully convolutional networks [3] have brought significant leaps forward in semantic segmentation against traditional non-CNN models [15], [16]. Some state-of-the-art FCN-based models (e.g., Deeplab v2 [1], PSPNet [2]) achieved impressive performance on semantic segmentation. Meanwhile, benefitting from the increasing number of pixel-wise annotated datasets (e.g., CamVid [17], KITTI [18], MS COCO [19], PASCAL VOC 2012 [20], Mapillary [21], Cityscapes [22]), FCN-based models leverage a supervised network training approach and obtain high accuracy segmentation results. However, the side effect of supervised training is its heavy dependence on well annotated data, while it is impractical to collect datasets with widespread variability that cover various testing scenarios in the real world. Moreover, the annotation of datasets at the pixel level is extremely time-consuming (e.g., annotating one image in Cityscape would require more than 90 minutes). Based on the cost of access to labeled datasets, FCN-based segmentation networks were alternatively trained on synthetic datasets [4], [23], which are readily rendered and annotated by Graphic engines. Nevertheless, due to the domain discrepancy in the visual appearance, models trained on synthetic data tend to perform poorly when running on real-world datasets and scenarios [5].

B. Unsupervised Pixel-Wise Adversarial Domain Adaptation

Domain adaptation (DA) [6] has been successfully leveraged to approach the generalization issue of computer vision tasks such as cross-domain image classification [24], [25] and detection tasks [26], [27]. It is intuitive to implement DA on FCN-based segmentation networks to tackle the problem of domain-shift. A prevalent family of approaches [7]–[14], [28]–[42] specialize in leveraging adversarial adaptation methods to address the pixel-wise adaptation problem by embedding a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. For the DA model, the segmentation network not only learns discriminative representations, but also acts as a generative network (generator) interacting with the discriminator for invariant representation learning. In other words, adaptation progress can be realized by adversarial learning [43], where the discriminator is trained to best distinguish the source and target domains encodings, and the generator simultaneously tries to produce invariant outputs that emphasize cases where the discriminator is confused on their classification. As a result of adversarial interaction with the discriminator, the generator manages to produce invariant- discriminative features, such that the segmentation network can be applied on both domains.

There are three major types of DA models for segmentation networks, i.e., image-space based adaptation [8], [28], [30], [40] that considers each pixel of the input image as an adapting instance aiming to generate cross-domain similar looking images; feature-space based adaptation [7]–[9], [14], [29] that considers each spatial grid on the feature layer as an adapting instance; and output-space based adaptation [10]–[13], which considers each element on the soft-max output layer as an adapting instance. For example,

the method of Vu *et al.* [12] is an output-space adaptation method which utilizes entropy-based adversarial loss, while Zhou *et al.* [13] recently proposed an output-space adaptation method that introduces affinity space adaptation for semantic segmentation. In comparison to image-space and feature-space based approaches, adversarial DA relying on output space recently yielded better performance due to the usage of spatial information in the output space, though the method shows a certain degree of overfitting adaptation on the target domain.

C. Cluster Assumption for Domain Adaptation

The method proposed in this paper also relates to cluster assumption-based domain adaptation approaches. The cluster assumption states that the feature projections of samples should be clustered around category prototypes so that decision boundaries can go through low-density regions with low entropy. In cross-domain image classification tasks, Shu *et al.* [44] incorporate a virtual adversarial training and conditional entropy minimization to push the decision boundaries away from the target domain. Saito *et al.* [45] utilize the cluster assumption to develop a semi-supervised domain adaptation strategy, while Pan *et al.* [46] compute category prototypes on the source domain, target domain and source-target domain, in order to push the prototypes to be close in the feature space. With regard to cross-domain semantic segmentation tasks, Vu *et al.* [12] utilize entropy minimization and adversarial adaptation simultaneously to encourage unambiguous cluster assignments. The approach was evolved by Chen *et al.* [47] who introduce a maximum squares loss based on cluster assumption. Besides, Lee *et al.* [48] employ adversarial dropout to enforce the cluster assumption on the target domain, by which the decision boundaries are pushed away from the target domain features.

III. PRELIMINARIES

In this section, we introduce mathematical preliminaries on supervised segmentation model settings and unsupervised output-space based domain adaptation settings. Building upon that, the proposed approach is detailed in the next section.

Under the cross-domain adaptation for semantic segmentation settings, let's consider a source domain $\mathcal{D}_S = \{(X_s^{[i]}, Y_s^{[i]})\}_i^{n_S}$, where $X_s^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is the i th of n_S synthetic dataset images with a size of $W \times H \times 3$ in $\mathcal{D}_S$, and $Y_s^{[i]} \in \mathbb{R}^{W \times H \times L}$ is a corresponding pixel-wise annotation label with L categories. Also, we denote the target domain in the same way, i.e., $\mathcal{D}_T = \{(X_t^{[i]})\}_i^{n_T}$, where $X_t^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is a real-world dataset image. Note that there are no available labels in $\mathcal{D}_T$. The goal of this work is to train an FCN-based segmentation network on the source domain and transpose it to the target domain unsupervisedly. For clarity, (X_s, Y_s) and X_t correspond to a random sample from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively in the rest of this paper. We use $m^{(w,h)}$ to denote the element of a matrix M , e.g., $x_s^{(w,h)}$ is the pixel of X_s at the location (w, h) .

A. Supervised Semantic Segmentation Problem Settings

An FCN-based segmentation network trains a feature encoder $F(\cdot)$ that learns discriminative features and a dense

classifier $C(\cdot)$ that produces structural prediction (i.e., classifying each pixel into a specific category). During network training, a source domain pair (X_s, Y_s) is fed into F , then C takes high-dimensional feature encodings from F and produces a final label soft-max output $C(F(X_s))$. The segmentation loss is the multi-class cross entropy, formulated as:

$$\mathcal{L}_{\text{seg}}^S = -\mathbb{E}_{(X_s, Y_s) \in \mathcal{D}_S} \left[\sum_{(w, h)} \sum_l \mathbb{I}[l = y_s^{(w, h)}] \times \log C(F(x_s^{(w, h)}))^{(l)} \right] \quad (1)$$

where $l \in \{1, \dots, L\}$, $\mathbb{E}[\cdot]$ is the statistical expectation. Ground truth label Y_s is correspondingly encoded into one-hot vector by the indication function $\mathbb{I}[\cdot]$.

As for a target domain sample, $X_t \in \mathcal{D}_T$, the source-trained segmentation network under the parameter distribution of $\mathcal{D}_S$ can also generate a direct label soft-max output $C(F(X_t))$, while it would hardly reflect the genuine label distribution of $\mathcal{D}_T$ because of the presence of domain-shift. Meanwhile, the absence of pixel-wise annotation in $\mathcal{D}_T$ does not allow for fine-tuning on the trained segmentation network.

B. Unsupervised Output-Space Based Domain Adaptation Problem Settings

The output-space adaptation scheme considers the entire FCN-based segmentation model as a generator ($G = C[F(\cdot)]$), which generates as invariant soft-max output distributions as possible with regard to the two domains. This generalizing capability of G is endowed by a discriminator D , which is embedded on the output space, at the end of the segmentation network, to classify the structured prediction from different domains into a specific domain label. The G and D are updated alternatively by adversarial optimization as follows:

First, the discriminator D has the capability of classifying each element on the soft-max outputs $G(X_s)$ and $G(X_t)$ into its original domain label (we denote source domain label as 1, and target domain as 0). During training, samples from the two domains with hard domain label $(X_s, 1)$ and $(X_t, 0)$ are propagated forward into the generator and the discriminator successively. In this step, the segmentation network G just takes part in the forward propagation and the discrimination network D is updated with binary cross entropy by back propagation, formulated as:

$$\mathcal{L}_{\text{adv}}(D) = -\mathbb{E}_{X_s \in \mathcal{D}_S} \left[\sum_{(w, h)} \log(D(G(x_s^{(w, h)}))) \right] - \mathbb{E}_{X_t \in \mathcal{D}_T} \left[\sum_{(w, h)} \log(1 - D(G(x_t^{(w, h)}))) \right] \quad (2)$$

Second, as part of the DA model, the segmentation network G is also expected to generate invariant cross-domain distributions in order to eliminate the inter-domain discrepancy. Optimizing Eq. (3) with sample $(X_t, 1)$, G must manage to produce “source-style” soft-max output while it is actually trained with data from the target domain. In this stage, the back propagation is only applied on G , while D is fixed. Combining with supervised training using Eq. (1), the segmentation network G manages to generate discriminative-invariant distributions among pixels within the source and

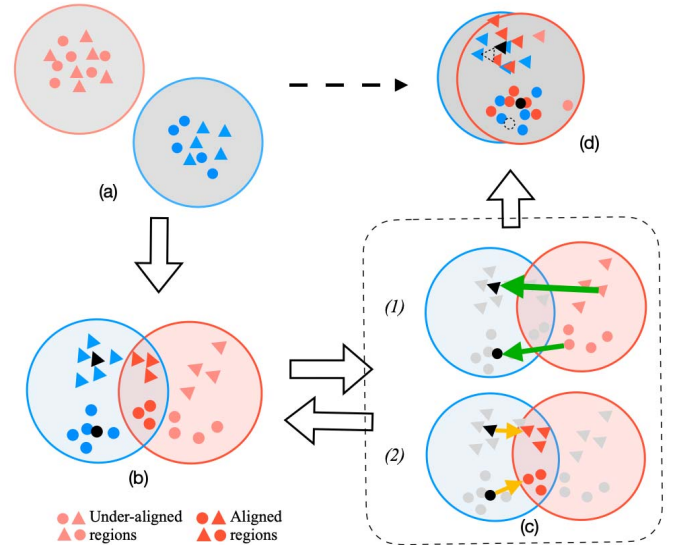


Fig. 2. Visualization of the prototypical knowledge adaptation (PKA) process. Only distribution projections of two classes are illustrated for clarity, and symbols are defined in Fig. 1. Four stages are considered in the adaptation process: (a) before adaptation; (b) supervised learning with prototypes on the source domain; (c) domain adaptation on the target domain; and (d) after adaptation. The training interaction between (b) and (c) is detailed in Section IV.C. The green and orange arrows denote the optimization of objective functions (7) and (9) respectively.

target domains.

$$\mathcal{L}_{\text{adv}}(G) = -\mathbb{E}_{X_t \in \mathcal{D}_T} \left[\sum_{(w, h)} \log(D(G(x_t^{(w, h)}))) \right] \quad (3)$$

IV. METHOD

In this section, the proposed prototypical knowledge oriented adaptation framework is defined with the objective to seek more precise guidance on the adaptation process in two ways. First, it extends the idea of conventional output-space DA approaches to a soft domain label-based pixel-specific adaptation strategy. Second, it resorts to a unilateral discriminator to refine the prototypical knowledge to further improve its adaptation performance.

A. Prototypical Knowledge Adaptation (PKA)

In conventional adaptation methods to approach domain shift problems for semantic segmentation, target domain distributions, along with hard domain label, are treated under the same distance from the source domain, by which gradients of the adaptation network are backpropagated in a global manner. After adaptation, the segmentation model would segment aligned target domain regions incorrectly and cause over-adaptation performance (see an example in Fig. 1). In this section, we first investigate the use of prototypical knowledge as an interpreter of soft domain label, then introduce a soft adversarial loss on under-aligned regions and a semi-supervised learning strategy on aligned regions. Our model is built upon conventional output-space DA models without introducing extra parameters.

1) *Prototypical Knowledge Interpretation (PKI)*: After the supervised training of a FCN on the source domain using Eq. (1), a pixel-wise classifier C is obtained and can be degraded into L estimated category prototypes

$\{C_1, \dots, C_L, \dots, C_L\}$, such that $\sum_{l=1}^L C_l(F(x_s^{(w,h)})) = 1$. From the perspective of cluster assumption [44], [45], source instances are clustered around class prototypes on the feature embedding (as shown in Fig. 2(b)) with supervision from the source domain ground truth. The category label of each instance $x_s^{(w,h)}$ is assigned to the nearest prototype according to the confidence $C_l(F(x_s^{(w,h)}))$. Due to the absence of labels on the target domain, each estimated prototype can be regarded as a representative point of a specific category within the source domain while keeping in the center of the source domain distributions. Therefore, given a target instance projected onto the feature embedding, even though not allowing to make precise prediction, prototypes measure the confidence of a source-trained FCN when generating source-like predictions on the target domain, which is referred to as $C_l(F(x_t^{(w,h)}))$.

Building upon the above observation, we define the prototypical knowledge of the target domain in Eq. (4).

$$\rho_l^{(w,h)} = \Phi[C_l(F(x_t^{(w,h)})) \cdot \tau_l] \quad (4)$$

where $\Phi(\cdot)$ represents a nonlinear transformation followed with sigmoid activation. To relieve the influence of the dominance of high frequency classes, we introduce the balance coefficient, $\tau_l = \sqrt{N^{\psi_l} / \sum_{\psi_l} C_l(F(x_t^{(w,h)}))}$, proposed in [14], where ψ^l is the l -th layer of one-hot outputs of X_t , and N^{ψ_l} is the nonzero pixel number of ψ^l . The prototypical knowledge indicates the extent of source-consistent semantic knowledge that a target instance learns from the source domain prototypes on the output space. We also illustrate this prototypical knowledge interpretation (PKI) process in Fig. 3.

2) *Soft Adversarial Loss*: As mentioned in Section III.B, conventional adaptation methods utilize hard domain labels $((X_s, \mathbf{1})$ and $(X_t, \mathbf{0})$) to develop the binary adversarial loss function. However, the usage of hard domain labels assigns the membership of domains to a Boolean set regardless of the existence of semantically consistent regions across domains. With the access to prototypical knowledge, it is possible to relax the binary domain membership to a continuous space, where we take the prototypical knowledge of the nearest prototype as the soft domain label of the target domain as:

$$\rho^{*(w,h)} = \max\{\rho_1^{(w,h)}, \dots, \rho_L^{(w,h)}\} \quad (5)$$

With the target domain label (X_t, ρ^*) , the binary cross-entropy term over a target domain sample in Eq. (2) is transformed into a soft cross entropy loss as defined in Eq. (6):

$$\mathcal{L}_{\text{soft}}^T = - \sum_{(w,h)} [\rho^{*(w,h)} \cdot \log(D(G(x_t^{(w,h)})))] \quad (6)$$

3) *Domain Adaptation Process*: In order to deploy precise guidance on adaptation, we split the soft-max output of target domain into two regions by a threshold T : the under-aligned regions (where $\rho^{*(w,h)} < T$) and the aligned regions (where $\rho^{*(w,h)} \geq T$). We then consider specific adaptation processes on under-aligned regions and on aligned regions respectively, as follows and shown in Fig. 2.

a) *PKA on under-aligned regions using soft adversarial loss*: Combined with soft adversarial loss $\mathcal{L}_{\text{soft}}^T$, we evolve the

conventional adaptation loss from Eq. (2) to an adversarial adaptation loss defined as:

$$\begin{aligned} \mathcal{L}_{\text{soft_adv}}(G, D) &= -\mathbb{E}_{X_s \in \mathcal{D}_S} [\sum_{(w,h)} \log(D(G(x_s^{(w,h)})))] \\ &\quad - \mathbb{E}_{X_t \in \mathcal{D}_T} [\sum_{(w,h) \in [UN]} \rho^{*(w,h)} \cdot \log(D(G(x_t^{(w,h)})))] \quad (7) \end{aligned}$$

This loss function (7) allows to localize adaptation operation over the source domain and the under-aligned regions [UN] of the target domain, meanwhile preventing aligned regions from adaptation by extinguishing gradients backpropagation on these regions.

b) *PKA on aligned regions using semi-supervised loss*: The semi-supervised learning strategy is used for adapting aligned regions in a non-adversarial way. Semi-supervised learning [41] is already commonly leveraged in DA methods as an efficient supplementary alignment approach with adversarial adaptation. Collaborating with adversarial adaptation that proceeds on under-aligned regions, semi-supervised learning operates the adaptation process on aligned regions.

To conduct the semi-supervised learning of prototypical knowledge adaptation (PKA) on aligned regions [AL], pseudo labels $\hat{y}_t^{(w,h)}$ are first distilled from prototypical knowledge by Eq. (8), where the indication function $\mathbb{I}[\cdot]$ encodes prototypical knowledge into one-hot vectors.

$$\hat{y}_t^{(w,h)} = \mathbb{I}[\text{argmax}\{\rho_1^{(w,h)}, \dots, \rho_L^{(w,h)}\}] \quad (8)$$

Second, since entropy minimization has demonstrated complementary effectiveness [42], [47] on the development of semi-supervised learning by encouraging unambiguous cluster alignments, we employ the negative entropy [42] on prototypical knowledge, $H^{(w,h)} = - \sum_{l=1}^L \rho_l^{(w,h)} \cdot \log(\rho_l^{(w,h)})$, to penalize over-confident prototypical knowledge values so as to prevent the network from early overfitting on easy-to-transfer regions.

Finally, the semi-supervised learning objective function is formulated in Eq. (9):

$$\begin{aligned} \mathcal{L}_{\text{seg}}^T &= -\mathbb{E}_{X_t \in \mathcal{D}_T} [\sum_{(w,h) \in [AL]} (\hat{y}_t^{(w,h)} \log G(x_t^{(w,h)}) \\ &\quad + \alpha H^{(w,h)})] \quad (9) \end{aligned}$$

where α controls the strength of entropy minimization. The overall network training is operated by minimaxing optimization among loss functions, which will be further detailed in Section IV.C.

B. Prototypical Knowledge Refinement (PKR)

The development of soft domain labels builds upon the cluster assumption that prototypical knowledge adaptation is highly related to the intra-domain semantic inconsistency. Nevertheless, the prototypical knowledge would contain implicit uncertainty information when involving uncertain predictions [53] on the output space. For instance, a target-domain pixel might be segmented incorrectly by a FCN learnt on the source domain because of the FCN's inherent inability in segmenting some regions (e.g., class boundaries and regions

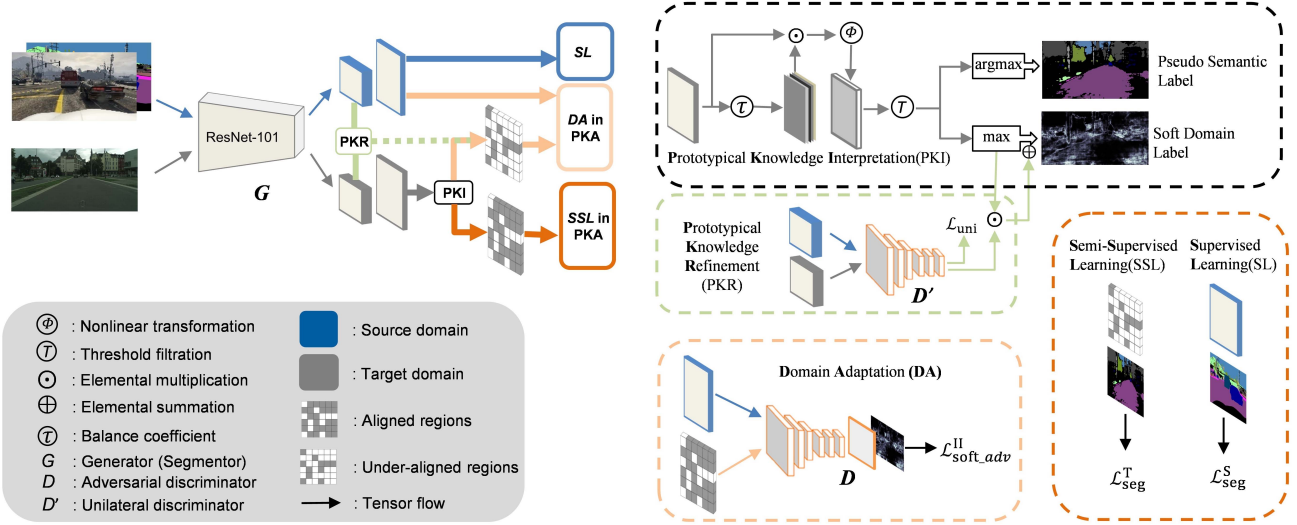


Fig. 3. A conceptual overview of the proposed prototypical knowledge oriented adaptation (PKA) framework. Entire network is parameterized by 3 different fully convolutional networks: generator G (ResNet-101 based segmentation network), and 2 discrimination networks, D and D' .

having large inter-class confusion [49]). In other words, under-aligned regions would contain aligned pixels that are far away from all prototypes, which eventually causes a side effect on the PKA process. To alleviate the influence of uncertain information on the prototypical knowledge, we propose a prototypical knowledge refinement strategy to enhance the performance of PKA.

1) *Unilateral Discriminator*: A unilateral discriminator D' on the feature space builds upon the assumption that discrimination confidence encoded by this unilateral discriminator can provide intra-domain discrepancy information from a different perspective. Different from conventional discriminators [7], [9], the unilateral discriminator D' is not trained in an adversarial manner but is trained separately by circumventing the interaction with the generator network. As a result, D' will remain discriminative to interpret intra-domain discrepancy (further discussed in Section V.F and Fig. 8).

In detail, the unilateral discriminator D' is trained by Eq. (10) to best discriminate the source and target domains encodings on the feature space.

$$\mathcal{L}_{\text{uni}}(D') = -\mathbb{E}_{X_s \in \mathcal{D}_S} [\sum_{(w,h)} \log(D'(F(x_s^{(w,h)})))] - \mathbb{E}_{X_t \in \mathcal{D}_T} [\sum_{(w,h)} \log(1 - D'(F(x_t^{(w,h)})))] \quad (10)$$

In each iteration, it allocates discrimination confidence $D'(F(x_t^{(w,h)})) \in [0, 1]$ to measure the intra-domain discrepancy of each feature grid. With discrimination confidence, we realize a control gate to refine prototypical knowledge, as defined in Eq. (11):

$$\hat{\rho}^{*(w,h)} = (1 + D'(F(x_t^{(w,h)}))) \cdot \rho^{*(w,h)} \quad (11)$$

Finally, the soft adversarial loss from Eq. (7) with $\hat{\rho}^{*(w,h)}$ is updated as in Eq. (12).

$$\mathcal{L}_{\text{soft_adv}}(G, D) = -\mathbb{E}_{X_s \in \mathcal{D}_S} [\sum_{(w,h)} \log(D(G(x_s^{(w,h)})))] - \mathbb{E}_{X_t \in \mathcal{D}_T} [\sum_{(w,h) \in [\text{UN}]} \hat{\rho}^{*(w,h)} \cdot \log(D(G(x_t^{(w,h)})))] \quad (12)$$

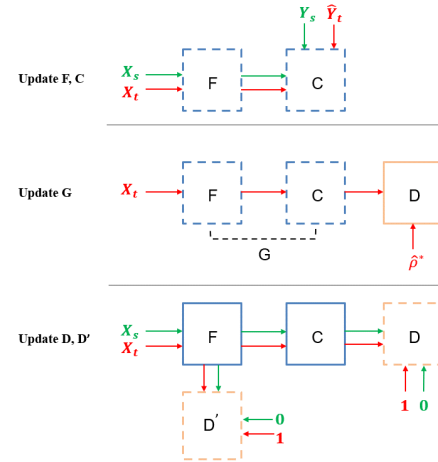


Fig. 4. Illustration of iterative updates of the proposed network. Input data and tensor flow on the source domain are denoted in green. Input data and tensor flow on the target domain are denoted in red. Network embeddings are denoted with dotted lines during back-propagation.

C. Network Overview and Optimization

We detail the proposed prototypical knowledge oriented adaptation framework in Fig. 3. The framework is parameterized by three convolutional networks: a generator G and two discriminators D and D' . In particular, D' is embedded on the final feature layer, and D is embedded on the soft-max output space. Note, G is the identical expression of $C \cdot F$ in the context of an output-space adaptation network.

The interactions between different network embeddings and adversarial optimization are critical to the network performance. During the network training, Eq. (1), (9), (10) and (12) are used to alternatively optimize the adaptation network according to the stages below. Note that there is no specific optimization order for stages in each iteration. The optimization stages are also shown in Fig. 4.

- **Segmenter updating.** At this stage, there are available labeled data from $\mathcal{D}_S$ and pseudo labels (produced by

Eq. (8)) from $\mathcal{D}_T$. The segmentation network is trained in a supervised way such that parameters in F and C are updated by minimizing the loss function Eq. (1) and Eq. (9) as follows:

$$\min_{F, C} [\mathcal{L}_{\text{seg}}^S(F, C) + \mathcal{L}_{\text{seg}}^T(F, C)] \quad (13)$$

- **Generator updating.** At this stage, unlabeled data in $\mathcal{D}_T$ are used to optimize the generator G , by maximizing the loss function Eq. (12) as in Eq. (14). The D is fixed at this step. λ_{adv} is the trade-off weight used to balance the segmenter and generator updating processes.

$$\max_G [\lambda_{\text{adv}} \mathcal{L}_{\text{soft_adv}}^{\text{II}}(G, D)] \quad (14)$$

- **Discriminators updating.** Data and domain labels in $\mathcal{D}_S$ and $\mathcal{D}_T$ are used to update D and D' simultaneously, so as to best distinguish soft-max output and feature encodings from the source and target domains. At this stage, the generator network is fixed. The loss functions Eq. (10) and Eq. (12) are jointly minimized as in Eq. (15).

$$\min_{D, D'} [\mathcal{L}_{\text{soft_adv}}^{\text{II}}(G, D) + \mathcal{L}_{\text{uni}}(F, D')] \quad (15)$$

D. Analysis

In this section, we theoretically evaluate the efficacy of the proposed prototypical knowledge oriented adaptation framework compared to conventional adaptation approaches.

Proposition 1: Semi-supervised learning in the PKA introduces a class-wise threshold filtration strategy in the process of pseudo label generation.

Proof: The class-wise threshold filtration strategy was initially proposed in [41], showing better performance than a global thresholding strategy. It sets a class-specific threshold k_l to filtrate pseudo labels for class l by the inequality operation $C_l(F(x_t^{(w,h)})) > k_l$.

In the PKA, prototypical knowledge on pseudo labels complies with the inequality (16):

$$\rho_l^{(w,h)} > T \quad (16)$$

$$\Leftrightarrow \Phi[C_l(F(x_t^{(w,h)})) \cdot \tau_l] > T$$

$$\Leftrightarrow C_l(F(x_t^{(w,h)})) > \Phi^{-1}(T)/\tau_l \quad (17)$$

The derivation in (17) shows that, even though a global threshold, T , is used in semi-supervised learning, the PKA utilizes a class-specific threshold $k_l = \Phi^{-1}(T)/\tau_l$ for pseudo label generation.

Proposition 2: PKA relaxes constraints from hard domain labels to a soft semantic similarity space, thereby achieving a tighter bound than conventional output-space based adaptation approaches.

Proof: We first revisit the domain adaptation theory which is initially proposed in [50]:

Theorem: Let H be a hypothesis space. $\epsilon_S(\cdot)$ and $\epsilon_T(\cdot)$ are corresponding generalization error functions for domains

$\mathcal{D}_S$ and $\mathcal{D}_T$. For all $h \in H$:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}d_H(\mathcal{D}_S, \mathcal{D}_T) + \lambda \quad (18)$$

where: $\lambda = \epsilon_S(h^*) + \epsilon_T(h^*)$, $\exists h^* = \underset{h \in H}{\operatorname{argmin}} \epsilon_S(h) + \epsilon_T(h)$.

In Eq. (18), distance $d_H(\mathcal{D}_S, \mathcal{D}_T)$ measures the supremum of the domain distribution divergence over $\mathcal{D}_S$ and $\mathcal{D}_T$, which can be further formalized by:

$$d_H(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h \in H} |Pr_{\mathcal{D}_S}[h((X_s))=1] - Pr_{\mathcal{D}_T}[h((X_t))=1]| \quad (19)$$

In conventional output-space based adaptation tasks, the minimization of $\epsilon_S(h)$ in Eq. (18) is solved by cross-entropy minimization (as formulated in Eq. (1)). Domain distribution divergence $d_H(\mathcal{D}_S, \mathcal{D}_T)$ in Eq. (19) is measured by learning a discriminator using Eq. (2), and then $d_H(\mathcal{D}_S, \mathcal{D}_T)$ is minimized by optimizing the generator G (as formulated in Eq. (3)). The interactions between the discriminator and the generator form the iterative adversarial learning process in conventional methods.

Taking a closer look on the relationship between $d_H(\mathcal{D}_S, \mathcal{D}_T)$ and adversarial adaptation loss in Eq. (2), hypothesis $h(\cdot)$ in Eq. (19) can be denoted as $D(G(\cdot)) \in H$, and given hard domain labels $(X_s, \mathbf{1})$ and $(X_t, \mathbf{0})$, the domain distribution divergence $d_H(\mathcal{D}_S, \mathcal{D}_T)$ can be rewritten as in Eq. (20):

$$d_H(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{D \in H} \left| Pr_{\mathcal{D}_S}[D(G((X_s)))=1] + Pr_{\mathcal{D}_T}[D(G((X_t)))=0] - 1 \right| \quad (20)$$

We shall see that PKA will achieve a tighter bound in $H \Delta H$ space [51] than conventional adversarial adaptation methods do in H space.

In PKA, soft domain labels $(X_s, \mathbf{1})$ and (X_t, ρ^*) are used to evolve binary adversarial loss. If we set:

$$\Gamma(X) = \begin{cases} \mathbf{1}, & \text{if } X \in \mathcal{D}_S \\ \rho^*, & \text{if } X \in \mathcal{D}_T, \end{cases} \quad (21)$$

the proposed soft adversarial loss (7) is subject to the $H \Delta H$ distance [51]:

$$d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{D \in H, \Gamma \in H} \left| Pr_{\mathcal{D}_S}[D(G(x_s(X_s))) = \Gamma(X_s)] - Pr_{\mathcal{D}_T}[D(G((X_t))) = \Gamma(X_t)] \right| \quad (22)$$

If we define $D(G(X)) \Delta \Gamma(X) = \{D^* : D^*(X^*) = 1\}$, where $X^* = \{X : D(G(X)) \oplus \Gamma(X)\}$, $\oplus$ is XOR operator, we can denote $d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T)$ as in Eq. (23):

$$d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{D^* \in H \Delta H} \left| Pr_{\mathcal{D}_S}[D^*(G((X_s))) = 1] + Pr_{\mathcal{D}_T}[D^*(G((X_t))) = 0] - 1 \right| \quad (23)$$

$$\leq 2 \sup_{D \in H} \left| Pr_{\mathcal{D}_S}[D(G((X_s))) = 1] + Pr_{\mathcal{D}_T}[D(G((X_t))) = 0] - 1 \right| \quad (24)$$

In Eq. (23), PKA shows a similar theoretical basis with conventional adaptation methods which is formulated in Eq. (22),

aiming to develop a discriminator D^* that is able to best distinguish encodings from source and target domains. However, in the inequality of Eq. (24), by reducing $d_{H\Delta H}$, PKA yields a tighter bound [51] over conventional methods that are developed on H hypothesis space. In this way, by relaxing constraints from hard domain labels, PKA obtains better adaptation performance compared with conventional methods.

V. EXPERIMENTS

In this section, we experimentally evaluate the efficacy of the proposed framework by exploiting cross-domain semantic segmentation tasks, and analyze the results qualitatively and quantitatively with comparisons to state-of-the-art approaches.

A. Datasets

Cityscapes [22] is a real-world street scene dataset collected by dash cameras mounted on a moving car wandering in European cities. It contains 2,975 training images and 500 validation images, with high resolution (2048×1024), and pixel-wise labels in 34 categories of street objects. This dataset is not used for training but rather the training set (without labels) is considered as the target domain.

GTA5 [23] is a synthetic street scene dataset extracted from a realistically rendered computer game: Grand Theft Auto V. As rendered and annotated by a Graphic engine, it forms a large dataset with 24,966 images with high resolution (1914×1052), and pixel-wise labels in 19 of the 34 categories of Cityscape. The entire image set with ground truth labels is used as the source domain in the task “GTA5 to Cityscapes.”

SYNTHIA-RAND-CITYSCAPES [4] contains 9,400 synthetic images with 16 of the 19 categories of GTA5. The resolution of each image is 760×1280 . While this dataset is also rendered by a Graphic engine, it is less realistic than GTA5 and uses different viewing angles. That presents more severe domain shift that will challenge the adaptation capability of the proposed model. This dataset is used as the source domain in the task “SYNTHIA to Cityscapes.”

B. Implementation Details

The proposed network is deployed using PyTorch on a NVIDIA GTX 1080Ti GPU. We use DeepLab-v2 [1] with ResNet-101 as FCN backbone for the segmentation network. Discriminator D comprises four convolutional layers with stride 2 and kernel size 4 followed by a leaky ReLU, and a classifier layer with same stride and kernel size. The discriminator D' comprises three convolutional layers with stride 1 and kernel size 1 followed by a leaky ReLU, and a classifier layer with kernel size 1. Before feeding into D and D' , the encodings from output space and feature space are up-sampled to the size of the input image: $W \times H$. The hyperparameter λ_{adv} is initialized as 0.001 and decayed by a damping policy with the multiplier $(1 - \frac{iter.}{max_iter.})$. In our best model, α is set as 0.005.

C. Baselines

The proposed method is able to readily combine with vanilla DA paradigms, while keeping the same number of network parameters. We consider two output-space based domain adaptation models as baselines for our experiments:

- **ASNet** [10] (CVPR 2018) is the first output-space based DA method which has been widely adopted as the baseline model in the state-of-the-art works [11]–[13]. We use ASNet (single mode) with implementation details reported in [10] as our “Baseline I.”
- **MRNet** [53] (IJCAI 2020) is a recently proposed method to evolve the ASNet with a simple but efficient memory regularization module. It does not introduce any extra parameters compared to [10]. We adopt MRNet (State1) with corresponding implementations as our “Baseline II.”

D. Adaptation From GTA5 to Cityscape

The overall quantitative experimental results over 19 classes are detailed in Table I. It shows that when the model is trained on data from the source domain (GTA5) only (“without adaptation”) and inferred on the target domain (Cityscapes), the performance measured as the mean Intersection over Union (mIoU) [18] reaches 36.6%. The rest of Table I are adaptation strategies from the literature compared to the proposed PKA and PKA + PKR (denoted as PKA+ in the rest of this paper) respectively. In this section, we experimentally compare the proposed methods with the state-of-the-art approaches in two ways.

1) *Comparison With the Baselines:* Results in Table I show that both baselines [10] and [53] surpass the without-adaptation model by a clear margin. The latter achieves an 8.9% improvement on mean IoU from the without-adaptation model. The PKA models (PKA (Baseline I) and PKA (Baseline II)) are developed upon Baseline I [10] and Baseline II [53], respectively, while keeping the same number of network parameters. By exploiting the prototypical knowledge on the output space and proceeding with pixel-specific adaptation strategy, PKA outperforms the corresponding Baseline I and Baseline II by a margin of 5.7% and 3.7% on mean IoU, respectively. Close inspection of the specific classes reveals that, PKA outperforms the baselines in most categories (e.g., “wall,” “terrain,” “bus” and “bike”).

2) *Comparison With State-of-the-Art:* Given that PKA relies on a different joint domain alignment strategy than another recently introduced pixel-wise weighted adaptation approach CLAN [11] that evolves the method in Baseline I [10], we also include a comparison with [11]. From the perspective of methodology, Luo *et al.* [11] employ two classifiers and consider the cousin-distance as the inter-domain discrepancy metric, while PKA only employs one classifier and considers more accessible prototypical knowledge on the output space, demonstrating a more intuitive and simpler adaptation network. From the experimental results in Table I, PKA (Baseline I) demonstrates better adaptation performance than [11] both in terms of overall mIoU and classes-specific IoU. In the overall comparison with other methods, PKA models outperform recently proposed output-space based methods [12], [13], [54]

TABLE I

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK “GTA5 TO CITYSCAPES” UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE. THE HIGHLIGHTED RESULTS SHOW BEST IOU AND BEST MEAN IOU IN EACH COLUMN

-GTA5 to Cityscapes-		road	side.	build.	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	moto.	bike	mIoU
Without adaptation		75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Adversarial Domain Adaptation																					
feature space	ASNet (fea.) [10]	83.7	27.6	75.5	20.3	19.9	27.4	28.3	27.4	79.0	28.4	70.1	55.1	20.2	72.9	22.5	35.7	8.3	20.6	23.0	39.3
	SIBAN [29]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
	SSF-DAN [14]	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
	ASNet [10]	86.5	25.9	79.8	22.1	20.0	23.0	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
output space	CLAN [11]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
	AdvEnt [12]	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
	AdvEnt+Minent [12]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
	ASA [13]	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	16.9	34.5	30.8	45.1
	MRNet [53]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.7	85.3	33.7	76.4	60.2	33.7	86.0	36.1	43.3	5.9	22.8	30.8	45.5
	IntraDA [54]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
	PKA (baseline I)	88.1	30.7	82.2	33.1	23.3	32.2	35.6	28.0	84.4	43.2	77.2	61.2	28.2	85.2	34.2	49.2	0.0	34.2	43.3	47.1
	PKA (baseline II)	90.4	33.1	84.6	31.7	24.4	36.4	42.2	41.3	86.8	40.6	78.8	64.1	36.5	87.9	37.4	48.0	2.3	26.7	40.9	49.2
Non-adversarial Domain Adaptation																					
Context Aware [55]		91.3	46.0	84.5	34.4	29.7	32.6	35.8	36.4	84.5	43.2	83.0	60.0	32.2	83.2	35.0	46.7	0.0	33.7	42.2	49.2
PIT [56]		87.5	43.4	78.8	31.2	30.2	36.3	39.9	42.0	79.2	37.1	79.3	65.4	37.5	83.2	46.0	45.6	25.7	23.5	49.9	50.6
MaxSquare [47]		89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	84.6	36.4	43.0	5.5	34.7	33.5	46.4
CBST [42]		91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
CRST (MRENT) [42]		91.8	53.4	80.6	32.6	20.8	34.3	29.7	21.0	84.0	34.1	80.6	53.9	24.6	82.8	30.8	34.9	16.6	26.4	42.6	46.1
PKA+ (baseline I)		91.6	50.5	84.4	34.5	27.5	32.5	36.4	31.1	85.5	37.1	84.8	60.8	30.2	81.2	28.5	38.5	0.0	38.4	45.7	48.4
PKA+ (baseline II)		93.0	55.0	85.9	32.7	30.0	34.5	40.0	36.3	85.7	40.6	86.4	62.5	32.9	82.8	26.7	50.5	0.2	39.8	49.4	50.8

and feature-space based methods [10], [14], [29] in terms of mean IoU, demonstrating the superior effectiveness of the proposed soft adversarial adaptation. When taking a close look on the results of IoU on infrequent classes and small objects in Table I, both the proposed PKA methods and other adversarial learning methods vary in DA performance on different classes, which reflects the different underlying adaptation strategies used by them. Besides, the PKA methods are not only more efficient on specific classes, such as “person,” “wall” and “terrain,” but also show effectiveness for over-adaptation alleviation, e.g., baselines [10] and [53] exhibit obvious over-adaptation on the class “bike” and “fence” compared to the without-adaptation model, while the PKA methods address it efficiently. Nevertheless, it also can be observed that the PKA methods occasionally underperform other approaches, as exhibited on the class “train.” We infer that prototypical knowledge on that category is not effectively utilized as there is a large inter-class confusion [49] on that category in the source domain, which inspires our future works to improve this problem.

With regard to PKA+, which is proposed to further boost the performance of PKA by employing a unilateral discriminator, the study compares PKA+ (PKA+ (Baseline I) and PKA+ (Baseline II)) with a variety of state-of-the-art non-adversarial adaptation methods. As shown in Table I, the two PKA+ models yield comparable results with non-adversarial DA methods [42], [47], [55], and [56] in terms of mean IoU, showing the convincing effectiveness of the proposed method in the domain adaptation community.

In terms of qualitative experimental evaluation, Fig. 5 illustrates the efficacy of the proposed method. First, it is seen that,

compared to a trained model that does not involve adaptation, the proposed methods provide a noticeable improvement in the segmentation. In comparison to [10], there are visible improvements on specific classes in the segmentation results, demonstrating the fundamental adaptation capability of the proposed models on those classes. Second, we visualize the feature clusters formed in the embedding space in Fig. 6 by using t-SNE [57]. We can observe that even though visible adaptation results are achieved from ASNet [10], a certain degree of domain misalignment exists. The proposed PKA and PKA+ provide more uniform and separable feature clusters, resulting in better performance in the target domain.

E. Adaptation From SYNTHIA to Cityscapes

Table II refers to the experimental results of the PKA and PKA+ methods on the task “SYNTHIA to Cityscapes.” Similar to [10]–[13], the proposed models are evaluated on 13 classes on Cityscapes. As detailed in Table II, without adaptation, the model (trained on SYNTHIA only) reaches 38.6% on mean IoU. PKA outperforms the two baselines consistently, demonstrating the fundamental efficacy of PKA on this different cross-domain scenario.

When compared to the adversarial methods ([11]–[13], and [54]), similar effectiveness as discussed in “GTA5 to Cityscapes” is observed in Table II. More specifically, the proposed PKA (Baseline I) method shows comparable overall performance to [11] on mean IoU with slightly different performance on IoU for specific classes while leveraging different underlying principles. Besides, the proposed PKA+ (Baseline II) outperforms the non-adversarial methods, Context Aware [55], PIT [56], MaxSquare [47], CBST [42], and

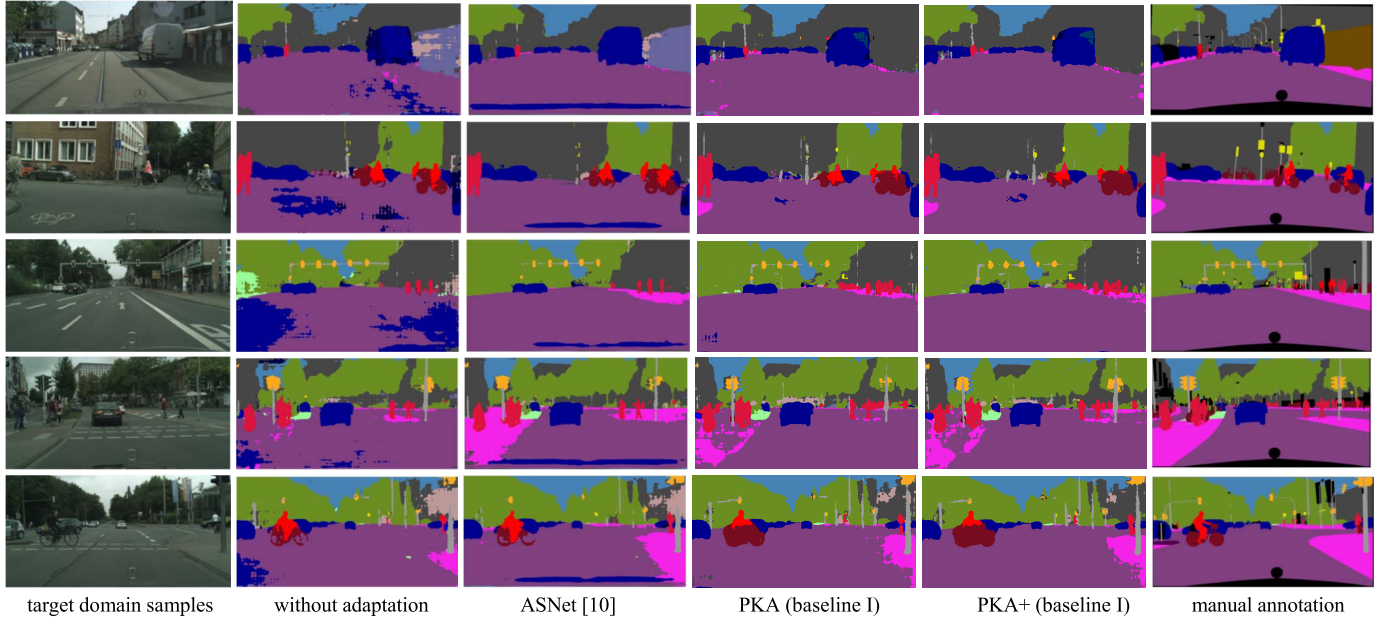


Fig. 5. Qualitative results obtained for semantic segmentation without and with domain adaptation. The third column is the adaptation results shown in ASNet. The fourth and fifth columns are adaptation results obtained with the proposed PKA and PKA+. The last column is the ground truth manual annotations.

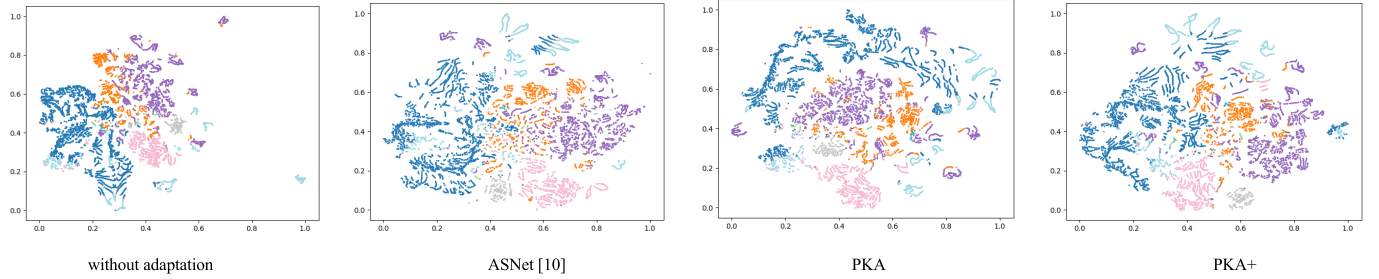


Fig. 6. Feature clusters visualization in the embedding space by t-SNE [57].

MRENT [42] on mean IoU while being more efficient on some classes, such as “building,” “sky,” and “car.”

F. Ablation Studies

Experimental ablation studies are also conducted in the task “GTA5 to Cityscapes” to investigate the efficacy of the proposed framework.

1) *Learning of the Hyper-Parameter T* : Parameter T in the formulation (16) is used to split the soft-max output of target domain into the under-aligned regions and the aligned regions. Setting an appropriate value of T not only involves sufficient aligned pixels for semi-supervised learning, but also alleviates the over-adaptation problem in the adversarial adaptation process. We tested PKA+ on both baselines by altering T over the range [0.7, 1]. As illustrated in Fig. 7, the adaptation performance is affected as T varies. The best performance is achieved when T is around 0.92. When $T = 1$, PKA+ degrades to a single adversarial adaptation strategy.

2) *Evaluation of the Efficacy of Each Component*: Without placing any constraints on the baseline networks and introducing additional parameters, PKA can facilitate the development of output-space based DA works with soft adversarial

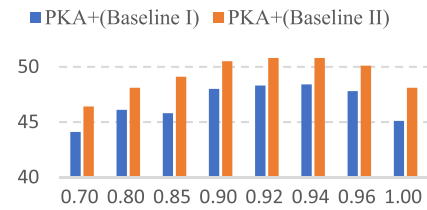


Fig. 7. Performance (on mIoU) of PKA+ while altering the hyperparameter T .

adaptation loss and lead to efficient adaptation improvements. Experimental results in Table III first show that the proposed PKA achieves a significant gain on mIoU from Baseline I and Baseline II, respectively. Second, a domain confidence (DC) based refinement strategy is proposed for PKA to improve the adaptation performance by refining prototypical knowledge. When combining with DC (shown as PKA+ in Table III), the PKA’s mIoU elevates from 47.1% to 48.4% on Baseline I, and from 49.2% to 50.8% on Baseline II, which demonstrates the fundamental efficacy on the refinement component. Furthermore, given that discrimination confidence (DC) is able to measure the intra-domain discrepancy, we experimented

TABLE II
EXPERIMENTAL RESULTS OF THE ADAPTATION TASK “SYNTIA TO CITYSCAPES” UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE

-SYNTIA to Cityscapes-		road	side.	build.	light	sign	vege.	sky	person	rider	car	bus	moto.	bike	mIoU
Without adaptation		55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
Adversarial Domain Adaptation															
feature space	ASNet (fea.) [10]	62.4	21.9	76.3	11.7	11.4	75.3	80.9	53.7	18.5	59.7	13.7	20.6	24.0	40.8
	SIBAN [29]	82.5	24.0	79.4	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	46.3
	SSF-DAN [14]	84.6	41.7	80.8	11.5	14.7	80.8	85.3	57.5	21.6	82.0	36.0	19.3	34.5	50.0
output space	ASNet [10]	84.3	42.7	77.5	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
	CLAN [11]	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.1	73.0	32.9	22.6	30.7	47.8
	AdvEnt [12]	87.0	44.1	79.7	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	47.6
	AdvEnt+Minent [12]	85.6	42.2	79.7	8.7	5.4	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	48.0
	ASA [13]	91.2	48.5	80.4	5.5	5.2	79.5	83.6	56.4	21.0	80.3	36.2	20.0	32.9	49.3
	MRNet [53]	82.0	36.5	80.4	18.0	13.4	81.1	80.8	61.3	21.7	84.4	32.4	14.8	45.7	50.2
	IntraDA [54]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9
	PKA (Baseline I)	86.1	43.1	81.6	9.4	12.4	80.6	84.1	61.1	23.4	81.4	35.1	23.0	36.4	50.6
	PKA (Baseline II)	84.1	37.7	82.2	13.1	14.1	81.7	81.0	63.4	24.9	87.2	35.9	24.9	44.1	51.9
Non-adversarial Domain Adaptation															
Context Aware [55]		82.5	42.2	81.3	18.3	15.9	80.6	83.5	61.4	33.2	72.9	39.3	26.6	43.9	52.4
PIT [56]		83.1	27.6	81.5	26.4	33.8	76.4	78.8	64.2	27.6	79.6	31.2	31.0	31.3	51.8
MaxSquare [47]		82.9	40.7	80.3	12.8	18.2	82.5	82.2	53.1	18.0	79.0	31.4	10.4	35.6	48.2
CBST [42]		68.0	29.9	76.3	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	48.9
CRST (MRENT) [42]		69.6	32.6	75.8	23.3	29.5	77.7	78.9	60.0	28.5	81.5	25.9	19.6	41.8	49.6
PKA+ (Baseline I)		86.4	42.0	84.4	12.4	14.3	81.1	84.9	63.1	21.4	85.8	37.3	22.4	39.7	51.9
PKA+ (Baseline II)		86.8	40.3	83.3	21.6	18.4	83.2	84.2	61.0	25.6	85.7	34.5	20.4	44.3	53.0

TABLE III
ABLATION STUDIES. FDA REFERS TO FEATURE SPACE BASED ADVERSARIAL ADAPTATION [7]. CGAN REFERS TO *Cycle-GAN* [52]. PK AND DC ARE SHORTED FROM PROTOTYPICAL KNOWLEDGE AND DISCRIMINATION CONFIDENCE

Method	Vanilla DA			Components			mIoU
	ASNet	MRNet	FDA	PK	DC	CGAN	
Baseline I [10]	✓						41.4
PKA	✓			✓			47.1
PKA+	✓			✓	✓		48.4
Baseline II [53]		✓					45.5
PKA		✓		✓			49.2
PKA+		✓		✓	✓		50.8
Baseline with DC	✓				✓		43.3
PKA with FDA [7]	✓		✓	✓			47.5
PKA+ with CGAN[52]		✓		✓	✓	✓	51.2

a separate adaptation approach, **Baseline with DC**, that utilizes the discrimination confidence as soft domain labels and operates the same adaptation strategy as **PKA**. As shown in Table III, **Baseline with DC** provides a mere improvement of 1.9% on mean IoU over Baseline I [10], and shows an inferior performance with respect to **PKA**(Baseline with prototypical knowledge (PK)). It indicates that discrimination confidence interpreted on the feature space is less reliable to estimate soft domain labels than prototypical knowledge.

3) *Combination With Other DA Methods*: We further explored the efficacy of the proposed method when combining with recent DA methods. First, given that [7]–[9] have demonstrated the effectiveness of DA on the feature space, we modified the D' from a unilateral training discriminator

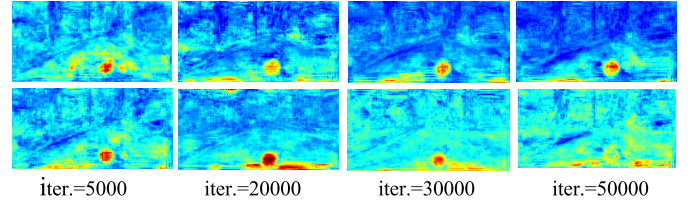


Fig. 8. Discrimination confidence visualization at different iteration steps (Unilateral discriminator (first row) vs. Conventional adversarial discriminator (second row)).

to a general adversarial discriminator for feature-space based domain adaptation [7] (**FDA**), to evaluate the superiority of the two kinds of discriminator when combined with **PKA**. Results in Table III show that, although **PKA with FDA** provides a slight improvement over the sole **PKA** model, it underperforms the proposed **PKA +** (PKA with DC) method. In Fig. 8, we visualize discrimination confidence maps encoded by the unilateral discriminator and the adversarial discriminator at different iterations. It can be observed that the adversarial discriminator tends to be indiscriminate when the generator produces domain-invariant features with increasing adversarial training iterations number. However, the proposed unilateral discriminator is trained independently without interactions with the generator G , thereby being able to remain discriminative on domain discrepancy information. Lastly, we assembled the **PKA+** with an image-space based DA method [8], which transferred the GTA5 dataset into a Cityscapes-style dataset using *Cycle-GAN* [52]. With this strategy, the results in Table III show a further 0.4% improvement in the mean IoU over **PKA+** alone.

4) *Choice of the Entropy Regularization Term*: Chen *et al.* [47] propose Maximum Squares Loss as a variation of negative

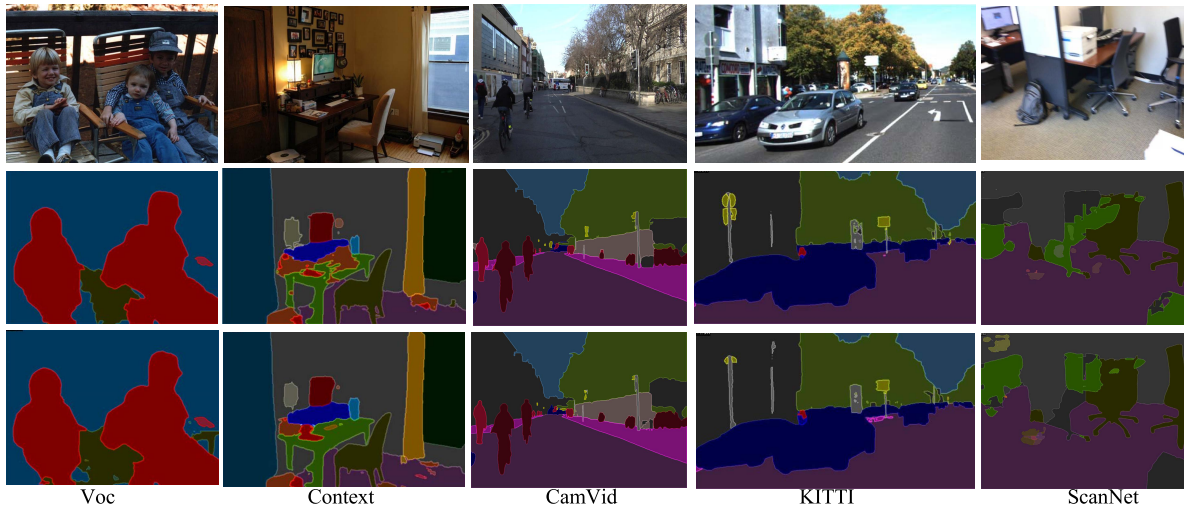


Fig. 9. PKA in real-world domain shifts. From top to bottom: target domain samples, segmentations without domain adaptation, adaptation results obtained with the proposed PKA.

TABLE IV

EXPERIMENTAL COMPARISON BETWEEN ENTROPY MINIMIZATION LOSS ($H(x)$) AND MAXIMUM SQUARES LOSS ($MS(x)$). *Vanilla Model* REFERS TO PKA MODELS TRAINED WITHOUT USING $H(x)$ NOR $MS(x)$

GTA5 to Cityscapes		
	PKA (Baseline I)	PKA (Baseline II)
<i>Vanilla model</i>	46.0%	48.4%
$H(x)$	47.1%	49.2%
$MS(x)$	46.9%	49.3%

entropy to achieve class-balanced cluster alignments. We also experiment PKA with the Maximum Squares term (defined in (25)) in our experiments.

$$H^{(w,h)} = -\frac{1}{2} \sum_{l=1}^L (\rho_l^{(w,h)})^2 \quad (25)$$

Experimental results with two types of entropy minimization terms are summarized in Table IV. It can be observed that the effectiveness of the maximum squares loss is on par with that of the negative entropy loss over prototypical knowledge regularization. Therefore, we just use the common negative entropy in all our models.

G. Domain Adaptation Experiments on Practical Scenarios

In this section, we further explore the potential applications of the proposed prototypical knowledge adaptation in more practical and general scenarios. With experimental details, we discuss the effectiveness and generalization ability of the proposed framework.

1) *PKA for Real-World Domains*: In this part, we exploit PKA on multiple real-world domains with more challenging settings. *MSeg* [58] is a new cross-domain semantic segmentation composite dataset that unifies a variety of real-world datasets from different scenes (“Everyday objects,” “Driving,” and “Indoor”) at a unified taxonomy of 194 categories. It contains 190,231 training images and 12,561 validation images

with flexible image resolutions, providing various realistic cross-domain shifts for real-world domain adaptation and generalization applications. Its elaborated taxonomy allows PKA to learn a universal segmentation model over multiple domains and adapt to the testing split (target domain) without using labels from that target domain. In our experiments, the training split [58] of *MSeg* with ground truth labels are set as the source domain. We revisit all training sets of the testing split of *MSeg* to build the target domain. Note that we do not use any images from the testing split of *MSeg* for training, e.g., the testing dataset of *CamVid* still is used for evaluation and the training dataset of *CamVid* is used in the target domain. In implementations, we follow [58] to train a HRNet-W48 [59] segmentation network and construct intra-scene DA pipelines with the proposed PKA method. Experimental results are summarized in Table V. The PKA method outperforms the source-only method on all target datasets with slight variations in different scenes. Meanwhile, qualitative results visualized in Fig 9 show that even though domain shift is less evident in the real-to-real scenarios, PKA achieves noticeable improvements on specific classes in the segmentation results.

PKA demonstrates its independent effectiveness while dealing with the domain shift across the large-scale label space. On the one side, PKA can readily combine with elaborated FCNs, e.g., HRNet-W48 [59], to aggregate semantic knowledge over a diversity of source domains. On the other side, although multiple adaptation scenes are involved over the large-scale label space, e.g., three distinct test scenes are predefined over 194 categories in *MSeg* [58], PKA allows to transfer semantics across domains by utilizing prototypical knowledge from the unlabeled target scenes without placing any constraints, thereby aligning intra-scene domain shifts and obtains better performance than the method without domain adaptation.

2) *PKA for Cross-City Test Case*: *Oxford RobotCar* [60] is a real-world street dataset containing 895 training images and 271 validation images collected in rainy scenes, providing another practical test case for our method. We follow

TABLE V

ADAPTATION RESULTS OVER THE *MSeg* DATASET UNDER THE METRICS MEAN IOU ON EACH DATASET AND THE HARMONIC MEAN OVER ALL DATASETS (THE TESTING SPLIT). WE DO NOT INCLUDE *WildDash* IN THE TARGET DOMAIN SINCE IT DOES NOT PROVIDE A TRAINING DATASET. THE SOURCE-ONLY MODEL IS TRAINED WHILE FULLY DEPENDING ON THE SOURCE DOMAIN DATASETS. BOTH THE SOURCE-ONLY MODEL AND PKA ARE TRAINED UNDER UNIFIED TAXONOMY AND EVALUATED ON TEST DATASET TAXONOMY AFTER LABEL MAPPING [57]. BALANCE COEFFICIENT τ_l IN EQ. (4) IS SET AS 1 DURING TRAINING

Method	Backbone	Everyday objects		Driving		Indoor	<i>h. mean</i>
		VOC	Context	CamVid	KITTI	ScanNet	
Source-only	<i>HRNet-W48</i>	66.1	40.3	83.3	61.5	44.8	55.2
PKA		68.0	41.4	84.9	61.7	46.2	56.5

TABLE VI

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK “CITYSCAPES TO OXFORD ROBOTCAR” UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE

Method	road	sidewalk	building	light	sign	sky	person	automobile two-wheel	mIoU
Source-only	79.2	49.3	73.1	55.6	37.3	36.1	54.0	81.3	49.7
PatchAlign [61]	94.4	63.5	82.0	61.3	36.0	76.4	61.0	86.5	58.6
MRNet [53]	95.9	73.5	86.2	69.3	31.9	87.3	57.9	88.8	61.5
PKA	93.8	77.9	86.1	70.2	40.3	89.1	56.0	91.1	60.4

TABLE VII

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK ON VIDEO SEMANTIC SEGMENTATION. WE SET THE 1ST FRAME AS OUR INITIAL KEY FRAME WITH A DURATION LENGTH OF 5, AND MEASURE mIoU ON THE 20TH FRAMES

Backbone	Per-frame approaches				Key-frame approach	
	<i>Resnet-101</i>		<i>MobileNet-V2</i>		<i>Resnet-101-FlowNetV2</i>	
Metric	mIoU	runtime	mIoU	runtime	mIoU	runtime
Source-only	36.6		35.2		36.5	
PKA	47.1	1.5 fps	44.5	20.7 fps	46.9	10.5 fps

Tsai *et al.* [61] to conduct the domain adaptation experiment: “Cityscapes to Oxford RobotCar” with PKA (Baseline II). Experimental results in terms of mean IoU are summarized in Table VI. It can be observed that the PKA method surpasses the source-only model by a clear margin, demonstrating the efficacy of our method on this different cross-domain scenario. When compared to state-of-the-art methods, PKA performs more efficiently on both general and class-specific adaptation improvements.

3) *PKA for Video Semantic Segmentation*: We further show that the proposed PKA can facilitate video segmentation scenario by being free from using any frame annotations. Due to the exceptional cost of frame-wise labelling for video sequences, video semantic segmentation models are generally trained on sparsely-labeled benchmark datasets in a semi-supervised manner [61]–[66]. To deploy PKA for video domain adaptation, the whole image set of *GTA5* [23] is taken as the source image domain. *Cityscapes* [22] contains 2,975 and 500 snippets for training and validation. Each snippet has 30 frames, of which only the 20th frame is annotated. We randomly extract one frame from each snippet to form the target video domain.

Based on the proposed PKA framework, we propose two schemes to develop unsupervised video domain adaptation. First, we propose a simple but efficient approach, per-frame inference, via combining PKA with a compact FCN backbone. In detail, we build PKA on a MobileNet-V2 [67] backbone, to train the **PKA-MobileNet** model over the source

domain and the target domain. During the inference phase, we directly deploy the trained **PKA-MobileNet** model on each frame of the validation video set of Cityscapes. We observe the unsupervised video segmentation performance in terms of mean IoU and the average of frames per second (fps). Experimental results are summarized in Table VII. Second, we propose a key-frame based solution, **PKA-FlowNet**, to reduce redundant computation among neighboring frames by using frame-to-frame optical flow estimate. It utilizes a strong image segmentation model at keyframes but propagates the semantic features of keyframes to non-key ones. It is built upon the observation that available temporal correlations of high-level semantic concepts exist on consecutive frames [62]–[64]. To train such a key-frame based model, we reuse the off-the-shelf ResNet-101 model (yielded from the “GTA5 to Cityscapes” task) as the semantic backbone on key frames and deploy a pretrained optical flow network, FlowNet V2 [68], to encode the temporal consistency across frames, and then propagate semantic patterns from the temporal consistency to non-key frames by using a wrap operation [62]. Even though the absence of labels in the target domain, pseudo labels distilled from prototypical knowledge are utilized to train FlowNet V2 to obtain temporal correlations. During inference, we are able to deploy the **PKA-FlowNet** model for instant video sequences segmentation.

Experimental results with the proposed approaches are reported in Table VII. First, when directly applying a trained Resnet-101 model onto each frame of the target domain (the

validation video set of *Cityscapes* [22]), both the source-only and the PKA method show slow inference runtime (1.5 fps). However, when trained with MobileNet, the source-only model is able to run faster with 20.7 fps. PKA further adapts MobileNet to the target domain and achieves higher accuracy (44.5% on mIoU) by learning prototypical knowledge from the target domain. Lastly, the key-frame based approach enables PKA to trade-off between inference accuracy and runtime. With estimated optical flow, the key-frame based model eventually achieves 46.9% of mIoU with 10.5 fps.

VI. CONCLUSION

In this work, we first propose a prototypical knowledge adaptation strategy by utilizing a soft adversarial loss to regularize the output-space adaptation process at a finer detail level. It realizes joint distribution alignment on the output space without introducing additional parameters beyond that of the conventional methods. In order to further improve the adaptation performance, we secondly propose a prototypical knowledge refinement strategy by introducing a unilateral discriminator in the proposed adaptation framework.

The theoretical analysis, experimental results on two baselines and comparison of performance against alternative DA strategies demonstrate that the proposed method is effective for adaptation improvements and is comparable to both adversarial and non-adversarial state-of-the-art methods. Besides, the proposed method is also evaluated under practical test settings with detailed experiment results on two real-world cross-domain scenarios. Future work will involve defining more efficient domain discrepancy metrics and regularization schemes to continue improve the performance of DA on cross-domain scenarios.

REFERENCES

- [1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [4] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
- [5] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," 2015, *arXiv:1505.01257*.
- [6] B. Gong, F. Sha, and K. Grauman, "Overcoming dataset bias: An unsupervised domain adaptation approach," in *Proc. NIPS Workshop Large Scale Visual Recognit. Retr. (LSVRR)*, vol. 3, 2012, pp. 1–5.
- [7] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*. [Online]. Available: <https://arxiv.org/abs/1612.02649>
- [8] J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1989–1998.
- [9] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C.-F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–8.
- [10] Y.-H. Tsai, W.-C. Hung, S. Schuster, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.
- [11] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.
- [12] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.
- [13] W. Zhou, Y. Wang, J. Chu, J. Yang, X. Bai, and Y. Xu, "Affinity space adaptation for semantic segmentation across domains," *IEEE Trans. Image Process.*, vol. 30, pp. 2549–2561, 2021.
- [14] L. Du *et al.*, "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.
- [15] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 708–721.
- [16] J. Tighe and S. Lazebnik, "Superpixels: Scalable nonparametric image parsing with superpixels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 352–365.
- [17] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2008, pp. 44–57.
- [18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [19] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.
- [20] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
- [21] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kotschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 22–29.
- [22] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
- [23] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 102–118.
- [24] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 102–118.
- [25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 213–226.
- [26] B. Sun and K. Saenko, "From virtual to reality: Fast adaptation of virtual object detectors to real domains," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 3.
- [27] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 797–809, Apr. 2014.
- [28] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.
- [29] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6778–6787.
- [30] Y. Chen, W. Li, and L. V. Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7892–7901.
- [31] Z. Wu *et al.*, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 535–552.
- [32] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1335–1344.
- [33] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," 2017, *arXiv:1711.01575*.
- [34] H. Huang, Q. Huang, and P. Krahenbuhl, "Domain transfer through deep activation matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 611–626.

- [35] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.
- [36] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.
- [37] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin, "Penalizing top performers: Conservative loss for semantic segmentation adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 587–603.
- [38] Y. Zhang, M. Ye, Y. Gan, and W. Zhang, "Knowledge based domain adaptation for semantic segmentation," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105444.
- [39] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, "Simplified unsupervised image translation for semantic segmentation adaptation," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107343.
- [40] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.
- [41] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 289–305.
- [42] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5982–5991.
- [43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, p. 27.
- [44] R. Shu, H. Bui, H. Narui, and S. Ermon, "A DIRT-T approach to unsupervised domain adaptation," in *Proc. Int. Conf. Learn. Represent. (ICRL)*, Feb. 2018, pp. 1–19.
- [45] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8050–8058.
- [46] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2239–2247.
- [47] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.
- [48] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 91–100.
- [49] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1823–1841, Aug. 2020.
- [50] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. NIPS*, 2007, p. 137.
- [51] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.
- [52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [53] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization *in vivo*," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2–9.
- [54] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3764–3773.
- [55] J. Yang, W. An, C. Yan, P. Zhao, and J. Huang, "Context-aware domain adaptation in semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 514–524.
- [56] F. Lv, T. Liang, X. Chen, and G. Lin, "Cross-domain semantic segmentation via domain-invariant interactive relation transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4334–4343.
- [57] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 1, pp. 2579–2605, 2008.
- [58] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "MSeg: A composite dataset for multi-domain semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2879–2888.
- [59] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.
- [60] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6556–6565.
- [61] Y.-H. Tsai, K. Sohn, S. Schuster, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.
- [62] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2349–2358.
- [63] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 352–368.
- [64] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8818–8827.
- [65] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5997–6005.
- [66] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8866–8875.
- [67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [68] F. Reda, R. Pottor, J. Barker, B. Catanzaro. (2017). *FlowNet2-Pytorch: Pytorch Implementation of FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks*. [Online]. Available: <https://github.com/NVIDIA/flowNet2-pytorch>

3.3 Target-dependent classifier adaptation

3.3.1 Summary

This section introduces a new perspective for **ESDA**, that aims at learning a target-dependent classifier for semantic segmentation in domain adaptation. It builds upon the observation that traditional domain adaptation can be efficiently uprooted when disjoint domain manifolds exist across domains where the joint-domain adaptation hypothesis could be biased to the source domain, thereby violating the cluster assumption in the target domain. By enclosing the cluster violation term, this research realizes loose domain alignment but a tight bound for the adaptation to the target domain. **ESDA** in this context exploits a target-dependent classifier to efficiently adapt to the target domain clusters even given a certain degree of domain misalignment.

In its implementation, the method proposes an unsupervised fine-tuning strategy, which takes the conventional DA model (joint-domain classifier) as an initialization and progressively pushes the classification boundaries away from the domain misalignment regions, to fine-tune a target-dependent classifier in accordance with the cluster assumption in the target domain. The proposed learning method is general to conventional domain adaptation models such that it may recycle the pre-trained conventional DA models and learn target-dependent classifiers with efficiency.

3.3.2 Original contributions

The main contributions of **ESDA** related to this work are:

- A target-dependent classifier for cross-domain semantic segmentation able to better fit with the target domain clusters even under imperfect domain alignment;
- An innovative linkage of meta-learning with domain adaptation in semantic segmentation;

- A framework that offers a general solution for a diversity of DA models. Without placing any constraints on the conventional DA framework involved, this strategy can readily be applied to a range of vanilla DA models and lead to improvement in adaptation.

3.3.3 Article

This research has been reported in “Learning a Target-Dependent Classifier for Cross-domain Semantic Segmentation: Fine-tuning versus Meta-learning” (H. Tian, S. Qu, P. Payeur), which is published at *Elsevier Pattern Recognition*, 2023 [104]. The online version can be found at <https://www.sciencedirect.com/science/article/abs/pii/S0031320323007884>. The original paper is provided below as published. Permission to reprint the paper in this thesis was obtained from the publisher Elsevier Ltd.



Learning a target-dependent classifier for cross-domain semantic segmentation: Fine-tuning versus meta-learning

Haitao Tian^{a,b,*}, Shiru Qu^a, Pierre Payeur^b

^a School of Automation, Northwestern Polytechnical University, Xi'an, China

^b School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Keywords:

Domain adaptation
Semantic segmentation
Unsupervised learning
Meta-learning

ABSTRACT

Recently proposed domain adaptation arts have dominated the field of cross-domain semantic segmentation by operating domain manifolds alignment and learning an optimal joint hypothesis (joint-domain classifier) for both source and target domains. However, a joint-domain classifier can still violate the cluster assumption in the target domain in case domain manifolds are not fully aligned after domain adaptation. In this work, we raise the intractability of perfect domain alignment and turn to exploit a novel hypothesis: a target-dependent classifier, to efficiently adapt to the target domain clusters even given a certain degree of domain misalignment. Specifically, we first propose an unsupervised fine-tuning strategy, which optimizes the joint hypothesis of vanilla domain adaptation into a target-dependent hypothesis to better fit with the target domain clusters. Second, we connect the “learning to learn” concept of meta-learning with pixel-wise domain adaptation, which serves as a reliable hypothesis initialization, providing an alternative solution to learning a more generalized target-dependent classifier. The proposed learning method is general to conventional domain adaptation models. In experiments, we recycle the pre-trained conventional DA models and learn target-dependent classifiers with the proposed method. Experimental results on synthetic-to-real adaptation and cross-city adaptation benchmarks demonstrate that the target-dependent classifier leads over state-of-the-art performance.

1. Introduction

Cross-domain semantic segmentation aims to learn transferable semantic patterns in changing scenarios for pixel-wise prediction tasks, playing a pivotal role in real-world applications such as autonomous driving and robotic navigation. In a deep neural network setting, this task tends to be intractable once pixel-wise labeled data is absent in a new domain where supervised fine-tuning is impossible. However, with the goal of dealing with covariate shift problem from domain divergence [1], the domain adaptation (DA) technique [2] tackles this problem by adapting segmentation network learnt on a source domain (source-dependent classifier) into a target domain without using any labeled data from that target domain. The foundation of domain adaptation lies in the assumption that if domain manifolds can be well aligned, a single classifier can operate well in the source domain, and also achieve low generalization error in the target domain [3].

Limitations on conventional DA. A family of domain adaptation works have been proposed for tackling cross-domain semantic segmentation problems, which can be grouped into two streams: the

adversarial adaptation strategy [2,4–10] and the self-supervised adaptation strategy [11,12]. Both streams of DA attempt to apply an optimal joint hypothesis (*joint-domain* classifier, see an illustration in Fig. 1) across domains after domain alignment. However, (i) for adversarial adaptation, aligning domain manifolds in an embedding space is solved by Jensen-Shannon divergence minimization [13], while if the distributions of two domains are disjoint, or have supports that lie on low dimensional manifolds, output probability measures of overlapped regions will be invalid [14–16], such that the *joint-domain* classifier will not converge on a reliable hypothesis [17]. Conversely, (ii) for self-supervised adaptation, pseudo-labeling [11] the target domain is equivalent to constructing a new labeled dataset for supervised learning of the target domain. But confident pseudo labels from the target domain would still contain a portion of noisy labels. The dominance of source domain ground truth will entangle the support of the joint hypothesis to overfit around the source domain. The *joint-domain* classifier thus hurts classification performance in the target domain.

The work in this paper is built upon the observation that the mainstream of state-of-the-art DA works suffer from imperfect domain

* Corresponding author at: School of Automation, Northwestern Polytechnical University, Xi'an, China.
E-mail address: htian026@uottawa.ca (H. Tian).

<https://doi.org/10.1016/j.patcog.2023.110091>

Received 18 May 2023; Received in revised form 10 October 2023; Accepted 29 October 2023

Available online 4 November 2023

0031-3203/© 2023 Elsevier Ltd. All rights reserved.

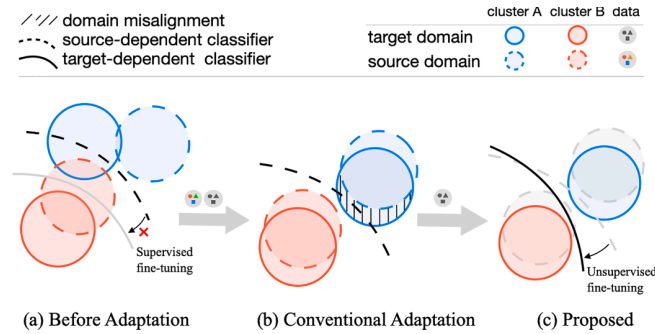


Fig. 1. Illustration of conventional domain adaptation and the proposed method. Before domain adaptation (a): available source labels allow to learn a classifier with supervision on the source domain (source-dependent classifier). However, given the absence of target labels, supervised fine-tuning for the target domain is intractable. Conventional adaptation methods (b) aim to adapt the source-dependent classifier to a joint hypothesis which is expected to work for both domains (*joint-domain* classifier). However, under imperfect domain alignment from the conventional methods, the *joint-domain* classifier could be biased by misaligned regions on the target domain. Conversely, the proposed method (c) aims to seek a *target-dependent* classifier that better generalizes to the target domain even given the existence of a certain degree of domain misalignment.

alignment, which causes the resulting *joint-domain* classifier to corrupt adaptation performance in the target domain. Rather than further complicating conventional DA paradigms, we present a general solution for semantic segmentation domain adaptation: learning a *target domain dependent* classifier (as shown in Fig. 1). Learning a target domain dependent classifier allows to optimize conventional DA to adapt to a specific hypothesis depending upon target statistics that will better generalize to the target domain even under a certain degree of domain misalignment.

Specifically, we propose **Pixel-wise Conditional Entropy Minimization (PCEM)**, a strategy that seeks to minimize the cluster assumption violation on the target domain. First, *PCEM* takes the conventional DA model (*joint-domain* classifier) as an initialization and progressively pushes the classification boundaries away from the domain misalignment regions, to fine-tune a target-dependent classifier in accordance with cluster assumption in the target domain. Second, we connect the “learning to learn” concept of meta-learning [18,19] to *PCEM*, as an alternative approach for learning a generalized target-dependent classifier. Meta-learning provides pivotal competency for learning reliable network initializations [20–22]. The proposed combination of *PCEM* and meta-learning provides a reliable initial hypothesis, meta-initialization, for the target-dependent classifier development. In experimental results, it is shown that meta-learning facilitates further adaptation improvement.

The main contributions of this work are: (i) the introduction of a *target-dependent* classifier for cross-domain semantic segmentation able to better fit with the target domain clusters even under imperfect domain alignment; (ii) an innovative linkage of meta-learning with domain adaptation in semantic segmentation; and (iii) a framework that offers a general solution for a diversity of DA models. Without placing any constraints on the conventional DA framework involved, this strategy can readily be applied to a range of vanilla DA models and lead to improvement in adaptation.

2. Related work

2.1. Domain adaptation for semantic segmentation

(1) **Adversarial Adaptation:** Generative adversarial network [13] has a pivotal competency for eliminating domain discrepancy among

distribution encodings and learning domain-agnostic models. Image-space based adaptation [23–31] considers each pixel of the input image as an adapting instance and eliminates domain shift by generating cross-domain similar looking images. Feature-space based adaptation [2,5,29,32–35] was proposed to tackle domain discrepancy on the feature space. However, aligning the feature encodings does not always guarantee classification accuracy on the output space, especially when probability measures of distributions are disconnected. A family of output-space based adaptation [6–9,36] was then proposed to address this issue by developing structural domain alignment on the output space. However, the improvement on adversarial adaptation remains limited.

(2) **Self-Supervised and Cluster-Based Adaptation:** One alternative to adversarial adaptation is self-supervised learning. It builds upon the observation that a source-dependent classifier allows to construct highly confident pseudo labels by a thresholding strategy and realizes supervised learning in the target domain. Zou et al. proposed a class-balanced learning strategy in [11] and then developed a confidence regularized learning strategy in [12]. Recently, the self-supervised learning strategy is more widely integrated in general domain adaptation paradigms as a complementary component to enhance domain adaptation performance [10,27,28,37]. The cluster assumption has been well aggregated in a stream of DA works [8,38,39] to maintain the target domain’s clusters separable during DA progress and complement the generic domain adaptation. Vu et al. [8] utilized entropy minimization and adversarial adaptation to encourage unambiguous cluster assignments. The approach was evolved by Chen et al. [39] who introduced a maximum squares loss based on the cluster assumption.

(3) **Recent Improvements:** Despite the formulation of a wide range of domain adaptation approaches for semantic segmentation, the state-of-the-art performance of domain adaptation remains limited in recently reported works. Therefore, a stream of methods [16,40–43] were proposed to improve the efficacy of conventional domain adaptation. First, aiming to leverage a more efficient metric to deal with the probability measures issue existing in generic adversarial loss, Lee et al. [16] developed a reliable adaptation loss upon Wasserstein distance [44] to replace the genetic binary cross-entropy loss. Second, Wang et al. [40] dealt with the instability in adversarial adaptation by applying a L1 loss to minimize the distance between the target domain features with the most similar source domain counterparts. Besides, our work is also related to an adversarial intra-domain adaptation strategy (IntraDA) [43]. It is built on the *joint-domain* classifier from Advent [8] and conducts a secondary adaptation optimization on the target domain to yield better generalization. We also present a comparison with IntraDA in Section 4.

2.2. Non-conservative domain adaptation

Recent research on domain adaptation also identifies possible limitations on conventional domain adaptation and aims to propose non-conservative domain adaptation. Shu et al. suggest that training a classifier which generalizes well on the source domain can hurt performance on the target domain even with domain adaptation [17]. Recent advances in Transformer [45] allow to use a different network architecture [46] for domain adaptation in semantic segmentation. In order to utilize more available well-labeled homogeneous domains, multi-source domain adaptation was proposed [47]. Liu et al. believe that the traditional pair-wise domain adaptation has the limitation to apply to a novel domain in practice and an open compound domain adaptation strategy is proposed [48].

2.3. Meta-learning

Meta-learning recently attracted major attention because of its competency of “learning to learn” on reliable network initialization [20–22] and network optimizer [49–51], as well as for network

architecture search [52,53].

The second emphasis in this paper is related to recently proposed meta-learning paradigms. MAML [20] and Reptile [21] are devoted to learning reliable weights initialization for few-shot learning network that can lead the network to adapt to new tasks in a few gradient steps. MLDG [19] then extends this strategy for domain generalization from multi-source domains to an unseen domain. Online Meta-learning [54] is the first work applying meta-learning in domain adaptation, which is proposed to tackle the computationally heavy graph problem involved in meta-learning. Guo et al. [55] proposes a domain-aware meta-learning strategy for domain adaptation to sample reliable source pixels to facilitate self-supervised adaptation progress with pseudo labels.

3. Methods

In this section, we begin with a formal description of the target-dependent classifier for semantic segmentation. We then detail our proposed solutions to learn a target-dependent classifier by an unsupervised fine-tuning strategy, and a meta-learning based strategy, respectively.

3.1. Target-dependent classifier for semantic segmentation

The cross-domain semantic segmentation problem considers a dataset composed of: a labeled source domain $\mathcal{S} = \{X_s, Y_s\}$ and a unlabeled target domain $\mathcal{T} = \{X_t\}$, where X_s is an image sample in the source domain; Y_s is a corresponding pixel-wise annotation label pre-defined in L categories; X_t is an image sample in the target domain. We use x_s and x_t to denote a random pixel sample on X_s and X_t , respectively. y_s denotes the corresponding class label of x_s .

With research works providing various DA paradigms and variations, state-of-the-art DA research [8,10,12,27,40] relies on a conventional domain adaptation assumption that a fully convolutional network (FCN) based classifier trained on a source domain will transpose easily into a target domain if the adapted distribution encodings are domain-agnostic. This condition can be characterized by finding a *joint-domain* classifier (hypothesis h parameterized by θ^*) that has low generalization error on both the source and target domains, formulated as:

$$h_{\theta^*} = \underset{h_{\theta} \in H}{\operatorname{argmin}} [\operatorname{err}_{\mathcal{S}}(h_{\theta}(X_s)) + \operatorname{err}_{\mathcal{T}}(h_{\theta}(X_t))], \quad (1)$$

where $\operatorname{err}_{\mathcal{S}}(\cdot)$ and $\operatorname{err}_{\mathcal{T}}(\cdot)$ are corresponding generalization error functions for the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$ in the hypothesis space H . However, solving a *joint-domain* hypothesis does not guarantee a perfect domain distribution alignment due to the weak constraint on Jensen-Shannon divergence commonly used in conventional DA methods [14–16]. If the domain clusters offer a certain degree of misalignment after domain adaptation, the learnt classifier h_{θ^*} would corrupt the adaptation performance on the target domain, a phenomenon referred to as the cluster assumption violation problem [17].

This work exploits a novel hypothesis: *target-dependent* classifier, h_{θ^*} . It aims to efficiently adapt the joint-domain classifier to better fit the target domain clusters even given a certain degree of domain cluster

violation, which can be formulated as:

$$h_{\theta^*} = \underset{h_{\theta} \in H}{\operatorname{argmin}} [\operatorname{err}_{\mathcal{S}}(h_{\theta}(X_s)) + v_{\mathcal{T}}(h_{\theta}(X_t))], \quad (2)$$

where $\operatorname{err}_{\mathcal{S}}(h_{\theta}(X_s))$ is the generalization error for the joint-domain classifier h_{θ} on $\mathcal{S}$, while $v_{\mathcal{T}}(h_{\theta}(X_t))$ measures the extent of cluster assumption violation involved by h_{θ} on the misalignment region on $\mathcal{T}$.

Novelty. The development of a target-dependent classifier hinges on the intractability of perfect domain alignment: disjoint domain manifolds exist across domains where the joint-domain adaptation hypothesis could be biased to the source domain, thereby violating the cluster assumption in the target domain. By enclosing the cluster violation term, $v_{\mathcal{T}}$, in Eq. (2), the proposed method realizes loose domain alignment but a tight bound for the adaptation to the target domain, as illustrated in Fig. 2.

Benefits. First, learning a target-dependent classifier promises better adaptation performance to the target domain. Clearly, while $v_{\mathcal{T}}$ in Eq. (2) is rigorously zero, the target-dependent classifier degrades into a normal joint-domain classifier. The proposed method seeks fine target-specific adaptation near the cluster violation areas in the target domain, thereby gaining adaptation improvements (illustrated as Fig. 2a compared to Fig. 2d and detailed as the PCEM strategy in the next section). Second, the formulation of Eq. (2) streamlines a compatible paradigm to conventional DA arts. On one side, Eq. (2) allows to incorporate existing well-adapted models as the network initial condition, h_{θ^*} , into the learning of the target-dependent classifier. On the other side, as the formulation of $v_{\mathcal{T}}$ does not introduce any complication in the network architecture, a future joint-domain adaptation design can be readily transposed into a target-dependent hypothesis by the sequential optimization on Eqs. (1) and (2).

3.2. Pixel-wise conditional entropy minimization with fine-tuning

In this section, we detail the Pixel-wise Conditional Entropy Minimization (PCEM) strategy, an unsupervised fine-tuning scheme that takes a conventional DA model as an initial condition and attempts to progressively fine-tune a target-dependent classifier (also see an illustration in Fig. 3). It will be formed of the following two stages.

- (1) **Initialization:** The first phase is driven by a conventional domain adaptation loss $\mathcal{L}_{DA}$:

$$\mathcal{L}_{DA}(x_s, y_s, x_t; \theta) = \mathcal{L}_{\text{seg}}[(h_{\theta}(x_s), y_s)] + \mathcal{L}_{\text{adp}}[(h_{\theta}(x_s), h_{\theta}(x_t))], \quad (3)$$

where $\mathcal{L}_{\text{seg}}$ is the typical supervised segmentation loss (pixel-wise cross entropy), and $\mathcal{L}_{\text{adp}}$ is the objective function of the adaptation component, which can be adapted to mainstream domain adaptation paradigms, such as adversarial adaptation loss [8] and self-learning loss [12]. A range of state-of-the-art domain adaptation models are considered in Section 4.2. The use of $\mathcal{L}_{DA}$ aims to form an optimal joint-domain hypothesis, h_{θ^*} , to provide an initial condition for the next fine-tuning phase:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} [\mathbb{E}_{\mathcal{S}, \mathcal{T}}(\mathcal{L}_{DA}(x_s, y_s, x_t; \theta))]. \quad (4)$$

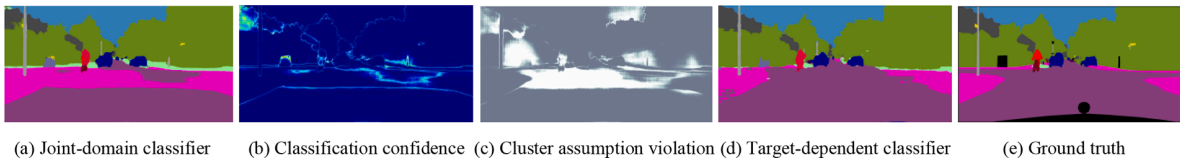


Fig. 2. Cluster assumption violation in the target domain. A state-of-the-art joint-domain classifier [27] obtains **compromising performance** on the target domain, that is (a) compared to (e), even though reaching **high confidence** (dark blue) on classification (b). This work aims to formulate the **cluster assumption violation** in the joint-domain classifier (c) and proposes a method to minimize the pixel-wise cluster assumption violation (d) (further detailed in Section 3.2) while learning a target-dependent classifier. Dark areas in (b) and bright intensity in (c) refer to high confidence value and high cluster violation, respectively.

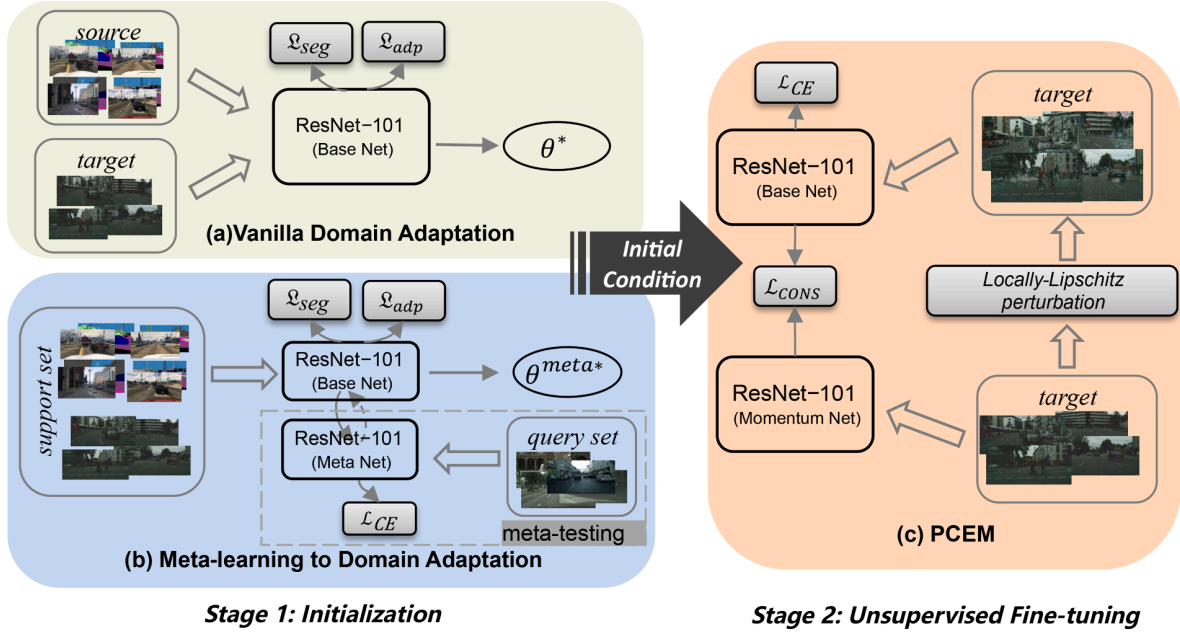


Fig. 3. Overview of the proposed two-phase strategy of learning a target-dependent classifier.

- (2) *Cluster Assumption Violation Formulation, Minimization and Fine-tuning*: We then learn a target-dependent classifier on the initial condition θ^* by minimizing cluster assumption violation in the target domain. It builds upon the observation that entropy minimization conditioned by a locally-Lipschitz constraint offers the effectiveness of iterative refinement on classifier boundaries that are in accordance with cluster assumption in the target domain [17]. In detail, *PCEM* first defines a *pixel-wise conditional entropy loss* $\mathcal{L}_{CE}$ on the target domain:

$$\mathcal{L}_{CE}(x_i; \theta^*) = - \sum_l \left[h_{\theta^*}(x_i)^{(l)} \log \left(h_{\theta^*}(x_i)^{(l)} \right) \right] + \mathcal{L}_{Lip}. \quad (5)$$

where $\mathcal{L}_{Lip}$ is the locally-Lipschitz constraint. We follow [17,56,57] for the implementation of $\mathcal{L}_{Lip}$ (see more details in Section 4.3, Section 4.8, and Fig. 8). The use of conditional entropy loss seeks to monitor the pixel-wise cluster assumption violation in the target domain (as illustrated in Fig. 2) and fine-tune the classification boundaries away from the domain misalignment regions by minimizing $\mathcal{L}_{CE}$.

Sibling Network. Aiming to formulate a pixel-wise adaptation strategy, *PCEM* secondly exploits the structural semantic adaptation information in the fine-tuning process. In particular, *PCEM* employs a *KL divergence based constraint* (D_{KL}) on the output space of the base network to regularize the adaptation process (see a conceptual view in Fig. 4):

$$\mathcal{L}_{CONS}(x_i; \theta^*, \bar{\theta}) = D_{KL}(h_{\theta^*}(x_i) \parallel h_{\bar{\theta}}(x_i)), \quad (6)$$

where $h_{\bar{\theta}}$ is a sibling network with the base network. The sibling network is not involved in network back-propagation but is updated by the momentum update:

$$\bar{\theta}_n = m\bar{\theta}_{n-1} + (1-m)\theta_n^*, \quad (7)$$

where m is the momentum set as 0.99 in this work, and $\bar{\theta}_0 = \theta^*$. The motivation for using the sibling network is to provide guidance for optimization on the base network. It forces the base network to stay close to the sibling network such that *PCEM* is able to incrementally fine-tune the classifier's decision boundaries within a local semantic-aware parameter neighborhood and push the classifier away from cluster assumption violated regions.

In general, driven by *PCEM*, the unsupervised fine-tuning optimization can be solved by:

$$\theta^T = \underset{\theta^*}{\operatorname{argmin}} \mathbb{E}_{\mathcal{T}} [\mathcal{L}_{CE}(x_i; \theta^*) + \alpha \mathcal{L}_{CONS}(x_i; \theta^*, \bar{\theta})], \quad (8)$$

where α is an hyperparameter to weight the KL divergence constraint. The influence of hyperparameter α on fine-tuning is examined in Section 4.7.

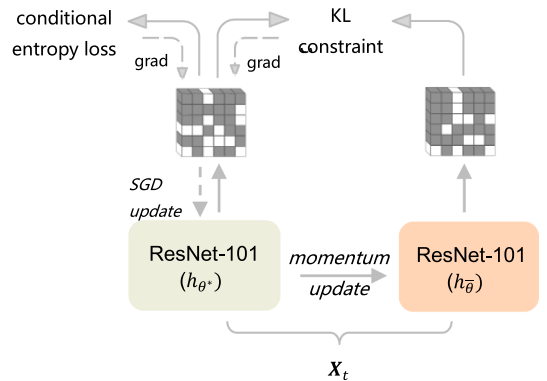


Fig. 4. A conceptual overview of the *PCEM* fine-tuning phase. The semantic-aware adaptation for semantic segmentation is realized by employing a KL based constraint over the base backbone (ResNet-101 (θ^*)) and the sibling backbone (ResNet-101 ($\bar{\theta}$)). *PCEM* is processed only in the target domain, such that the target-dependent classifier will not be entangled by source domain ground truth labels.

3.3. Pixel-wise conditional entropy minimization with meta-learning

A solution to implement *PCEM* on an initialized joint hypothesis of a conventional DA framework has been introduced in Section 3.2. However, we assume that the success of *PCEM* depends on a reliable initialization optimized from conventional DA. As discussed in Section 3.1, the dominance of source domain ground truth labels potentially forces the support of the joint hypothesis to overfit around the source domain, resulting in high cluster assumption violation in the target domain and thereby passing biased source domain knowledge to the initial condition. Learning from a source-biased initial condition, *PCEM* would likely fall into a suboptimal solution.

In this section, we propose a meta-learning solution to facilitate the proposed *PCEM*. Meta-learning has shown its efficacy in few-shot learning [20,49] and domain generalization [19]. It learns reliable network initializations upon which the network optimization is more likely to proceed to a global minimum. Connecting with domain adaptation, meta-learning provides episodic initializations during the DA process, named meta-initialization, upon which *PCEM* learns a more generalized target-dependent classifier (see an illustration in Fig. 3 (b)).

To establish meta-learning on conventional DA, we randomly sample a support set $Sup = \{X_s^{SUP}, Y_s^{SUP}, X_t^{SUP}\}$ and a query set $Que = \{X_t^{QUE}\}$ from the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$, where $Sup \cap Que = \emptyset$. The learning strategy proceeds as follows:

- (1) *Meta-initialization*: Different from the naïve initialization process formulated in Eq. (4), meta-initialization is implemented by episodic training (Fig. 5) upon the DA loss. First, we meta-initialize a vanilla DA model over the support set Sup by:

$$\theta^{meta} \leftarrow \theta - \lambda_{DA} \nabla_{\theta} [\mathcal{L}_{DA}(X_s^{SUP}, Y_s^{SUP}, X_t^{SUP}; \theta)] \quad (9)$$

where λ_{DA} is the learning rate of conventional domain adaptation (tuning λ_{DA} is discussed in IV.C). Eq. (9) can be solved with a mini batch of normal gradient decent optimizations. Note that gradient backpropagation upon Eq. (9) is held until the meta-testing stage (see Algorithm 1).

Algorithm 1

Learning of meta-initialization.

Input: $\{X_s^{SUP}, Y_s^{SUP}, X_t^{SUP}\}, \{X_t^{QUE}\}$
Initialize: random initialization: $\theta_{1,1}$; hyperparameter: β ; learning rate: λ_{DA} and λ_{META}
for $i = 1, 2, \dots, I$ # of training iterations: I
 # meta-initialization
 for $j = 1, 2, \dots, J$ # of meta-learning iterations: J
 $\mathcal{L}_{DA}^{i,j} \leftarrow \mathcal{L}_{DA}(X_s^{SUP}, Y_s^{SUP}, X_t^{SUP}; \theta_{i,j})$
 $\theta_{i,j+1} = \theta_{i,j} - \lambda_{DA} \nabla_{\theta_{i,j}} \mathcal{L}_{DA}^{i,j}$
 end
 # meta-testing
 $\mathcal{L}_{CE}^{i,j} \leftarrow \mathcal{L}_{CE}(X_t^{QUE}; \theta_{i,j}^{meta})$
 $\theta_{i+1,1} = \theta_{i,1} - \lambda_{META} \nabla_{\theta_{i,1}} [\beta \mathcal{L}_{CE}^{i,j} + \mathcal{L}_{DA}^{i,1}]$
end
Output: θ^{meta^*}

- (2) *Meta-testing*: Second, the conditional entropy loss, $\mathcal{L}_{CE}$, operates fast entropy evaluation on the query set by:

$$\mathcal{L}_{CE}(X_t^{QUE}; \theta^{meta}). \quad (10)$$

With the combination of the gradients generated from $\mathcal{L}_{DA}$ and $\mathcal{L}_{CE}$, the meta-initialization θ^{meta} is updated by:

$$\theta^{meta} \leftarrow \theta - \lambda_{META} \nabla_{\theta} [\beta \mathcal{L}_{CE}(X_t^{QUE}; \theta^{meta}) + \mathcal{L}_{DA}(X_s^{SUP}, Y_s^{SUP}, X_t^{SUP}; \theta)] \quad (11)$$

where β is employed to control the gradients amplitude for meta-testing. Eq. (11) can be solved by existing episodic gradient decent optimization paradigms, e.g., *MAML* [20] (specific implementation details are reported in IV.C). By means of Eq. (11), meta-testing provides effective regulations (dynamic gradients feedback) to guide the meta-initialization to proceed in a coordinate way with the cluster assumption in the target domain.

The resulting *meta-initialization*, θ^{meta^*} , can be obtained by alternating between steps of meta-initialization and meta-testing (Algorithm

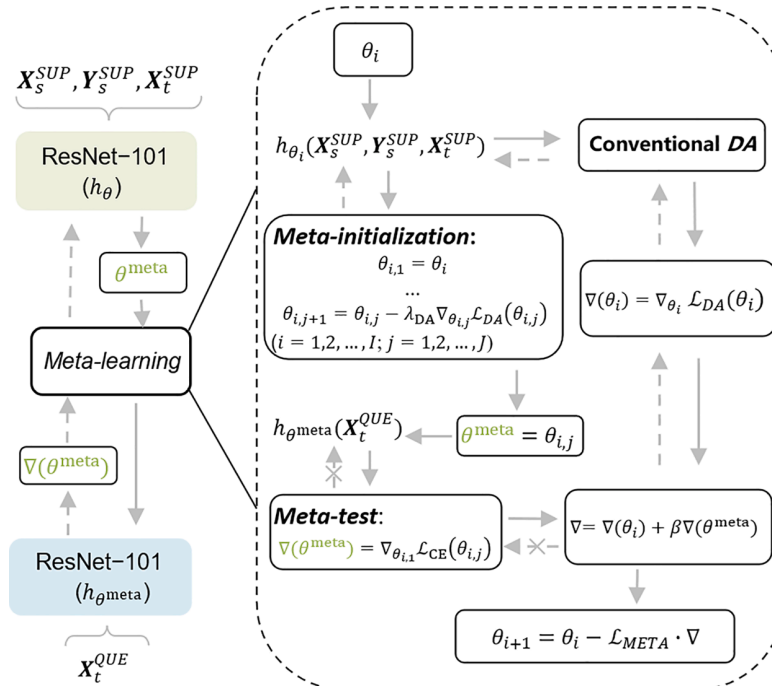


Fig. 5. Illustration of meta-learning based target-dependent classifier learning strategy.

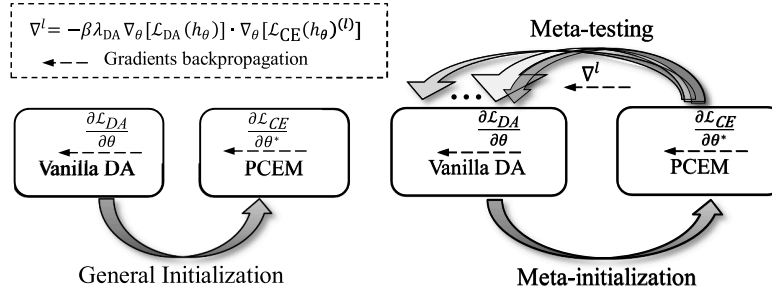


Fig. 6. **Left:** unsupervised fine-tuning proposes a sequential learning process upon vanilla domain adaptation and conditional entropy minimization. **Right:** meta-learning strategy provides the class-specific gradients regularizations (∇^l , where $l \in 1, \dots, L$), resulting in a reliable initial condition: meta-initialization.

1). Lastly, *PCEM* learns a target-dependent classifier by normal gradients descent optimization on Eq. (8) with the initialization, θ^{meta^*} .

Discussion. In order to further demonstrate the benefit that *meta-learning* has on the development of a target-dependent classifier, we make the following considerations.

Proposition 1. Meta-learning provides class-dependent dynamic gradients regularization (see a conceptual view in Fig. 6) during the pixel-wise domain adaptation process.

Proof. We first rewrite the pixel-wise meta-testing function $\mathcal{L}_{CE}(\mathbf{X}_t^{\text{QUE}}; \theta^{\text{meta}})$ in Eq. (10) as:

$$\mathcal{L}_{CE}(h_{\theta^{\text{meta}}}) = \mathcal{L}_{CE}(h_{\theta - \lambda_{DA} \nabla_{\theta} \mathcal{L}_{DA}}). \quad (12)$$

By means of the first-order Taylor expansion, Eq. (12) can be approximated as:

$$\mathcal{L}_{CE}(h_{\theta^{\text{meta}}}) \approx \mathcal{L}_{CE}(h_{\theta}) + \lambda_{DA} \nabla_{\theta} [\mathcal{L}_{CE}(h_{\theta})] \nabla_{\theta} [\mathcal{L}_{DA}(h_{\theta})]. \quad (13)$$

Considering that h_{θ} is the soft-max output of L categories on the output space, we then rewrite Eq. (13) as:

$$\mathcal{L}_{CE}(h_{\theta^{\text{meta}}}) \approx \sum_{l=1}^L \mathcal{L}_{CE}(h_{\theta})^{(l)} - \lambda_{DA} \sum_{l=1}^L [\nabla_{\theta} \mathcal{L}_{DA}(h_{\theta})^{(l)}] \cdot \sum_{l=1}^L [\nabla_{\theta} \mathcal{L}_{CE}(h_{\theta})^{(l)}]. \quad (14)$$

Finally, we rewrite the two objective functions in Eq. (11) as:

$$\begin{aligned} & \beta \cdot \mathcal{L}_{CE}(\mathbf{X}_t^{\text{QUE}}; \theta^{\text{meta}}) + \mathcal{L}_{DA}(\mathbf{X}_s^{\text{SUP}}, \mathbf{Y}_s^{\text{SUP}}, \mathbf{X}_t^{\text{SUP}}; \theta) \\ & \approx \beta \sum_{l=1}^L \mathcal{L}_{CE}(h_{\theta})^{(l)} + \sum_{l=1}^L \mathcal{L}_{DA}(h_{\theta})^{(l)} \\ & \quad - \beta \lambda_{DA} \sum_{l=1}^L [\nabla_{\theta} \mathcal{L}_{DA}(h_{\theta})^{(l)}] \cdot \sum_{l=1}^L [\nabla_{\theta} \mathcal{L}_{CE}(h_{\theta})^{(l)}] \\ & = \sum_{l=1}^L [\beta \mathcal{L}_{CE}(h_{\theta})^{(l)} + \mathcal{L}_{DA}(h_{\theta})^{(l)} + w^{(l)} \cdot R] \end{aligned} \quad (15)$$

where $w^{(l)} = \nabla_{\theta} \mathcal{L}_{CE}(h_{\theta})^{(l)}$, and $R = -\beta \lambda_{DA} \nabla_{\theta} \mathcal{L}_{DA}(h_{\theta})$.

In Eq. (15), we have the class-wise objective form of the meta-learning optimization. The term $\beta \mathcal{L}_{CE}(h_{\theta})^{(l)} + \mathcal{L}_{DA}(h_{\theta})^{(l)}$ acts as a general entropy minimization-based domain adaptation method [8,38], by which not only are the domain manifolds aligned, but also the target-domain cluster assumption is guaranteed. The second term: $w^{(l)} \cdot R$ can be considered as a class-dependent regularizing term which is weighted by a class-wise weight $w^{(l)}$, which orients the adaptation optimization in a direction where conditional entropy is also minimized. In other words, the meta-learning component identifies “where” to optimize, while the domain adaptation indicates “how” to optimize.

Proposition 2. The proposed unsupervised meta-testing loss

demonstrates unique contributions to meta-learning in domain adaptation.

Analysis. Given the absence of supervision on the target domain, it is untraceable to formulate a meta-testing phase on the target domain to extract beneficial meta-knowledge. Current research [54,55] attempts to connect meta-learning to domain adaptation by utilizing source domain labels to formulate the query set so as to drive the meta-testing progress, e.g., Online Meta-learning [54] splits a partial source domain data to compose a query set, while Guo et al. [55] utilizes a pre-trained domain predictor to select target-like source pixels for the query set. However, the full dependence on source domain in meta-testing will inevitably result in domain misalignment information, which potentially compromises the meta-learning performance. In this research, we rather utilize the target domain pixels to compose the query set. Combining with the unsupervised meta-testing loss, $\mathcal{L}_{CE}$, the meta-testing phase is fully dependent on the target statistics, resulting in beneficial target-biased initial conditions.

4. Experiments

In this section, we experimentally evaluate the performance of the two proposed types of *target-dependent* classifiers (*fine-tuning* based and *meta-learning* based) on two “*synthetic to real-city*” scenarios and a “*cross real-city*” scenario.

4.1. Datasets

Cityscapes [58] is a real-world street scene dataset containing 2,975 training images and 500 validation images, with high resolution (2048×1024). GTA5 [59] is a synthetic street scene dataset extracted from a realistically rendered computer game. It forms a large dataset with 24,966 images with high resolution (1914×1052). SYNTHIA [60] contains 9,400 synthetic images with a resolution of 760×1280 . It is less realistic than GTA5 and uses different viewing angles, presenting more severe domain shift. The entire image sets of GTA5 and SYNTHIA with ground truth labels are used as two source domains. The training set of Cityscapes (without labels) is considered as the target domain. This combination allows to learn Cityscapes-dependent classifiers from each source domain.

Oxford RobotCar [61] is a real-world dataset, which consists of four sets of real street scenes with 895 training images and 271 validation images collected in rainy scenes. A cross real-world domain scenario is considered as an additional test case for the proposed target-dependent classifiers. The training set of the real-world Cityscapes dataset (with labels) is considered as the single source domain while testing is conducted on Oxford RobotCar dataset. This scenario allows to learn city specific classifiers from the Cityscapes domain.

4.2. Baselines

As discussed in Section 3, the principle of the proposed approach provides benefits when combined with vanilla DA paradigms. To perform a comprehensive evaluation, we consider a range of state-of-the-art domain adaptation models as baselines for experiments.

Source-dependent classifier (without adaptation) is trained while fully depending on data from the source domain without seeing any target data.

ADVENT [8] (CVPR 2019) is an output-space based adversarial adaptation approach. We consider an “AdvEnt” (ResNet101 based) model, with their same implementation details, as a baseline framework.

DTST [40] (CVPR 2020) considers the difference of appearance variance between the stuff regions and the instances of things, and thus treats them differently with a semantic and instance matching module in the adaptation process. We consider their reported most efficient model (ResNet101 based) as a baseline framework.

CAA [62] (TPAMI 2021) raises the semantic inconsistency problem in pixel-wise domain adaptation and proposes particular modules to conduct category-level alignment. We consider the ResNet101 based model, CLAN (full), as a baseline framework.

SAC [27] (CVPR 2021) utilizes a style transfer algorithm and an image translation algorithm to conduct domain adaptation upon intermediate translated image spaces. We adopt all strategies used in the resulting ResNet101 model to train a baseline framework.

RPLL [63] (IJCV 2021) proposes to estimate the prediction uncertainty during training to rectify the pseudo label learning for self-supervised segmentation adaptation. We consider MRNet (ResNet101 based) as a baseline.

PKA [7] (TIP 2021) proposes a prototypical knowledge adaptation strategy by utilizing a soft adversarial loss to regularize the output-space adaptation process at a finer detail level. We consider the most efficient reported model as a baseline framework.

4.3. Implementation details

In our first experiment, we train joint-domain classifiers with the implementations reported in baselines and target-dependent classifiers (fine-tuning based, FT) with the strategy presented in Section 3.2, to evaluate whether baselines’ adaptation performance can be improved after the proposed unsupervised fine-tuning process. For fair comparison, we use DeepLab-v2 [64] with ResNet-101 [65] backbone as a fine-tuning network with all baselines, using the SGD method [66] (the momentum is set as 0.9, weight decay is 10^{-4} and the learning rate is 2.5×10^{-4}). We utilize Virtual

Adversarial Training [56] to realize the locally-Lipschitz constraint:

$$\mathcal{L}_{Lip} = \max_{\|r\|_2 < \varepsilon} D_{\mathcal{H}}(h_{\theta^*}(X_i) \| h_{\theta^*}(X_i + r)), \quad (16)$$

where $\|r\|_2 < \varepsilon$ defines a norm-ball where the input perturbation r is constrained (a visual presentation of input perturbation is provided in Fig. 8). Following [17,57], we set the hyperparameters ε as 3.5. We also experimenting alternatives to the locally-Lipschitz constraint in Section 4.8.

The goal of the second experiment is to evaluate whether meta-learning (ML) can further facilitate the development of the target-dependent classifier. In this way, we first combine meta-learning with baselines to gain meta-initializations, and then we learn target-dependent classifiers upon meta-initializations. During training, we set the meta-learning rate λ_{META} to 2.5×10^{-4} and it is decreased using the polynomial decay with power of 0.9. For the learning rate λ_{DA} over vanilla DA, we follow the values claimed by the baselines. For single-round based adaptation networks [8,27,62], we perform meta-testing operation at the end of conventional adaptation loss in each iteration before the network optimization. For multi-round baseline networks [7, 40,63], meta-testing is placed in each self-supervised round.

Table 1
Experimental results for the first “synthetic to real-city” adaptation task “GTA5 to Cityscapes” in terms of mean IoU. For each baseline, the joint-domain classifier is first presented. The two proposed target-dependent classifiers (Fine-tuning based (Proposed (FT)) and Meta-learning based (Proposed (ML))) follow. In addition, we indicate the various DA strategies used in baselines: A(adversarial DA), S(self-supervised learning), E (entropy minimization) and T(image translation based DA).

GTA5 to Cityscapes	road	side.	Build.	Wall	fence	pole	light	sign	vegetation	terra.	Sky	person	rider	car	truck	bus	train	motorbike	bike	mIoU	gain
Source-dependent Classifier	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6	-
Proposed (FT)	82.0	19.9	81.3	20.9	23.8	26.7	36.5	22.1	79.2	11.7	79.3	58.9	28.3	77.1	36.8	23.6	0	27.5	38.9	40.7	4.1
ADVENT [8] (A+E)	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8	-
IntraDA [43] (A+S+E)	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0	30.2	35.8	46.3	2.5
Proposed (FT)	86.9	31.1	82.7	31.8	23.7	30.0	35.3	24.0	84.1	41.7	77.8	60.3	28.5	84.3	41.5	49.7	0	32.2	44.1	46.8	3.0
Proposed (ML)	89.7	32.8	82.2	26.5	27.7	32.4	36.7	29.7	85.9	42.1	78.4	60.9	32.5	85.8	32.6	43.4	0.2	39.9	36.1	47.2	3.4
DTST [40] (T+A+S)	90.6	44.7	84.8	34.3	28.7	31.6	32.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2	-
Proposed (FT)	92.9	51.4	85.7	36.7	29.2	35.9	39.5	43.1	85.1	42.5	86.3	62.0	31.3	84.9	31.1	50.0	0.0	38.1	39.5	50.8	1.6
Proposed (ML)	93.2	52.6	85.6	33.3	30.7	33.3	39.9	38.1	85.6	42.8	87.6	63.1	34.2	83.7	30.3	51.7	0.4	36.3	40.2	50.7	1.5
CAA [62] (A)	88.7	35.5	80.3	27.5	25.0	29.3	36.4	28.1	84.5	37.0	76.6	58.4	29.7	81.2	38.8	40.9	5.6	32.9	28.8	45.5	-
Proposed (FT)	91.4	34.8	82.7	28.6	29.1	32.4	36.6	31.1	85.8	36.1	76.1	58.8	31.2	83.4	40.6	42.4	4.4	37.6	31.3	47.1	1.6
Proposed (ML)	91.6	33.9	84.4	31.0	29.9	32.5	36.5	30.4	87.1	39.4	78.0	60.6	30.9	82.4	42.1	42.6	5.0	36.4	32.3	47.7	2.2
SAC [27] (S+T)	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8	-
Proposed (FT)	90.4	53.6	86.8	41.9	28.0	40.8	46.9	43.2	87.5	42.1	86.1	66.3	30.0	88.2	45.8	56.4	25.9	28.8	47.9	54.5	0.7
Proposed (ML)	90.4	53.8	87.1	42.3	28.2	41.8	49.5	45.4	87.7	42.8	86.0	68.3	30.3	88.4	46.6	56.8	27.6	32.0	50.2	55.6	1.8
RPLL [63] (S)	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	71.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3	-
Proposed (FT)	93.4	36.6	89.6	37.1	30.6	40.1	49.6	49.9	88.5	35.6	73.9	66.1	39.0	88.1	41.6	51.1	0.9	33.2	47.1	52.2	1.9
Proposed (ML)	91.2	37.4	90.4	35.1	28.7	42.2	51.9	50.0	85.0	38.1	74.1	66.8	40.4	87.6	41.2	53.6	0.7	34.1	46.6	52.4	2.1
PKA [7] (A+S+E)	93.0	55.0	85.9	32.7	30.0	34.5	40.0	36.3	85.7	40.6	86.4	62.5	32.9	82.8	26.7	50.5	0.2	39.8	49.4	50.8	-
Proposed (FT)	92.4	56.9	85.5	35.1	30.9	42.0	45.4	39.8	89.6	43.1	89.9	64.4	37.6	85.5	29.7	52.6	0.8	46.5	51.2	53.4	2.6
Proposed (ML)	92.3	54.6	86.6	38.2	30.1	43.7	46.7	37.3	88.4	47.7	84.7	65.3	38.6	86.7	29.6	53.3	0.8	45.1	52.4	53.8	3.0

Table 2

Experimental results for the second “synthetic to real-city” adaptation task “SYNTHIA to Cityscapes”.

SYNTHIA to Cityscapes	road	side.	Build.	Light	sign	vegetation	sky	person	rider	car	bus	motorbike	bike	mIoU	gain
Source-dependent Classifier	55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6	-
Proposed (FT)	75.3	34.3	75.4	15.8	14.2	79.0	82.0	51.5	20.8	58.5	32.7	18.7	30.6	45.3	6.7
ADVENT [8] (A+E)	87.0	44.1	79.7	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	47.6	-
IntraDA [43] (A+S+E)	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9	1.3
Proposed (FT)	84.3	39.1	78.6	8.3	11.1	79.2	81.6	57.7	28.5	73.9	38.4	27.3	39.1	49.8	2.2
Proposed (ML)	83.3	36.4	77.3	9.1	12.0	78.0	82.5	58.4	26.7	77.9	42.5	26.5	44.2	50.2	2.6
DTST [40] (T+A+S)	83.0	44.0	80.3	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.7	28.5	45.8	52.1	-
Proposed (FT)	82.4	45.2	79.9	15.7	16.1	82.1	84.3	61.1	35.0	69.1	38.2	29.5	46.1	52.7	0.6
Proposed (ML)	85.1	43.8	82.6	16.0	14.5	81.6	84.4	60.2	35.6	71.9	37.6	27.6	47.0	52.9	0.8
CAA [62] (A)	82.7	37.2	81.5	17.1	13.1	81.2	83.3	55.5	22.1	76.6	30.1	23.5	30.7	48.8	-
Proposed (FT)	84.0	39.4	82.4	19.5	12.9	82.5	84.6	55.9	24.5	77.9	32.5	25.6	32.3	50.3	1.5
Proposed (ML)	85.1	38.8	83.1	18.8	13.2	80.6	85.2	55.4	25.1	77.0	31.8	24.9	33.3	50.2	1.4
SAC [27] (T)	89.3	47.2	85.5	45.5	32.0	87.1	89.3	63.6	25.4	86.9	35.6	30.4	53.0	59.3	-
Proposed (FT)	89.2	47.3	88.8	40.7	33.6	88.5	91.2	65.5	30.4	80.8	33.9	32.6	55.4	59.8	0.5
Proposed (ML)	90.1	49.2	90.0	44.1	36.1	90.2	88.8	64.4	31.1	88.1	35.5	33.1	56.4	61.3	2.0
RPLL [63] (S)	87.6	41.9	83.1	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	54.9	-
Proposed (FT)	90.4	42.6	82.0	32.3	21.4	82.7	80.7	62.7	22.4	85.5	42.6	25.1	55.5	55.8	0.9
Proposed (ML)	87.9	43.0	84.6	33.1	17.3	82.6	78.9	63.0	23.5	86.6	42.3	23.9	55.4	55.5	0.6
PKA [7] (A+S+E)	86.8	40.3	83.3	21.6	18.4	83.2	84.2	61.0	25.6	85.7	34.5	20.4	44.3	53.0	-
Proposed (FT)	89.1	42.6	86.6	23.2	20.6	86.4	77.4	64.0	28.0	87.6	37.4	23.1	47.8	54.9	1.9
Proposed (ML)	88.4	45.3	88.6	24.9	21.7	87.4	80.6	66.4	29.3	90.4	38.0	26.4	48.9	56.6	3.5

4.4. Synthetic to real-city test cases

Experimental results on the two “synthetic to real-city” test cases, “GTA5 to Cityscapes” and “SYNTHIA to Cityscapes”, are detailed in Table 1 and Table 2, respectively. We provide observations to three types of classifiers:

- Source-dependent classifier.* When the classifier is trained from the source domain (GTA5 or SYNTHIA) and directly infers on the target domain (Cityscapes), the performance under the metrics of mean Intersection over Union (mIoU) [67] reaches a mere 36.6% and 38.6% on the two tasks. It shows that domain shift deteriorates the classification accuracy in the target domain with severe performance drop.
- Joint-domain classifiers (baselines).* Results show that joint-domain classifiers for all baselines surpass the source classifiers (upper row of Tables 1 and 2) by a clear margin, which demonstrates that all baselines operate effective adaptation across domains. As an efficient DA integration models, SAC [27] achieves the best adaptation performance in Tables 1 and 2.
- Target-dependent classifiers (proposed).* We first observe that the fine-tuning (FT) based classifiers can achieve a direct adaptation improvement over the source-dependent classifier, which demonstrates the fundamental efficacy of the proposed unsupervised fine-tuning strategy on cross-domain semantic segmentation problems. When combined with DA baselines, fine-tuning boosts all baselines in terms of mean IoU (detailed in the “gain” column of Tables 1 and 2), with slight variations observed on the different vanilla DA baseline frameworks considered. The best improvement can be observed on the “ADVENT [8]” baseline, where the proposed method also outperforms IntraDA [43], an intra-domain adaptation method which employs a secondary adversarial adaptation strategy. Conversely, when integrating meta-learning (ML) with baselines to learn target-dependent classifiers based on meta-initializations, results in Tables 1 and 2 show that the ML classifiers achieve more compelling performance and outperform fine-tuning (FT) based classifiers in most cases. We also provide in Fig. 7 the visualization results on semantic segmentation adaptation upon baselines.

4.5. Cross real-city test case

Experimental results in terms of mean IoU on the task “Cityscapes to

Real Cities” are summarized in Table 3. The source-dependent classifier is trained while fully relying on data from the Cityscapes. It can be observed that the source-dependent classifier suffers from poor classification performance on the Oxford RobotCar dataset. Although the *joint-domain* classifiers [7,63] gain efficient domain adaptation and outperform the source-dependent classifier, the proposed *target-dependent* classifiers boost the baseline method in all cities, demonstrating the efficacy of the proposed methods on a challenging cross-domain scenario.

4.6. Ablation study

Experimental ablation studies are also conducted on the task “GTA5 to Cityscapes” to investigate the effectiveness of different components of the proposed target-dependent classifiers. We consider ASNet [6] as the vanilla DA model and train target-dependent classifiers with three particular loss functions: (i) the *entropy minimization loss* $\mathcal{L}_E$ [8], (ii) the *locally Lipschitz constraint* $\mathcal{L}_{Lip}$, and (iii) the *KL based constraint* $\mathcal{L}_{CONS}$. Experimental results are summarized in Table 4.

(1) *Effectiveness of locally-Lipschitz constraint.* Interacting with the entropy minimization (EntMin) progress, the Locally-Lipschitz constraint $\mathcal{L}_{Lip}$ in Eq. (5) is expected to enforce the target-dependent classifier consistency on predictions within a slight but dynamic input perturbation (see a visual interpretation in Fig. 8). As the ablation experiments shown in Table 4, EntMin does not demonstrate substantial effectiveness for unsupervised fine-tuning optimization, while the pixel-wise conditioned entropy minimization (vanilla PCEM) method ($\mathcal{L}_E + \mathcal{L}_{Lip}$) brings a clear adaptation improvement (2.1 %). To further evaluate the independent effectiveness of using the locally-Lipschitz constraint, we replace locally-Lipschitz perturbation noise with normal Gaussian noises $\mathcal{L}_{Gau}$ and conduct PCEM adaptation ($\mathcal{L}_E + \mathcal{L}_{Gau}$). The results in Table 4 show that such an attempt will not provide an improvement upon the normal entropy minimization.

(2) *Effectiveness of using a sibling network.* It is also worth to evaluate the independent effectiveness of using the *KL based constraint* $\mathcal{L}_{CONS}$ in Eq. (8). The result in Table 4 indicates that the regularized PCEM ($\mathcal{L}_E + \mathcal{L}_{Lip} + \mathcal{L}_{CONS}$) further boosts the performance, which suggests that the introduction of the sibling network facilitates the optimization of a target-dependent classifier.

4.7. Parameter analysis

We also make an analysis on the choice of two important hyper-

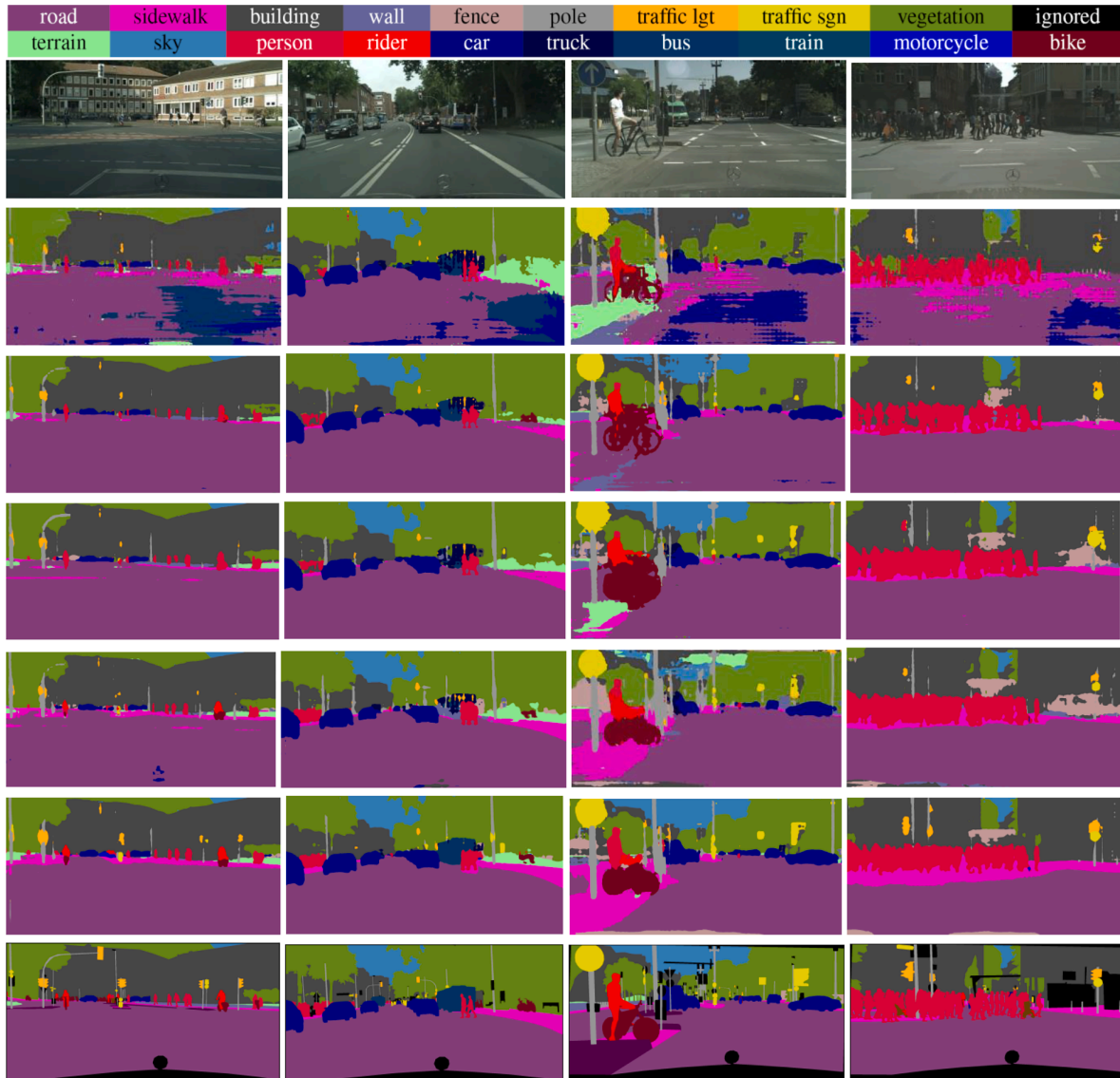


Fig. 7. Prediction results obtained for semantic segmentation on the task “GTA5 to Cityscapes” without and with domain adaptation. Upper two rows are original image samples and prediction results of the source-dependent classifier (without-adaptation). Prediction results with the proposed *target-dependent* classifiers follow: from the third row to the sixth row, results correspond to classifiers meta-learned (ML) on different baselines: ADVENT, DTST, CAA, and SAC. The last row shows ground truth labels.

Table 3

Experimental results of the adaptation task “Cityscapes to Oxford RobotCar” under the metrics of IoU on each class and mean IoU on overall performance.

Cross Real-City Test Case (Cityscapes to Oxford-RobotCar)											
Method	road	sidewalk	building	light	sign	sky	person	automobile	two-wheel	mIoU	gain
Source-only	79.2	49.3	73.1	55.6	37.3	36.1	54.0	81.3	49.7	61.9	-
Proposed (FT)	85.1	60.4	80.1	75.4	50.2	44.9	55.1	88.4	54.6	66.0	4.1
Proposed (ML)	88.4	63.6	85.4	74.9	51.4	50.6	49.5	90.4	58.6	68.1	6.2
PKA [7]	93.8	77.9	86.1	70.2	40.3	89.1	56.0	91.1	60.4	73.9	-
Proposed (FT)	92.3	80.6	85.5	71.9	38.6	90.4	58.6	92.3	66.7	75.2	1.3
Proposed (ML)	95.5	82.4	84.9	75.6	41.6	92.4	60.6	94.8	66.5	77.2	3.3
RPLL [63]	95.9	73.7	87.4	72.8	43.1	88.6	61.7	89.6	57.0	74.4	-
Proposed (FT)	96.1	79.5	86.8	73.5	44.9	90.1	66.4	92.3	61.9	76.8	2.4
Proposed (ML)	97.5	79.4	84.6	75.6	48.3	91.8	68.4	89.4	64.1	77.7	3.3

Table 4
Experimental ablation study results for the “GTA5 to Cityscapes” task.

Method	Baseline	Components			mIoU
	$\mathcal{L}_E$	$\mathcal{L}_{Gau}$	$\mathcal{L}_{Lip}$	$\mathcal{L}_{CONS}$	
EntMin [8]	✓				42.4
EntMin [8] with Gaussian	✓	✓			42.3
Vanilla PCEM (proposed)	✓		✓		44.5
Regularized PCEM (proposed)	✓		✓	✓	46.8

parameters α and β . First, α is employed to weight $\mathcal{L}_{CONS}$ in Eq. (8). We learn target-dependent classifiers from each baseline with the same fine-tuning implementation details but altering α over the range [0, 1.5]. As demonstrated by the results reported in Table 5, using a small α will let the fine-tuning degrade into a plain conditional entropy minimization strategy. Second, β introduced in Eq. (15) is employed to control the gradients amplitude for entropy regularization in meta-learning. When $\beta = 0$, meta-learning degrades into a plain fine-tuning strategy, as shown in Table 6. One can observe that both the fine-tuning and meta-learning processes can benefit from a proper value of α and β , respectively, e.g., the fine-tuning based classifiers can achieve the best performance when α is around 0.7, and the best meta-learning-based performance happens with β between 0.005 and 0.01.

4.8. Alternatives to the locally-Lipschitz constraint

(1) *Input gradients*. The experimental investigation examines alternatives to imposing input perturbation for the Lipschitz constraint. It first interprets the virtual adversarial training objective introduced in Eq. (16) in a sense of *Lipschitz continuous*. The $D_{\mathcal{H}\mathcal{L}}$ term in Eq. (16) can be rewritten as:

$$D_{\mathcal{H}\mathcal{L}}(h_{\theta^*}(X_t) \| h_{\theta^*}(X_t + r)) = |h_{\theta^*}(X_t)(\log(h_{\theta^*}(X_t)) - \log(h_{\theta^*}(X_t + r)))|. \quad (17)$$

By means of first order Taylor expansion in r , we have the approximation of $\mathcal{L}_{Lip}$ as:

$$\mathcal{L}_{Lip} = \max_{\|r\|_2 < \varepsilon} |r \cdot \nabla_x [h_{\theta^*}(X_t)] + h_{\theta^*}(X_t) \cdot O(r)| \quad (18)$$

$$\approx \max_{\|r\|_2 < \varepsilon} r \|\nabla_x [h_{\theta^*}(X_t)]\| = \varepsilon \cdot \|\nabla_x [h_{\theta^*}(X_t)]\|_2. \quad (19)$$

In Eq. (18), $O(\cdot)$ denotes the higher order terms of the expansion. Eq. (19) suggests that virtual adversarial learning implicitly imposes regularizations on the Frobenius norm of the Jacobian matrix of h_{θ^*} with respect to X_t . Imposing constraint on input gradients, the virtual adversarial training sufficiently guarantees *Lipschitz continuous* on the hypothesis h_{θ^*} .

Inspired by such an interpretation, we consider the input gradients as an alternative to $\mathcal{L}_{Lip}$ in Eq. (5) and compare the effectiveness of different versions of the Lipschitz constraint in experiments. Specifically, two types of input gradients constraints, i.e., global constraint (detailed on the second row of Table 7) and class-wise constraint (detailed on the third row of Table 7), are evaluated for comparison. Experimental results are concluded in Table 7. First, both input gradients constraints and cross-Lipschitz regularizer are effective to the resulting classifier, indicating that imposing smoothness near the data points over the optimization process does have the effectiveness for a better target-dependent classifier. Among three constraints, virtual adversarial learning achieves the best performance improvement. We are also interested in investigating more Lipschitz constraint terms for domain adaptation in our future works.

(2) *Input augmentation*. Furthermore, the work also investigates the efficacy of recently proposed strong image augmentation strategies [27, 46, 68, 69] into the PCEM strategy as an alternative to Lipschitz constraint. First, it reformulates Eq. (16) as:

$$\mathcal{L}_{Aug} = D_{\mathcal{H}\mathcal{L}}(h_{\theta^*}(X_t) \| h_{\theta^*}(\mathcal{T}(X_t))) \quad (20)$$

where $\mathcal{T}(\cdot)$ denotes the augmentation functions detailed as $\mathcal{T}_1$ and $\mathcal{T}_2$ in Table 7. The semantic consistency interpreted by KL divergence in Eq. (20) formulates an alternative objective function to the original virtual adversarial training objective in Eq. (16). In this experiment, we use all

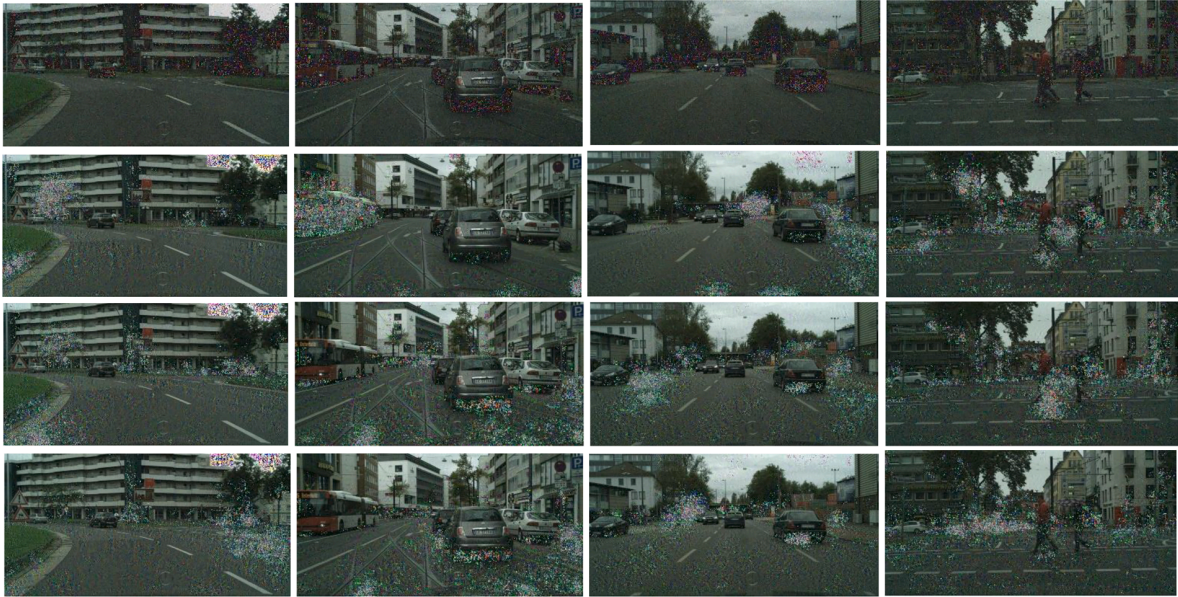


Fig. 8. Input perturbations with regard to different types of noise on the input image samples of the target domain (Cityscapes) (**Top row**: Gaussian noises on the input image samples; **Second row to bottom row**: Locally-Lipschitz perturbation noises on the input image samples at *Epoch 1*, *Epoch 10*, and *Epoch 25*, generated via virtual adversarial training [56]). It is observed that, different from the normal and global Gaussian noise pattern, the Locally-Lipschitz perturbation noises vary among adaptation epochs upon the local uncertainty of the classifier on particular categorical regions, as the response of the model on a slight input perturbation, which therefore forms the dynamic constraint (Locally-Lipschitz constraint $\mathcal{L}_{Lip}$ in Eq. (5)) on the proposed method, PCEM, and forces the *target-dependent* classifier to pay particular attention on categorical regions and preserves from reaching over-adaptation results.

Table 5
Tuning of the hyperparameter α .

[illegible]

Table 6
Tuning of the hyperparameter β .

Meta-learning (GTAS to Cityscapes)																
		RPIL			ADVENT			DTST			CAA			SAC		
β	0.001	0.005	0.007	0.01	0.001	0.005	0.007	0.01	0.001	0.005	0.007	0.01	0.001	0.005	0.007	0.01
mIoU	52.2	52.4	52.3	52.0	46.7	47.0	47.1	46.6	50.5	50.4	50.3	50.7	46.9	47.6	47.7	55.8

Table 7

Comparison on the variants of the locally-Lipschitz constraint. Experiments are conducted under the fine-tuning (GTA5 to cityscapes) evaluation for ADVENT. “Baseline (no constraint)” refers to the “EntMin [8]” model in Table 4. The class-wise constraint is revisited from cross-Lipschitz regularizer [70]. We follow [68] on the implementation of Fourier transform in the target domain samples. Using the image augmentation function $\mathcal{F}_2$, we follow [27] to exclude random colour jitter and Gaussian filter to the sibling network but the full augmentations to the base network.

Baseline (no constraint)		42.4
locally-Lipschitz constraint	$ h_{\theta'}(X_t) \cdot \log \frac{h_{\theta'}(X_t)}{h_{\theta'}(X_t + r)} $	46.8
	$\ \nabla_x [h_{\theta'}(X_t)]\ _2$	45.4
	$\frac{1}{L^2} \sum_{l,k=1}^L \ \nabla_x [h_{\theta'}(X_t)^l] - \nabla_x [h_{\theta'}(X_t)^k]\ _2$	46.1
image augmentation	$\mathcal{F}_1$: Fourier transform [68,69]	43.2
	$\mathcal{F}_2$: Random crop, Horizontal flips, Random colour jitter, and Gaussian filter [27,46]	44.7

Table 8

Tuning of the hyperparameter γ under the meta-learning (GTA5 to cityscapes) evaluation for ADVENT. β is set as 0.007 in experiments.

γ	0.0	0.3	0.5	0.7	1.0
mIoU	47.2	47.2	47.7	46.9	45.1

the same setting of *PCEM* proposed in Section 3.2 but alter the augmentation function to observe the particular contribution of image augmentation to target-dependent classifier. Experimental results are reported in Table 7. The two types of data augmentation strategies demonstrate a moderate effectiveness in *PCEM*, with slight variations on performance. It is relevant to pursue the investigation of more effective data augmentation strategies in future work.

4.9. Choice of the meta-testing objective

Online Meta-learning [54] proposes a semi-supervised meta-learning scheme for domain adaptation by utilizing label information in the target domain in the meta-testing phase. Given the absence of target domain labels, a workaround to unsupervised domain adaptation meta-learning is to leverage pseudo labels in the meta-testing phase. Specifically, in each iteration of the meta-learning phase, we yield pseudo label, $\hat{Y}_t^{QUE}$, of the query set sample X_t^{QUE} by utilizing KL divergence with regard to the input perturbation defined in Eq. (16). The element entry y of $\hat{Y}_t^{QUE}$ is obtained by $y = \argmax \{h_{\theta'}(x_t^{QUE})\}$, if $\tau \geq 0.5$; else 0, where $\tau = D_{\mathcal{H}\mathcal{L}}(h_{\theta'}(x_t^{QUE}) \| h_{\theta'}(x_t^{QUE} + r))$. The objective function of meta-testing term defined in Eq. (10) can be reformulated as: $\mathcal{L}_{CE}(X_t^{QUE}; \theta^{meta}) + \gamma \mathcal{L}_{seg}(\hat{Y}_t^{QUE}, X_t^{QUE}; \theta^{meta})$, where γ controls the weight of the pseudo label loss. In experiment, we use all the same setting of the meta-learning scheme proposed in Section 3.3 but tune γ to observe the particular contribution of the pseudo label loss to meta-testing. Table 8 demonstrates the experimental results by altering γ over the range [0.0, 0.1]. In conclusion, utilizing pseudo labels with a small weight γ , such as 0.05, could lead to incremental improvements to the resulting target-dependent classifier. However, a large weight, such as 0.1, is harmful to the meta-testing process. We suggest that the pseudo label loss in meta-testing should be used with caution as the noise in pseudo labels could corrupt the meta-testing phase.

4.10. Discussion

(1) *Computational cost for meta-learning*: As the experimental results suggest in Tables 1 and 2, learning meta-initialization for baselines guarantees performance boost to the target-dependent classifiers.

Table 9

Comparison on the computational costs of different implementations of meta-learning. All the implementations are conducted with PyTorch on a RTX 3090. The iteration time (seconds) and allocated memory are recorded with the batch size being set as 1. Last column concludes the model performance in terms of mIoU of the resulting target-dependent classifiers.

Models	Iteration Time (s)	GPU Memory (MB)	mIoU
ADVENT	0.49	7,111	43.8
ADVENT-FO-MAML	0.71	7,521	46.9
ADVENT-Sign-MAML	0.85	7,483	47.1
ADVENT-MAML	1.64	13,643	47.2

Nevertheless, by involving high-order derivative operations in the meta-testing phase, meta-learning initialization inevitably imposes computational overhead to the conventional DA baselines. For instance, connecting ADVENT, an adversarial learning DA model [8], to meta-learning, ADVENT-MAML, will lead to 9037 MB memory increase on a RTX 3090 GPU (detailed in Table 9). In order to circumvent the costs of the computational overhead, we resort to an efficient implementation of meta-learning in experiments. Specifically, we adopt FO-MAML, first order approximation of MAML, and Sign-MAML [71], for alternative implementations to MAML. Table 9 concludes the comparison on computational cost between the different versions of the meta-initializations. It is observed that both FO-MAML and Sign-MAML significantly reduce the computation overhead (in terms of iteration time and GPU memory) compared to MAML, while preserving performance improvements for the fine-tuning models. Experimental results suggest a practical trade-off between the model performance and computational cost in meta-initialization optimization.

(2) *Future works*: First, we aim to extend the proposed method to multi-source domain adaptation. With the substantial progress on unsupervised domain adaptation, a variety of pre-adapted models are released online for recycling. As discussed in this work, the proposed method can readily connect to the pre-adapted models and formulate a target-dependent classifier that fits better for the target domain. However, each pre-adapted model was optimized from a single source domain (e.g., GTA [59], Synscapes [72], and SYNTHIA [60]). Aggregating the different source domain adapted models to leverage the knowledge shared across domains could be difficult [47]. To circumvent the difficulty, without reformulating the adaptation philosophy, we resort to recent advances in *model reassembly* [73] and *model factorization* [74]. Specifically, it aggregates existing pre-adapted single-source models into *model zoo*. The *model zoo* consists of a collection of semantic segmentation models built on particular backbone architecture (e.g., DeepLabv2-ResNet101 [6,12,32,74], FCN8s-VGG16 [6,10,12,29,32,35], ResNet-38 [10,12], and DRN-D-105 [12,29]). Second, *model factorization* [74] allows to partition the pre-trained models into semantic headers and model equivalences of backbones. *Model reassembly* [73] then searches for an optimal classifier serving as an effective weight initialization for domain adaptation. Last, we adapt the reassembled model with the proposed method, implicitly realizing multi-source domain adaptation by only using the data from the target domain. The second path of the future works is to extend the proposed method to open compound domain adaptation [48] where the target is a union of multiple homogeneous domains. It helps in improving the practicality of the proposed method while operating in the condition where incoming data was collected from mixed and novel situations. The difficulty lies in dealing with multiple homogeneous domains in a compound. To realize open compound domain adaptation with the proposed method, we propose to leverage recent advances in data factorizing [75] and data condensation [76]. It allows to aggregate multiple homogenous domains into a synthetic dataset, on which the adaptation scenario degrades into pair-wise adaptation. The proposed method can be readily applied to optimize a compound target dependent classifier.

(3) *Limitations*: As presenting the first work for target-dependent

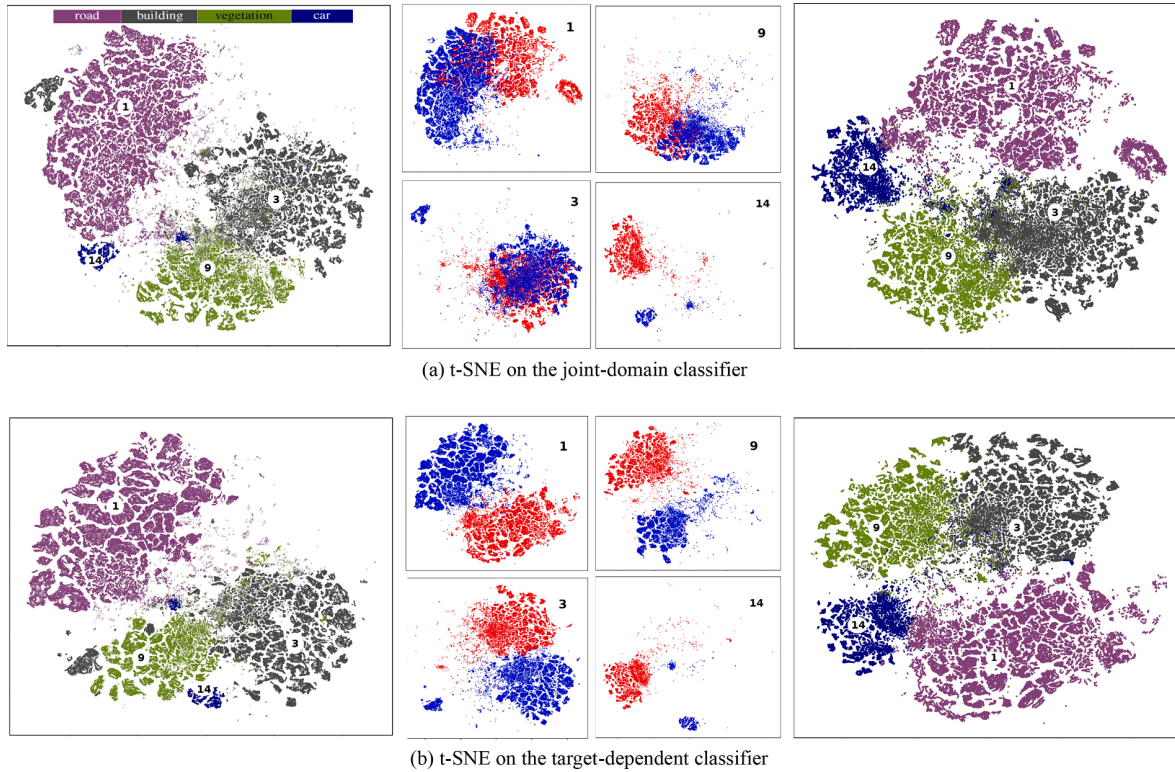


Fig. 9. Class clusters visualization using t-SNE [77] for two classifiers under the baseline ADVENT on the embedding space of the $L4$ layer of ResNet101. Left column: visualization on the source domain; Right column: visualization on the target domain; Center column: class pairs (blue cluster for the source domain and red for the target domain).

classifier in semantic segmentation, we further identify potential limitations of the proposed methodology. **First**, the proposed fine-tuning method is simple but effective to conventional baselines. However, only one semantic segmentation backbone, i.e., DeepLab-v2 [64] with ResNet-101 [65], is adopted in this work. More experiments on different backbone architectures deserve future investigations. **Second**, aiming to fine-tune from the conventional DA baselines, the proposed target-dependent classifier learning methods require the same model architecture of the baselines, which limits the scalability of this work. How to apply the proposed method to heterogeneous baselines need to further investigations. **Third**, the meta-testing objective used in this work are all revisited from current meta-learning paradigms, which is not tailored to unsupervised domain adaptation, which substantially increases the computational cost in the learning phase. **At last**, this work only focuses on single-source-single-target domain adaptation. Multi-source domain or multi-target domain adaptations can be challenging to the proposed methodology.

(4) *Deviations to conventional methods:* The work also studies the deviations of the proposed target-dependent classifier from conventional methods. Fig. 9 illustrates experimental results of t-SNE [77] obtained from the different types of classifiers while evaluating on image samples across domains. First, as shown in Fig. 9(a), the conventional method successfully realizes domain alignment on class clusters, e.g., cluster 3, while seeking to project domain-agnostic manifolds in the embedding space. However, such a paradigm happens to fail on the alignment of certain clusters, e.g., partial alignment on cluster 1 and cluster 9, and misalignment on cluster 14, which potentially leads the resulting classifier to cross the clusters of the target domain and converging to local optimum (as introduced in Fig. 1). Conversely, this proposed method rather disentangles such a hard domain alignment assumption and embraces loose domain alignment in adaptation. Even though the resulting class clusters can be deviated from the source

domain, as illustrated in Fig. 9(b), the proposed framework of PCEM is capable of adapting a specific hypothesis depending on target domain clusters to effectively reduce cluster violation, thereby realizing better adaptation to the target domain.

5. Conclusion

In this work, we observe that the fundamental assumptions of base domain adaptation can be efficiently uprooted when domain encodings fail to be perfectly aligned. This paper introduces two new perspectives for DA, that aim at learning a target-dependent classifier for semantic segmentation in domain adaptation. The first method proposes an unsupervised fine-tuning strategy, which recycles state-of-the-art pre-adapted models as initialization and adapts a target-dependent classifier to converge in the target domain with low generalization error regardless of the dominance of source domain ground truth labels. The second method connects meta-learning with domain adaptation for semantic segmentation by providing an effective meta-initialization to the training process, which facilitates the development of the target-dependent classifier. Experimental results demonstrate that the proposed strategies improve the domain adaptation performance over state-of-the-art paradigms and performs well on various synthetic-to-real and real-to-real domain adaptation scenarios. Possible limitations on the proposed method are also discussed with respect to its scalability and practicality in different domain adaptation scenarios. Last, it proposes possible paths on the future works, with the expectation of inspiring perspective research in this direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the China Scholarship Council (CSC) under Grant 201906290105 and in part by the Natural Sciences and Engineering Research Council of Canada under the Discovery Grants Program.

References

- [1] A.J. Storkey, M. Sugiyama, Mixture regression for covariate shift, *Adv. Neural Inf. Process. Syst.* 19 (2007) 1337.
- [2] J. Hoffman, D. Wang, F. Yu, and T. Darrell "FCNs in the wild: pixel-level adversarial and constraint-based adaptation." *arXiv preprint arXiv: 1612.02649* (2016).
- [3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, *Mach. Learn.* 79 (1) (2010) 151–175.
- [4] J. Hoffman, E. Tzeng, T. Park, J.Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: cycle-consistent adversarial domain adaptation, in: *Proceedings of the International Conference on Machine Learning*, 2018.
- [5] Y.H. Chen, W.Y. Chen, Y.T. Chen, B.C. Tsai, Y.C.F. Wang, M. Sun, No more discrimination: cross city adaptation of road scene segmenters, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [6] Y.H. Tsai, W.C. Hung, S. Schuster, K. Sohn, M.H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [7] H. Tian, S. Qu, P. Payeur, A prototypical knowledge oriented adaptation framework for semantic segmentation, *IEEE Trans. Image Process.* 31 (2021) 149–163.
- [8] T.H. Vu, H. Jain, M. Bucher, M. Cord, P. Perez, Advnt: adversarial entropy minimization for domain adaptation in semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [9] H. Tian, S. Qu, P. Payeur, Unsupervised pixel-wise weighted adversarial domain adaptation, in: *Proceedings of the 16th International Symposium on Visual Computing*, 2021.
- [10] L. Du, J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye, X. Zhang, SSF-DAN: separated semantic feature based domain adaptation network for semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [11] Y. Zou, Z. Yu, B.V.K. Vijaya Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [12] Y. Zou, Z. Yu, X. Liu, B.V.K. Vijaya Kumar, J. Wang, Confidence regularized self-training, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [13] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks." *NIPS*. 2014.
- [14] M. Arjovsky, and L. Bottou. "Towards principled methods for training generative adversarial networks." *arXiv preprint arXiv:1701.04862* (2017).
- [15] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: *Proceedings of the International Conference on Machine Learning*, 2017.
- [16] C.Y. Lee, S. Batra, M.H. Baig, D. Ulbricht, Sliced wasserstein discrepancy for unsupervised domain adaptation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [17] R. Shu, H.H. Bui, H. Narui, S. Ermon, A DIRT-T approach to unsupervised domain adaptation, in: *Proceedings of the International Conference on Learning Representations*, 2018.
- [18] S. Thrun, L. Pratt, Learning to learn: introduction and overview. *Learning to Learn*, Springer, Boston, MA, 1998, pp. 3–17.
- [19] D. Li, Y. Yang, Y.Z. Song, T.M. Hospedales, Learning to generalize: meta-learning for domain generalization, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 2018.
- [20] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: *Proceedings of the International Conference on Machine Learning*, 2017.
- [21] A. Nichol, J. Achiam, and J. Schulman. "On first-order meta-learning algorithms." *arXiv preprint arXiv: 1803.02999* (2018).
- [22] T. Munkhdalai, H. Yu, Meta networks, in: *Proceedings of the International Conference on Machine Learning*, 2017.
- [23] S. Sankaranarayanan, Y. Balaji, A. Jain, S.N. Lim, R. Chellappa, Learning from synthetic data: addressing domain shift for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [24] Y. Chen, W. Li, L. Van Gool, Road: reality oriented adaptation for semantic segmentation of urban scenes, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [25] Y. Li, Lu Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [26] L. Song, Y. Xu, L. Zhang, Bo Du, Q. Zhang, X. Wang, Learning from synthetic images via active pseudo-labeling, *IEEE Trans. Image Process.* 29 (2020) 6452–6465.
- [27] N. Araslanov, S. Roth, Self-supervised augmentation consistency for adapting semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [28] R. Li, W. Cao, S. Wu, H.S. Wong, Generating target image-label pairs for unsupervised domain adaptation, *IEEE Trans. Image Process.* 29 (2020) 7997–8011.
- [29] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.
- [30] S. Zhu, Y. Tian, Shape robustness in style enhanced cross domain semantic segmentation, *Pattern Recognit.* 135 (2023), 109143.
- [31] R. Li, W. Cao, Q. Jiao, Si Wu, H.S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, *Pattern Recognit.* 105 (2020), 107343.
- [32] Y. Luo, P. Liu, T. Guan, J. Yu, Y. Yang, Significance-aware information bottleneck for domain adaptive semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [33] Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, L. Zhang, Self-ensembling attention networks: addressing domain shift for semantic segmentation, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 2019.
- [34] D. Hu, J. Liang, Q. Hou, H. Yan, Y. Chen, Adversarial domain adaptation with prototype-based normalized output conditioner, *IEEE Trans. Image Process.* 30 (2021) 9359–9371.
- [35] W.L. Chang, H.P. Wang, W.H. Peng, W.C. Chiu, All about structure: adapting structural information across domains for boosting semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1900–1909.
- [36] D. Guan, J. Huang, S. Lu, A. Xiao, Scale variance minimization for unsupervised domain adaptation in image segmentation, *Pattern Recognit.* 112 (2021), 107764.
- [37] Q. Lian, F. Lv, L. Duan, B. Gong, Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: a non-adversarial approach, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [38] J. Huang, S. Lu, D. Guan, X. Zhang, Contextual-relation consistent domain adaptation for semantic segmentation, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer, 2020.
- [39] M. Chen, H. Xue, D. Cai, Domain adaptation for semantic segmentation with maximum squares loss, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [40] Z. Wang, Mo Yu, Y. Wei, R. Feris, J. Xiong, W.M. Hwu, T.S. Huang, H. Shi, Differential treatment for stuff and things: a simple unsupervised domain adaptation method for semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [41] Yi-H Tsai, K. Sohn, S. Schuster, M. Chandraker, Domain adaptation for structured output via discriminative patch representations, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [42] F. Lv, T. Liang, X. Chen, G. Lin, Cross-domain semantic segmentation via domain-invariant interactive relation transfer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [43] F. Pan, I. Shin, F. Rameau, S. Lee, I.S. Kwon, Unsupervised intra-domain adaptation for semantic segmentation through self-supervision, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [44] C. Villani, *Optimal Transport: Old and New*, 338, Springer Science & Business Media, 2008.
- [45] S. Liu, J. Ye, S. Ren, X. Wang, Dynast: dynamic sparse transformer for exemplar-guided image generation, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer Nature Switzerland, 2022, pp. 72–90.
- [46] L. Hoyer, D. Dai, L. Van Gool, Daformer: improving network architectures and training strategies for domain-adaptive semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.
- [47] J. He, X. Jia, S. Chen, J. Liu, Multi-source domain adaptation with collaborative learning for semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11008–11017.
- [48] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S.X. Yu, B. Gong, Open compound domain adaptation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12406–12415.
- [49] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: *Proceedings of International Conference on Learning Representations*, 2017.
- [50] Z. Li, F. Zhou, F. Chen, and H. Li. "Meta-SGD: learning to learn quickly for few-shot learning." *arXiv preprint arXiv:1707.09835* (2017).
- [51] K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [52] H. Liu, K. Simonyan, Y. Yang, Darts: differentiable architecture search, in: *Proceedings of the International Conference on Learning Representations*, 2018.
- [53] X. Chen, Y. Duan, Z. Chen, H. Xu, Z. Chen, X. Liang, T. Zhang, Z. Li, Catch: context-based meta reinforcement learning for transferrable architecture search, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer, 2020.

- [54] D. Li, T. Hospedales, Online meta-learning for multi-source and self-supervised domain adaptation, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer, 2020.
- [55] X. Guo, C. Yang, B. Li, Y. Yuan, Metacorection: domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [56] T. Miyato, S.I. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and self-supervised learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (8) (2018) 1979–1993.
- [57] S. Lee, D. Kim, N. Kim, S.G. Jeong, Drop to adapt: learning discriminative features for unsupervised domain adaptation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [58] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [59] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: ground truth from computer games, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer, 2016.
- [60] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A.M. Lopez, The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [61] Y.S. Xu, T.J. Fu, H.K. Yang, C.Y. Lee, Dynamic video segmentation network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [62] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, Y. Yang, Category-level adversarial adaptation for semantic segmentation using purified features, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (8) (2021) 3940–3956.
- [63] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, *Int. J. Comput. Vis.* 129 (4) (2021) 1106–1120.
- [64] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (4) (2017) 834–848.
- [65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [66] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: *Proceedings of the COMPSTAT'2010*, Physica-Verlag HD, 2010, pp. 177–186.
- [67] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the kitti dataset, *Int. J. Robot. Res.* 32 (11) (2013) 1231–1237.
- [68] L. Melas-Kyriazi, A.K. Manrai, Pixmatch: unsupervised domain adaptation via pixelwise consistency training, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12435–12445.
- [69] Y. Yang, S. Soatto, FDA: Fourier domain adaptation for semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.
- [70] M. Hein, M. Andriushchenko, Formal guarantees on the robustness of a classifier against adversarial manipulation, in: *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2017.
- [71] C. Fan, P. Ram, S. Liu, Sign-MAML: efficient model-agnostic meta-learning by SignSGD, in: *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2021.
- [72] M. Wrenninge, and J. Unger. "Synscapes: a photorealistic synthetic dataset for street scene parsing." *arXiv preprint arXiv:1810.08705* (2018).
- [73] X. Yang, D. Zhou, S. Liu, J. Ye, X. Wang, Deep model reassembly, in: *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022.
- [74] X. Yang, J. Ye, X. Wang, Factorizing knowledge in neural networks, in: *Proceedings of the European Conference on Computer Vision*, Cham, Springer Nature Switzerland, 2022, pp. 73–91.
- [75] S. Liu, K. Wang, X. Yang, J. Ye, X. Wang, Dataset distillation via factorization, *Adv. Neural Inf. Process. Syst.* 35 (2022) 1100–1113.
- [76] S. Liu, J. Ye, R. Yu, X. Wang, Slimmable dataset condensation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3759–3768.
- [77] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008) 2579–2605.

3.4 Supplementary Information

This section provides more details related to the experiments conducted in Chapter 3.

3.4.1 Additional implementation details

The proposed frameworks (Pixel-wise Weighted Adaptation (Section 3.1), Prototypical Knowledge-oriented Adaptation (Section 3.2), and Target-dependent Classifier Adaptation (Section 3.3)) are deployed using PyTorch on a NVIDIA GTX 1080Ti GPU with 11 GB memory. In implementation, we use the Stochastic Gradient Descent (SGD) optimizer with Nesterov acceleration (the momentum is set as 0.9, weight decay is $10e-4$). The learning rate is $25e-5$ and is decayed by a damping policy using the polynomial decay with power of 0.9. All models are trained with $1e5$ iterations.

3.4.2 Computational cost

This section discusses the computation cost during training of each model proposed in **ESDA**. Specifically, Figure 3.4.1 compares model accuracy for semantic segmentation

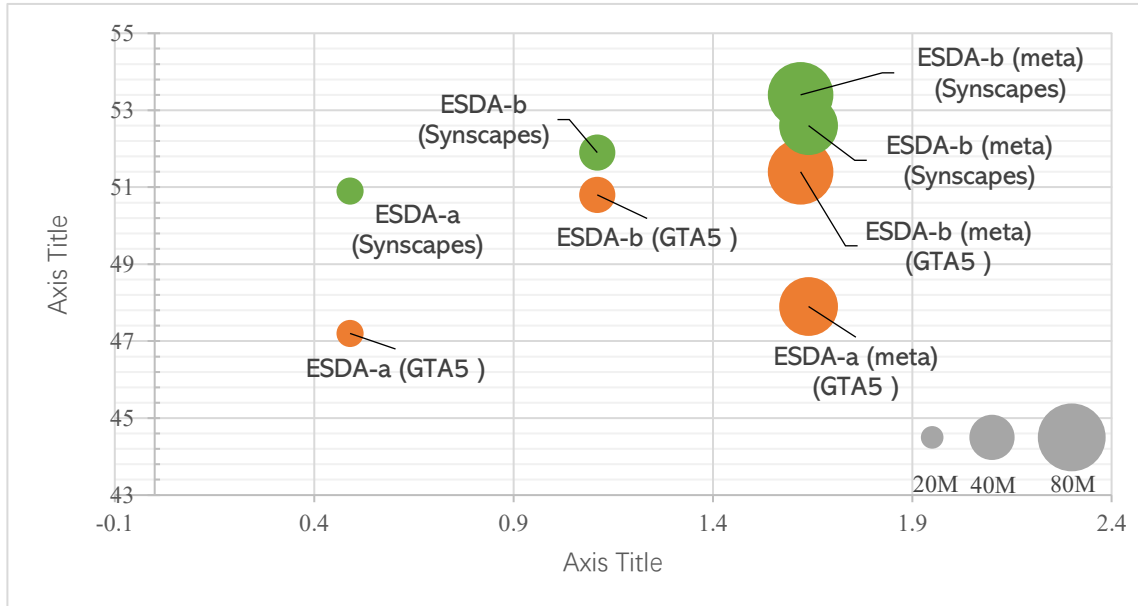


Figure 3.4.1. Comparison of training computational cost for different implementations. **Orange** circles stand for the models trained on GTA5 [65] and **green** ones on Synscapes [66]. Scales for 20M, 40M, and 80M parameters are displayed as reference.

(Mean IoU), training iteration time (sec/iteration), and number of parameters (bubble area). **ESDA-a** and **ESDA-b** correspond to the Pixel-Wise Adversarial Adaptation model proposed in Section 3.1 and the Prototypical Knowledge Adaptation model proposed in Section 3.2, respectively. First, benefiting from using a compact model structure as proposed in Section 3.1, **ESDA-a** is the fastest model, making it the most computationally efficient. However, this efficiency comes at the cost of lower accuracy, e.g., 47.2% in mIoU on the GTA5 dataset. Second, the **ESDA-b** model (proposed in Section 3.2) provides a balance between computational time and accuracy. The meta-learning implementation (**ESDA-a (meta)** and **ESDA-b (meta)**) that is formulated in Section 3.3 achieves further improvements in mIoU over both datasets but with longer training iteration time, indicating higher computational overhead compared to their “non-meta” counterparts. This analysis highlights the trade-off between accuracy and computational cost that can be achieved with the proposed approaches, suggesting that the choice of implementation depends on the specific application requirements.

3.4.3 Additional experiments

In this section, the three proposed **ESDA** frameworks are evaluated on more challenging real-world adaptation settings. NTHU [63] is a real-world dataset, which consists of four sets of real street scene images with high resolution (2048×1024) from four cities: Rio, Rome, Tokyo, and Taipei. It is considered to form four target domains, each of which has 3,200 images, providing a supplementary test case for the proposed **ESDA** frameworks.

In this experiment, the training set of Cityscapes [21] (with labels) is considered as the single source domain while the model is optimized to learn a segmenter specific to a city, e.g. Rio, Rome, Tokyo, or Taipei. Experimental results in terms of mean IoU on the

semantic segmentation task “Cityscapes to real cities” are summarized in Table 3.4.1. First, the baseline source-dependent segmenter is trained while fully depending on data from Cityscapes. It can be observed that such a source-dependent segmenter suffers from poor segmentation performance on the four real cities, reflecting the challenge introduced by domain shifts in real-world data. Second, two state-of-the-art works are considered as comparatives in the experiments. ASNet [60] employs adversarial learning to align high dimensional features on the soft-max layer. MaxSquare [105] raises the probability imbalance problem in conventional entropy-minimization domain adaptation and proposes a maximum squares loss to balance the gradient of easy-to-transfer target samples. Results in Table 3.4.1 show that both ASNet and MaxSquare surpass the baseline model by a clear margin, e.g., ASNet achieves a minimum of 2.7% improvement on mean IoU over the four cities. While comparing to the state-of-the-art works, both **ESDA-a** (Section 3.1) and **ESDA-b** (Section 3.2) models consistently improve over the baseline. Finally, the meta-learning implementations **ESDA-a** (meta) and **ESDA-b** (meta) (Section 3.3) further enhance performance in multiple cities (e.g., “ROME”, “TOKYO”, and “TAIPEI”). These results validate the efficacy of the proposed methods on diversified adaptation scenarios with substantial domain shift.

CITYSCAPES TO REAL CITIES			
CITY	Method	mIoU	gain
ROME	Source-dependent Segmenter	50.9	-
	ASNet [60]	53.8	2.9
	MaxSquare [105]	54.1	3.2
	ESDA-a	52.4	1.5
	ESDA-b	55.7	4.8
	ESDA-a (meta)	54.0	3.1
	ESDA-b (meta)	56.3	5.4
RIO	Source-dependent Segmenter	48.2	-
	ASNet [60]	51.6	2.6
	MaxSquare [105]	52.7	4.5
	ESDA-a	52.4	4.2
	ESDA-b	53.6	5.4
	ESDA-a (meta)	51.9	2.7
	ESDA-b (meta)	54.7	5.5
TOKYO	Source-dependent Segmenter	47.7	-
	ASNet [60]	49.9	2.2
	MaxSquare [105]	50.4	2.7
	ESDA-a	50.0	2.3
	ESDA-b	52.9	5.2
	ESDA-a (meta)	51.1	3.4
	ESDA-b (meta)	54.8	7.1
TAIPEI	Source-dependent Segmenter	46.5	-
	ASNet [60]	49.1	2.6
	MaxSquare [105]	51.0	4.5
	ESDA-a	52.6	6.1
	ESDA-b	53.1	6.6
	ESDA-a (meta)	52.9	6.4
	ESDA-b (meta)	54.1	7.6

Table 3.4.1. Experimental results for the semantic segmentation task “Cityscapes to Real Cities”. The table compares the performance of different methods in terms of mean Intersection over Union (mIoU) and the performance gain achieved.

Chapter 4. Enhanced Graph Domain Adaptation for Skeleton Videos

In Chapter 3, multiple methods for Enhanced Scene Domain Adaptation (**ESDA**) are introduced to deal with scene domain shift in FCN-based image semantic segmentation tasks. In contrast, Chapter 4 provides details on original contributions for Enhanced Graph Domain Adaptation (**EGDA**), which focuses on human action recognition using skeleton sequences. **EGDA** is built upon the key observation that Graph Convolutional Networks (GCNs) have significantly advanced state-of-the-art human action recognition. However, the deployment of trained GCN models is highly sensitive to graph domain shifts, including variations in camera viewpoints, human poses, sequence lengths, and joint permutations. In this context, and to address these challenges, Chapter 4 details the methodology of **EGDA** and experimental results with four proposed adaptation frameworks, respectively, Cross-view Adaptation, Cross-sensor Adaptation, Cross-sequence Adaptation and Cross-permutation Adaptation.

Section 4.1 on **Cross-View Adaptation** consists of a journal paper published in *Algorithm* in 2024 [106]. Section 4.2 on **Cross-Sensor adaptation** is a conference paper published at *CRV* 2024 [107]. Section 4.3 on **Cross-Sequence Adaptation** consists of a conference paper published at *AAAI* 2025 [108], and Section 4.4 on **Cross-Permutation Adaptation** is a conference paper submitted to *ICCV* 2025 which is under review. For the sake of clarity, the contents of each paper are briefly summarized in the aforementioned sections, emphasizing the corresponding original contributions. The publications (manuscripts) are then presented as published (submitted). At last, Section 4.5 presents additional experimental results that are not included in the papers.

4.1 Cross-View Adaptation

4.1.1 Summary

In practice, variations in camera view setups will cause domain shift to skeletal data related to orientation, scale, and the human subject’s pose. Cross-view shift can lead to significant intra-class discrepancy, where an action recognition model will suffer from extracting discriminative action patterns, consequently compromising the model in environments where observation views are changed. In this context, **EGDA** first introduces a cross-view adaptation framework that focuses on the practical deployment of a human action recognition model while processing skeletal data variations in between extrinsic environment-specific characteristics, such as camera view, orientation, height, and distance from the subject.

In implementation, this first component of **EGDA** introduces an unsupervised temporal-domain adaptation framework for human action recognition from skeleton-based data that combines Contrastive Prototype Learning (CPL) and Temporal Adaptation Modeling (TAM), with the aim of transferring the knowledge learned from a source domain to an unlabeled target domain. The CPL strategy, inspired by recent success in contrastive learning applied to skeleton data, learns a compact temporal representation from the source domain, from which the TAM strategy leverages the capacity for self-training to adapt the representation to a target application domain using pseudo-labels.

4.1.2 Original contributions

The contributions of this component of **EGDA** consist of the following:

- It investigates the core issues associated with skeletal data distribution shifts during model deployment, particularly focusing on cross-view adaptation. This

ensures that human action recognition models trained in controlled environments can effectively handle real-world variability;

- The proposed method requires only labeled data from the source domain and unlabeled data from the target domain for model adaptation, significantly reducing the effort invested in data annotation in the target domain;
- Experimental results demonstrate the effectiveness of the proposed strategy in comparison with fully supervised learning that requires voluminous labeled data from both the source and target domains.

4.1.3 Article

Research results have been presented in “Unsupervised Temporal Adaptation in Skeletal-Based Human Action Recognition” (H. Tian and P. Payeur), which was published in *Algorithms*, 2024 [106]. The online version can be found at <https://www.mdpi.com/1999-4893/17/12/581>. The original paper, published under the Creative Commons Attribution (CC BY) license where authors retain the copyright, is provided below as published. This paper is also a significantly extended version of a previous paper “Semi-Supervised Adaptation for Skeletal-Data-Based Human Action Recognition” (H. Tian, P. Payeur), which was published in *Engineering Proceedings*, 2023 [109] which is not included here.

Article

Unsupervised Temporal Adaptation in Skeleton-Based Human Action Recognition

Haitao Tian  and Pierre Payeur * 

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada; htian026@uottawa.ca

* Correspondence: ppayeur@uottawa.ca

Abstract: With deep learning approaches, the fundamental assumption of data availability can be severely compromised when a model trained on a source domain is transposed to a target application domain where data are unlabeled, making supervised fine-tuning mostly impossible. To overcome this limitation, the present work introduces an unsupervised temporal-domain adaptation framework for human action recognition from skeleton-based data that combines Contrastive Prototype Learning (CPL) and Temporal Adaptation Modeling (TAM), with the aim of transferring the knowledge learned from a source domain to an unlabeled target domain. The CPL strategy, inspired by recent success in contrastive learning applied to skeleton data, learns a compact temporal representation from the source domain, from which the TAM strategy leverages the capacity for self-training to adapt the representation to a target application domain using pseudo-labels. The research demonstrates that simultaneously solving CPL and TAM effectively enables the training of a generalizable human action recognition model that is adaptive to both domains and overcomes the requirement of a large volume of labeled skeleton data in the target domain. Experiments are conducted on multiple large-scale human action recognition datasets such as NTU RGB+D, PKU MMD, and Northwestern–UCLA to comprehensively evaluate the effectiveness of the proposed method.

Keywords: unsupervised domain adaptation; human action recognition; contrastive learning; human skeleton data



Citation: Tian, H.; Payeur, P.

Unsupervised Temporal Adaptation in Skeleton-Based Human Action Recognition. *Algorithms* **2024**, *17*, 581. <https://doi.org/10.3390/a17120581>

Academic Editors: Marcin Iwanowski and Bogusław Cyganek

Received: 31 October 2024

Revised: 6 December 2024

Accepted: 11 December 2024

Published: 16 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Skeleton-based human action recognition has been widely exploited for privacy-sensitive computer vision applications, such as human activity analysis [1], interactive gaming [2], and video surveillance [3]. Human actions can be represented by skeleton model videos [4] that encode the trajectories of skeleton joints captured by specialized acquisition systems, offering the merits of being agnostic to the environment and providing compact data representation. Graph Convolution Networks (GCNs) have dominated the recent research [5–7] on vision-based human action recognition by demonstrating their pivotal competence in aggregating action dynamics from spatial–temporal skeleton topologies and realizing end-to-end sequence-wise action classification.

Even though training a powerful GCN model for action recognition from large-scale public datasets [8,9] is promising, the transposition of the resulting model to applications in real-world environments remains challenging. For instance, the data distribution in a target deployment environment can deviate partially from that of the source training domain due to (i) the particular image acquisition configuration adopted or (ii) variability between subjects performing the same action. This consideration refers to the data domain covariate issue [10]. A common strategy consists of conducting fine-tuning rounds with full supervision using data collected in the target image acquisition configuration to reduce data discrepancy between the source domain and the target domain. However, the collection and annotation of a large volume of skeleton data for fine-tuning is extremely time consuming.

This work proposes an unsupervised domain adaptation (UDA) approach for skeleton-based human action recognition. It aims to utilize copious data from a controlled source domain, e.g., public datasets or laboratory environments, to pre-train an initial model. The approach then refines a target-adaptive action recognition model to deploy in a real-world (uncontrolled) domain with a fine-tuning phase using unlabeled data from the latter domain. UDA is achieved in two phases: Contrastive Prototype Learning (CPL) and Temporal Adaptation Modeling (TAM). Specifically, CPL utilizes supervised contrastive learning [11] to model discriminative action prototypes on trimmed data from the source domain. TAM formulates a self-training layer on the top of action prototypes. Since data in the target domain are unlabeled, TAM first leverages the pre-learned action prototypes to generate pseudo-labels on the target domain. Second, using both pseudo-labeled target data and labeled source data, it trains an action classification model that is adaptive to both domains. The contributions of this work consist of the following three elements:

- An original domain adaptation framework allows the action recognition network to initially learn human action knowledge from a labeled data domain, e.g., a public dataset or a controlled laboratory environment, to adapt to a target application domain;
- The proposed method requires only labeled data from the source domain and unlabeled data from the target domain for model adaptation, significantly reducing the effort invested in data annotation in the target domain;
- The proposed method is generalizable to state-of-the-art network architectures for unsupervised domain adaptation in action recognition without additional constraints.

2. Related Work

Skeleton data encode the dynamics of human movement over time by depicting the trajectories of the human body's skeleton joints. In this context, a human action can be represented by a sequence of skeleton frames that is trimmed to cover only the series of body movements that compose a specific action from the beginning to the end. By benchmarking the human action recognition task on a variety of large-scale skeleton-based datasets such as NTU RGB+D [8] and the PKU Multi-Modality Dataset [9], Graph Convolution Networks (GCNs) have attracted significant attention in recent works [5–7]. The concept of a GCN is to construct a graph upon skeleton frames in which each node corresponds to a human body joint and the edges correspond to the spatial connectivity among the joints, allowing the network to interpret the topological features of skeleton data. For instance, Li et al. [12] introduce actional-structural directed graph neural networks (AS-GCNs) to model the temporal dependencies among skeleton joints. Yan et al. [5] define spatial-temporal graph convolutional networks (ST-GCNs) to capture both spatial and temporal information in skeleton data, achieving competitive results in action recognition. Chen et al. [6] propose a framework (CTR-GCN) with the goal of interpreting relationships between joints.

However, the generalization of action recognition models in real environments is still challenging [13] due to the data distribution shift caused by variations in sensory configuration, e.g., camera views, heights, orientations, locations, and variations in data collection [8]. Skeleton contrastive learning [7,14] is an effective way to overcome the need for annotation on large volumes of data while demonstrating pivotal competency in learning invariant representations from unlabeled skeleton data for downstream tasks such as action recognition. These methods involve skeleton augmentation while constructing self-supervised contrastive learning pretext tasks, such as similarity measurement [15], which contrast positive pairs against negative pairs from a pre-defined dynamic dictionary [16,17].

3. Method

3.1. Preliminary

3.1.1. Skeleton-Based Action Recognition

A skeleton sequence, $X \in \mathbb{R}^{T \times V \times 3}$, consists of T frames in the shape of V human body joints, each one being defined in a calibrated camera reference frame with three-dimensional coordinates. A skeleton-based action recognition model F (Equation (1)) can be

decoupled into a feature subnetwork, $\mathcal{M} : \mathbb{R}^{T \times V \times 3} \rightarrow \mathbb{R}^{T \times V \times C_{fea}}$, which parses skeleton joint topologies from the input sequence X via spatial and temporal graph convolutions [5], and a task subnetwork, $\mathcal{C} : \mathbb{R}^{T \times V \times C_{fea}} \rightarrow [0, 1]^{L \times 1}$, which aggregates actions' semantic representations into a SoftMax output space where the multi-class cross-entropy loss is involved in optimization:

$$L_{CE} = -Y \cdot \log \left[\underbrace{\left(\mathcal{C} \circ \mathcal{M} \right)}_F (X) \right] \quad (1)$$

where Y is the sequence-wise action label of X , while $Y \in \mathbb{R}^L$ is defined in a range of L categories.

3.1.2. Domain Adaptation in Skeleton Data

However, end-to-end supervised learning on F requires the collection and annotation of a large volume of skeleton sequences $\{X, Y\}$, which is extremely time-consuming in practice. To circumvent the need for such data, this work proposes Unsupervised Domain Adaptation (UDA) based on two fundamental considerations:

- (1) The action recognition model, $F = \mathcal{C} \circ \mathcal{M}$, is modular such that its subnetworks ($\mathcal{M}$ and $\mathcal{C}$) can be treated individually and separately, which allows an adaptation framework to optimize them step by step.
- (2) In practical engineering applications, a target deployment environment (target domain $\mathcal{D}_t$) often presents misaligned data distributions, given the fact that the image acquisition configuration can differ from that of the original data collection domain (source domain $\mathcal{D}_s$). At the same time, individuals coming from different environments performing the same action may have significant variations in their movement patterns, leading to another form of discrepancy in data distribution. Either data domain shift (i.e., the data distribution variations across $\mathcal{D}_s$ and $\mathcal{D}_t$) tends to corrupt the model's performance when the model is well trained on $\mathcal{D}_s$ but is evaluated on $\mathcal{D}_t$.

To this end, UDA considers a source domain, i.e., $\mathcal{D}_s = \{(X_s, y_s)\}$, which is composed of the sequence X_s and its annotation y_s , and a target domain, $\mathcal{D}_t = \{X_t\}$, where the skeleton sequence X_t is unlabeled. UDA aims to utilize a significant quantity of labeled data from $\mathcal{D}_s$ to pretrain an intermediate model (e.g., feature encoder) and then refine a target-adaptive model F by using only unlabeled data from the target environment. By transferring skeleton action knowledge across different domains, UDA effectively avoids learning from scratch and saves the labor of annotating videos in the target domain.

Figure 1 illustrates the proposed framework. **Contrastive Prototype Learning (CPL)** optimizes the subnetwork $\mathcal{M}$ over the labeled data from $\mathcal{D}_s$. The pre-learned $\mathcal{M}$ is recycled into a successive **Temporal Adaptation Modeling (TAM)** phase that models temporal adaptation knowledge (pseudo-labels) on the unlabeled data from the target domain. Finally, the proposed **CPL** and **TAM** models are submitted to an optimization phase to progressively refine an action classifier.

3.2. Contrastive Prototype Learning (CPL)

In conventional skeleton-based human action recognition, a classification network receives a skeleton sequence X where its feature subnetwork (encoder), $\mathcal{M} : \mathbb{R}^{T \times V \times 3} \rightarrow \mathbb{R}^{T \times V \times C_{fea}}$, is responsible for parsing a spatiotemporal skeleton representation of X into a high-dimensional feature space where hidden states are closely clustered into *action prototypes* such that the task subnetwork (classifier), $\mathcal{C}$, is able to determine action-wise classification boundaries. Based on these observations, we propose to use Contrastive Prototype Learning (CLP) to learn *action prototypes* by optimizing $\mathcal{M}$ using skeleton data from the source domain $\mathcal{D}_s$.

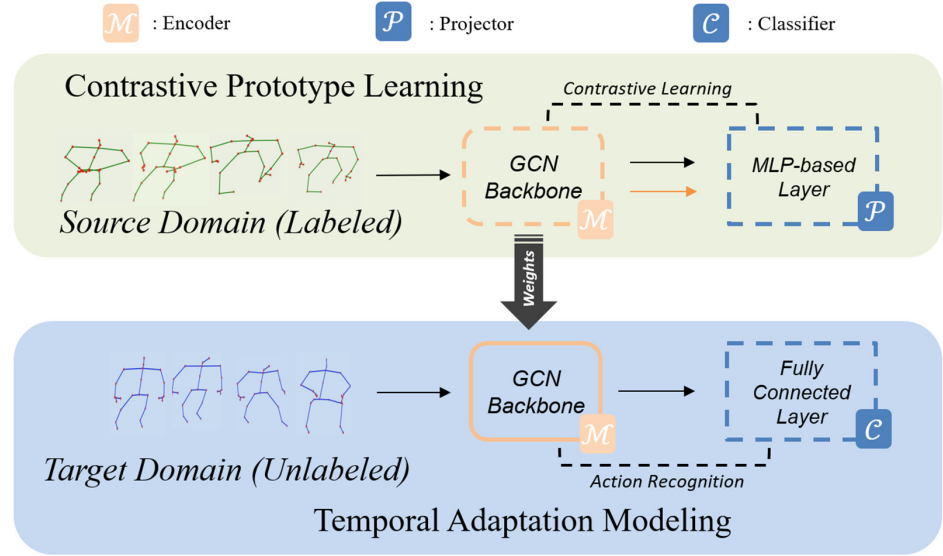


Figure 1. Unsupervised domain adaptation (UDA) framework combining the proposed CPL and TAM strategies. In CPL, the training data (labeled) from the source domain supports supervised representation learning. The learned backbone is reused in TAM for refining over data samples (initially unlabeled) from the target domain. Network embedding is denoted with dotted lines if it is updated during training and with solid lines otherwise.

Specifically, **CPL** utilizes a projection layer, $\mathcal{P} : \mathbb{R}^{T \times V \times C_{fea}} \rightarrow \mathbb{R}^{C_{prj}}$, on top of the feature subnetwork $\mathcal{M}$ to implement contrastive learning. Recent research has utilized a variety of data augmentation schemes [7,14] to operate skeleton-data-based contrastive learning. In contrast, **CPL** opts out of data augmentation and leverages action labels in contrastive learning. Since data labels carry strong supervisory signals, **CPL** helps learn explicit and discriminative action prototypes.

Let $\mathcal{B}_S = \left\{ \left(X_s^{(i)}, y_s^{(i)} \right) \right\}_{i=1}^N$ be a data batch from the source domain $\mathcal{D}_S$, where $N \ll |\mathcal{D}_S|$. Let $Q = \{1, \dots, N\}$ be the set of instance indices of $\mathcal{B}_S$. The data batch considers a semantically identical set $P(i)$ with respect to the anchor $X_s^{(i)}$, in which each element $p \in P(i)$ indices a sequence $X_s^{(p)}$ which has the same sequence label as the anchor, i.e., $P(i) = \left\{ p \in A(i) : y^{(p)} = y^{(i)} \right\}$, where $A(i) \in Q \setminus \{i\}$ indicates all elements in Q except for i . We call $X_s^{(p)}$ a *positive* counterpart of the anchor. In this way, the representation pair $(z_s^{(i)}, z_s^{(p)})$ consists of a dynamic skeleton dictionary from which **CLP** can learn underlying feature invariances by $z^i = (\mathcal{P} \circ \mathcal{M})(X_s^{(i)})$ and $z^{(p)} = (\mathcal{P} \circ \mathcal{M})(X_s^{(p)})$. In order to map semantically similar sequences to action prototypes, **CLP** uses a SupCon loss [11] to minimize the distance between feature embeddings of anchor and positives:

$$L_{\text{SupCon}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z^i \cdot z^{(p)} / \tau)}{\sum_{n \in \mathcal{N}} \exp(z^i \cdot n / \tau)} \quad (2)$$

where τ acts as the temperature parameter and n represents a negative counterpart sampled from a memory bank $\mathcal{N}$ (we follow the same implementation as in [14] on the construction of the memory bank).

3.3. Temporal Adaptation Modeling (TAM)

Upon the optimization of effective action prototypes, a target domain sequence $X_t^{(i)}$ can be encoded as prototypical distributions according to $\mathcal{M}(X_t^{(i)})$. However, since the target domain is unlabeled, it is intractable to directly learn an action classifier $\mathcal{C}$ using end-to-end supervised learning. At the same time, given the domain shift between the source domain and the target domain, directly learning a classifier exclusively on the source domain will lead to a biased classifier that compromises the testing performance on the target domain.

TAM proposes an episodic optimization paradigm that alternates a supervised learning iteration on the source domain and a self-training iteration on the target domain.

3.3.1. Supervised Learning on $\mathcal{D}_s$

It learns a classifier, $\mathcal{C} : \mathbb{R}^{T \times V \times C_{fea}} \rightarrow [0, 1]^{L \times 1}$, over the data sample $(X_s^{(i)}, Y_s^{(i)}) \in B_s$. The learning process is optimized by a typical multi-class cross-entropy loss using

$$L_{CE-s} = -Y_s^{(i)} \cdot \log((\mathcal{C} \circ \overline{\mathcal{M}})(X_s^{(i)})) \quad (3)$$

where $\overline{\mathcal{M}}$ denotes the action prototypes learned by Equation (2) which is fixed at this step.

3.3.2. Self-Training on $\mathcal{D}_t$

After training the model on the labeled source data, it is used to predict pseudo-labels for the unlabeled target domain data. Specifically, let $X_t^{(i)}$ be data sample from the target domain data batch $\mathcal{B}_t = \{X_t^{(i)}\}_{i=1}^N$ and $p_t^{(i,l)}$ be the probabilistic distribution of $X_t^{(i)}$ at category l where $p_t^{(i,l)} = ((\mathcal{C} \circ \mathcal{M})(X_t^{(i)}))^{(l)}$, we treat $\hat{Y}_t^{(i)}$ as the pseudo-label of $X_t^{(i)}$, which can be obtained by

$$\hat{Y}_t^{(i)} = \begin{cases} l, & \text{if } l = \arg \max_l p_t^{(i,l)}, \text{ and } p_t^{(i,l)} > T \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where T denotes the confidence threshold. Only predictions with confidence scores above T are considered valid pseudo-labels. This minimizes the inclusion of noisy labels. Therefore, we use $\hat{Y}_t^{(i)}$ in the self-training optimization:

$$L_{CE-t} = -\hat{Y}_t^{(i)} \cdot \log((\mathcal{C} \circ \overline{\mathcal{M}})(X_t^{(i)})) \quad (5)$$

The pseudo-labels help the model learn to recognize target domain patterns, gradually improving its performance. Pseudo-labels are refined in each iteration. As the model becomes more accurate, it generates better pseudo-labels, leading to a refinement of the model's optimization.

3.4. Optimization

The proposed UDA network (illustrated in Figure 1) consists of three components: an encoder $\mathcal{M}$, a projector $\mathcal{P}$, and a classifier $\mathcal{C}$. The network training is carried out by the interaction among (2), (3), and (5), which are alternatively optimized (Figure 2) according to the stages below:

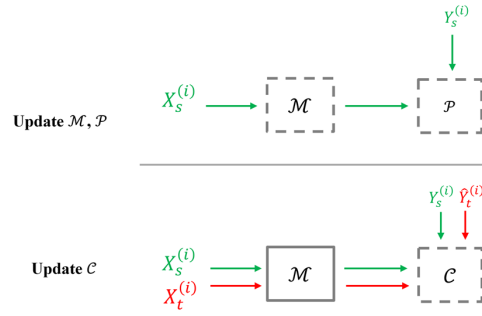


Figure 2. Tensor flow of the training pipeline. Green symbols relate to the source domain, while red ones relate to the target domain.

- **CPL ($\mathcal{M}, \mathcal{P}$) updating.** The CPL model is initially trained with supervised contrastive learning utilizing labeled source domain data $(X_s^{(i)}, Y_s^{(i)})$. Parameters in $\mathcal{M}$ and $\mathcal{P}$ are updated by minimizing the loss function (2) as follows:

$$\min_{\mathcal{M}, \mathcal{P}} \sum_{(X_s^{(i)}, Y_s^{(i)}) \in B_s} L_{\text{SupCon}}(X_s^{(i)}, Y_s^{(i)}; \mathcal{M}, \mathcal{P}) \quad (6)$$

- **TAM ($\mathcal{C}$) updating.** The classifier $\mathcal{C}$ is updated in an iterative refinement scheme.
 - (1) Data and domain labels $(X_s^{(i)}, Y_s^{(i)})$ are used to update $\mathcal{C}$ to improve semantic classification. It is achieved by minimizing the loss functions (3) as in (7), where λ is the trade-off weight used to balance the **TAM** training. At this step, $\overline{\mathcal{M}}$ is fixed.

$$\min_{\mathcal{C}} \left[(1 - \lambda) \sum_{(X_s^{(i)}, Y_s^{(i)}) \in B_s} L_{\text{CE-s}}(X_s^{(i)}, Y_s^{(i)}; \overline{\mathcal{M}}, \mathcal{C}) \right] \quad (7)$$

- (2) The pre-optimized $\mathcal{C}$ is utilized to obtain the pseudo-label $\hat{Y}_t^{(i)}$ of $X_t^{(i)}$ based on Equations (4) and (5) as in Equation (8). Note that $\overline{\mathcal{C}}$ and $\overline{\mathcal{M}}$ are fixed at this step.

$$\max_{\hat{y}_t^{(i)}} L_{\text{CE-t}}(X_t^{(i)}, \hat{Y}_t^{(i)}; \overline{\mathcal{M}}, \overline{\mathcal{C}}) \quad (8)$$

- (3) Afterwards, data and pseudo-labels $(X_t^{(i)}, \hat{Y}_t^{(i)})$ are used to update $\mathcal{C}$.

$$\min_{\mathcal{C}} \left[\lambda \sum_{X_t^{(i)} \in B_t} L_{\text{CE-t}}(X_t^{(i)}, \hat{Y}_t^{(i)}; \overline{\mathcal{M}}, \mathcal{C}) \right] \quad (9)$$

Steps (1), (2), and (3) are implemented in each iteration of the **TAM** stage.

4. Experiments

Experimentation is conducted to evaluate the effectiveness of the proposed method in cross-domain human action recognition scenarios. Moreover, experimental ablation studies examine best practices related to the proposed framework, facilitating its reproduction.

4.1. Datasets

NTU RGB+D [8] is a popular large-scale skeleton-form human action recognition dataset. It presents 56,880 samples covering 60 human daily actions recorded in indoor scenes with three Microsoft Kinect V2 cameras mounted in different locations to support “cross-view” (C-View) variations. It also involves “cross-subject” (C-Sub) variations by

involving 40 actors in action performance. This dataset encoded skeleton representations over 25 joints.

PKU-MMD [9] is another popular public skeleton dataset presenting fewer data samples. For instance, PKU-MMD Part I has 21,545 class-specific sequences covering 51 human daily activities (41 single-person actions and 10 two-person actions), while PKU-MMD Part II contains 7000 class-specific sequences covering the 41 single-person actions. This dataset was also recorded with three Microsoft Kinect V2 devices, which provided skeleton representations encoded over 25 joints. Compared to NTU RGB+D, it involves significant camera view variations in the samples, which is used to mimic practical environments where data collection configurations can be inconsistent. The dataset also involves C-View and C-Sub variations.

Northwestern–UCLA [18] is a smaller 3D skeleton dataset composed of 1494 short sequences covering 10 human actions. This dataset was recorded with one Microsoft Kinect V1, which provided skeleton representations encoded over 20 joints. The discrepancy on the number of joints with respect to PKU-MMD acquisition configuration provides an effective evaluation scenario for domain adaptation. The dataset also involves C-View and C-Sub variations.

4.2. Implementations

As components of the proposed network structure, we adopt ST-GCN [5] as the feature encoder ($\mathcal{M}$) due to its wide adoption in recent research. Second, the projector $\mathcal{P}$ is composed of a spatial pooling layer and a temporal pooling layer, followed by two linear layers equipped with a ReLU activation after the first linear layer. The classifier $\mathcal{C}$ is composed of a spatial pooling layer and a temporal pooling layer, followed by a linear layer and a SoftMax activation.

The proposed framework in Figure 1 is deployed using PyTorch on a NVIDIA RTX 3090 GPU. For the **CPL** phase, we adopt the same hyperparameter values as in previous work [14], i.e., C_{prj} is set as 128, τ as 0.07, momentum value as 0.999. The size of $\mathcal{N}$ is set as 32,768 for the source trimmed dataset NTU RGB+D and 16,384 for PKU MMDv1. The model is updated by using SGD with weight decay 0.0001 and learning rate 0.01. In the **TAM** phase, we use SGD to update the network with Nesterov momentum 0.9, weight decay 0.0005, and learning rate 0.05. We also use cosine annealing for the learning rate scheduler.

We use joint Normalization in data preprocessing. First, we establish a body frame system: (i) define the vector from the “spine” to the “left shoulder” as the x -axis; (ii) define the vector from the “spine” to the “spine base” as the y -axis; (iii) define the z -axis as the cross product of the x -axis and the y -axis. Second, we normalize the skeleton sequences by subtracting the trajectory of the “spine” joint.

4.3. Evaluation Protocols

First, multiple UDA tasks are considered to validate the performance of the proposed framework. In each UDA task, a pair of datasets, one labeled and the other unlabeled, are involved to represent the skeleton data distribution shift, i.e., the “C-View” evaluation protocol is used to represent the variation in image acquisition configuration and the “C-Sub” evaluation protocol is used to represent subject variations across domains.

- **NTU RGB+D to PKU-MMD.** We consider a single source domain, NTU RGB+D, and two target domains, PKU-MMD part I and PKU-MMD part II. In the first experimentation “NTU RGB+D to PKU-MMD part I” (Table 1), skeleton sequences under 50 actions common to both datasets are used for training and testing, while in “NTU RGB+D to PKU-MMD part II” (Table 2), 41 single-person actions are used for experiments.

Table 1. Experimental results on unsupervised domain adaptation (test case 1).

NTU RGB+D to PKU-MMD Part I	Source				Target			
	C-Sub		C-View		C-Sub		C-View	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Source only	69.07	41.83	72.86	44.59	57.32	27.40	59.02	29.09
UDA	70.05	45.09	72.45	46.34	69.34	37.35	72.17	36.12
Target only	34.41	18.67	35.30	17.36	85.06	49.21	87.34	52.18

Table 2. Experimental results on unsupervised domain adaptation (test case 2).

NTU RGB+D to PKU-MMD Part II	Source				Target			
	C-Sub		C-View		C-Sub		C-View	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Source only	75.34	38.46	80.98	39.21	23.58	10.06	22.46	9.5
UDA	74.15	36.53	80.02	38.50	30.55	12.42	31.29	14.53
Target only	35.23	12.35	32.73	11.45	45.91	24.68	47.58	25.39

- **NTU RGB+D to NW-UCLA.** This evaluation considers NTU RGB+D as the source domain and Northwestern–UCLA with 20-joint skeletons as the target domain (Table 3). The skeletons from the NTU RGB+D are preprocessed to remap from 25 to 20 joints. The datasets share seven common actions, i.e., “drop”, “pick up”, “throw”, “sitting down”, “standing up”, “wear jacket”, and “take off jacket”, that form a label space.

Table 3. Experimental results on unsupervised domain adaptation (test case 3).

NTU RGB+D to NW-UCLA	Source				Target			
	C-Sub		C-View		C-Sub		C-View	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Source only	88.39	49.92	93.03	65.28	65.36	34.65	69.42	37.61
UDA	87.74	43.80	90.57	64.63	72.69	42.19	76.44	40.53
Target only	44.31	23.11	47.49	30.12	89.24	43.23	92.06	65.28

- **(PKU-MMD) Part I to Part II.** We consider the trimmed PKU-MMD Part I as the source domain and the untrimmed PKU-MMD Part II as the target domain (Table 4). The skeleton sequences under 41 common single-person actions are used for training and testing.

Table 4. Experimental results on unsupervised domain adaptation (test case 4).

(PKU-MMD) Part I to Part II	Source				Target			
	C-Sub		C-View		C-Sub		C-View	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Source only	85.06	39.92	87.34	42.96	27.43	14.28	29.07	14.28
UDA	85.21	41.01	34.41	14.28	33.62	15.92	32.19	15.84
Target only	19.40	7.22	24.39	11.04	45.91	23.94	47.46	21.91

4.4. Results

Experimental results in terms of Top 1 accuracy (Acc) and F1 scores [8] are reported in Tables 1–4. First, we consider two baseline models, a source-only model and a target-only model, each of which is trained exclusively with supervision from a specific (source or target) domain by using the domain’s training set with annotations. Second, we trained our UDA model via the proposed **CPL** and **TAM** strategies.

The experimental results first demonstrate that either baseline model can yield convincing results in its original training domain (e.g., the source-only model in Table 1 trained on NTU RGB+D achieves 69.07% on the “C-Sub” testing set and 72.86% on the “C-View” testing set of NTU RGB+D, while the target-only model trained on PKU-MMD Part I achieves 85.06% and 87.34% on the testing sets of PKU-MMD Part I). However, given the domain shift between the source and the target domains, either baseline model fails when performing cross-domain action recognition (e.g., in Table 1 the source-only model (trained on NTU RGB+D) only achieves 57.32% on PKU-MMD Part I and similar results are observed for the target-only model when tested on NTU RGB+D data).

When the proposed UDA framework is used to learn a joint-domain classifier using the (labeled) training set of the source domain and the (unlabeled) training set of the target domain, the results in the second row of Tables 1–4 demonstrate that the resulting UDA model achieves significant improvement on cross-domain action recognition. It achieves a gain of 12.02% in Acc (from 57.32% to 69.34%) and 9.95% in F1 (from 27.40% to 37.35%) over the source-only model performance in Table 1. The UDA model still achieves comparable performance on the source domain compared to the source-only model in most of testing cases (e.g., 69.07% vs. 70.05% under C-sub in Table 1), demonstrating its effectiveness in learning action knowledge from two domains. However, given that target labels are absent, the UDA model expectedly underperforms the fully supervised (i.e., target-only) model. Similar performance is observed for every one of the four cross-domain test scenarios considered. Furthermore, even though the proposed UDA model obtains effective adaptation outcome under both the “C-View” and “C-Sub” evaluation protocols, it performs better in the “C-View” case, which we attribute to the different degree of skeleton data distribution shift.

4.5. Effects of Contrastive Prototype Learning

4.5.1. Training with Fewer Target Domain Samples

The critical contribution of Contrastive Prototype Learning (CPL) is the optimization of discriminative action prototypes via supervised contrastive learning using the source domain data. It is relevant to validate the transferability of the action prototypes that are learned in the source domain while only involving a subset of the target domain data for refinement. For this experiment, we first use CPL to learn an encoder $\mathcal{M}$ using the labeled data from NTU-RGB+D. Second, we fix the pre-trained $\mathcal{M}$ and use different proportions (varying from 5% to 100%) of labeled data from PKU MMD Part I to fine-tune a classifier $\mathcal{C}$. Note that this differs from the experiments reported in Section 4.4 that use only unlabeled data for model fine-tuning. Experimental results in Table 5 demonstrate that the model performance tends to increase monotonically with the ratio of target data samples involved but only until a saturation point. Interestingly, even using a very small number (e.g., 5%) of data samples from the target domain, the model still achieves compelling performance compared to the best model (85.06% in Table 1) when using 100% of the PKU-MMD Part I dataset.

Table 5. Performance (Top 1 accuracy) for different proportions of the target data samples in classifier C refinement training in contrastive prototype learning.

Percentage of Data Use	5%	10%	30%	50%	70%	100%
CPL	69.47%	72.65%	80.11%	83.86%	84.34%	85.06%

4.5.2. Comparison to State-of-the-Art Methods

The proposed method offers the advantage of remaining free of data augmentation for contrastive learning, unlike recently introduced approaches relying on a combination of skeleton augmentations. Therefore, we also include a comparison with similar studies. From the experimental results reported in Table 6, CPL demonstrates better Top1 accuracy

than the methods described in [19–21], revealing a promising adaptation performance and effectiveness in learning from skeleton data.

Table 6. Comparison to state-of-the-art works.

NTU RGB+D to PKU-MMD Part II	Top1 Acc
Zheng et al. [19]	44.8%
Lin et al. [20]	45.8%
Thoker et al. [21]	45.9%
CPL (proposed)	47.2%

4.5.3. Comparison to Vanilla Supervised Learning

An intuitive way to reach the same goal as **CPL** is to use the vanilla supervised pretraining paradigm. For instance, one can use a cross-entropy loss to train a regular action recognition model, $F = \mathcal{C} \circ \mathcal{M}$, using the labeled data from the source domain (Equation (1)), and then use the trained model $\mathcal{M}$ for transfer learning on the target domain (Equation (9)). This subsection presents two experiments to evaluate whether utilizing vanilla supervised learning can reach the same effectiveness as the proposed **CPL** strategy. First, we use conventional supervised method [5] and the proposed **CPL** method to train a comparative model F with full supervision from NTU RGB+D. Second, we fine-tuned $\mathcal{C}$ (with $\mathcal{M}$ fixed) using full supervision on 10% data samples from PKU-MMD Part I. Afterwards, we test the refined model on the test set of PKU-MMD Part I. The experimental results in Table 7 demonstrate that the fully supervised learning method offers limited transferability compared to the proposed **CPL** method. The fully supervised transfer learning only achieves 45.97% on the target domain, representing a 26.68% gap in accuracy compared to using **CPL**.

Table 7. Performance of “vanilla” fully supervised learning vs. semi-supervised learning. “NTU” stands for “NTU RGB+D”, and “PKU” stands for “PKU MMD Part I”.

	Full Supervision	CPL
	NTU and 10% PKU (Fine-Tuning)	NTU and 10% PKU (Fine-Tuning)
Accuracy	45.97%	72.65%

4.5.4. Visualization of Action Prototypes

For better understanding of the effectiveness of **CPL** in learning action prototypes, closer examination of the embedding features from the two domains is studied using t-SNE [22]. Figure 3 illustrates the action clusters of the two respective domains (NTU RGB+D and PKU-MMD) encoded on the last convolutional layer of $\mathcal{M}$ as interpreted by the two action recognition models (“vanilla” supervised pretrained on the upper row and “**CPL**” on the bottom row). First, even though the “vanilla” supervised pretrained model presents well-separated action clusters when tested on the source domain (upper left feature map), which reflects the model’s ability to interpret feature representations from the source domain, it presents less separable action clusters when tested on the target domain (upper right map) due to the impacts of domain shift. Such a discrepancy on action cluster interpretation leads to performance degradation across the two domains (from 69.07% to 57.32% in Table 1). Second, when applying the proposed **CPL** method, the model (bottom row) demonstrates more effective adaptation to the target domain whose action clusters are improved in terms of separability. Such improvement is observable on both test cases considering the source domain (bottom left map) and the target domain (bottom right map).

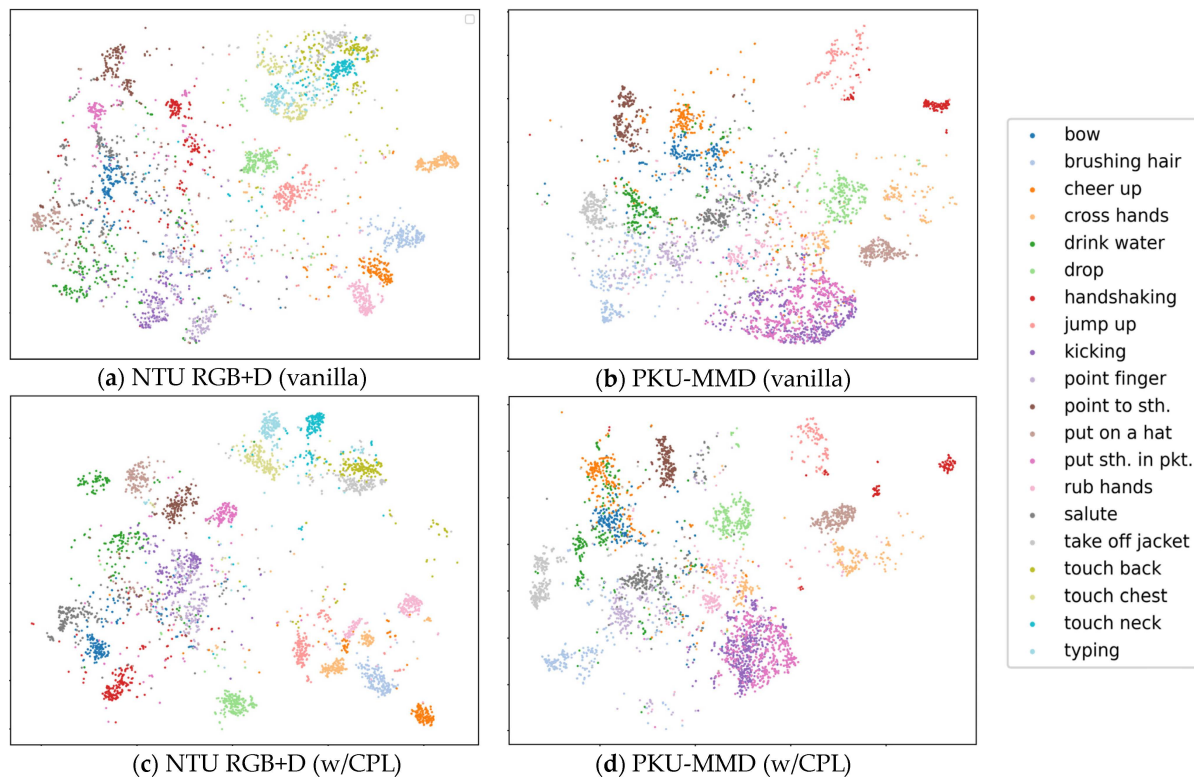


Figure 3. T-SNE visualization on action prototypes on the feature space of ST-GCN (upper row: “vanilla without CPL”; bottom row: “with CPL”). Action prototypes are distinguished by colors where twenty actions are randomly selected among fifty common actions for clarity. The left column represents action clusters of the source domain (NTU RGB+D), and the right column shows clusters of the target domain (PKU-MMD), respectively.

4.6. Ablation Study

We conduct ablation experiments to better understand the effectiveness of each component in the proposed framework.

4.6.1. Pseudo-Labels with TAM

In this work, **Temporal Adaptation Modeling (TAM)** is proposed to assist with integrating unlabeled target domain data into model training, allowing the joint-domain classifier to receive action knowledge from the target domain. To validate its effectiveness, we ablate the term L_{CE-t} proposed in Equation (9) (i.e., $\lambda = 0$ in Equation (9)) and use the exclusive source data in model training using Equation (7). Experimental results presented in Table 8 demonstrate a significant performance discrepancy on the two domains. Since only the source domain data is involved in model training, the resulting model is biased towards the source domain and achieves compromised performance in the target domain. The results demonstrate that **TAM** overcomes the limitation by using pseudo-labels and leads to effective adaptation to the target domain.

Table 8. The effectiveness of using pseudo-labels with TAM.

Models	Source	Target	Top1 Acc	Standard Deviation
without TAM	NTU RGB+D	-	59.35%	2.12%
with TAM	NTU RGB+D	PKU-MMD Part I	69.34%	1.70%

4.6.2. Hyperparameter Analysis for $\mathcal{T}$

In Equation (4), the parameter $\mathcal{T}$ acts as a threshold to distill reliable pseudo-labels for **TAM**. Setting a high value of $\mathcal{T}$ will seclude too many target data samples from model training and thus the source domain data will be dominant, giving rise to a biased training process. On the other hand, a low value of $\mathcal{T}$ may cause the inclusion of too many noisy labels from the target domain, thus corrupting the training process. To identify an appropriate value for $\mathcal{T}$, we alter $\mathcal{T}$ over the range [0.5, 1] in the **TAM** process while preserving other implementations parameters as defined in Section 4.2. As illustrated in Figure 4, the adaptation performance is affected as $\mathcal{T}$ varies. The best performance is achieved when $\mathcal{T}$ is around 0.85. Ultimately, when $\mathcal{T} = 1$, the model degrades to a source-only model (i.e., the “without TAM” model in Table 8).

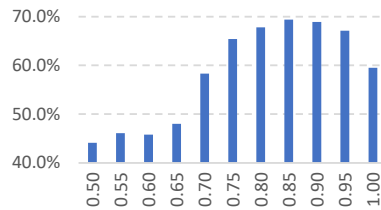


Figure 4. Performance (Top1 Acc) of TAM with varying hyperparameter $\mathcal{T}$.

4.6.3. Hyperparameter Analysis for λ

The implementation of **TAM** involves the determination of pseudo-labels on samples from the target domain and uses the parameter λ to weigh the importance of L_{CE-t} in Equation (9). Experimental investigation analyzes the selection of λ by conducting an experiment where the framework learns a classifier $\mathcal{C}$ with the same implementation detailed in Section 4.2 but altering λ over the range [0.1, 1]. As illustrated in Figure 5, the performance is clearly affected as λ varies with an eventual degradation in performance when the hyperparameter places too much weight on L_{CE-t} . Consequently, it is observed that the best performance is achieved when λ remains relatively low, that is, around 0.3.

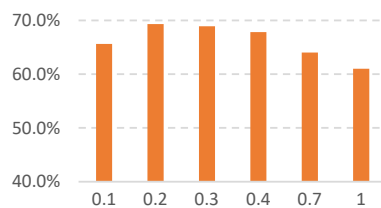


Figure 5. Performance (Top1 Acc) of TAM with varying hyperparameter λ .

5. Discussions

5.1. Future Work

As experimental results demonstrate in Tables 2 and 4, the model achieves limited adaptation results on the dataset PKU-MMD Part II. The latter presents significant domain shift compared to the source domain, NTU-RGB+D or PKU-MMD Part I. We conjecture that if the source and target domains are significantly different, initial pseudo-labels may be noisy, impacting the early training stages of TAM, thus leading to low accuracies in the target domain. In future works, it will be relevant to incorporate more efficient modules to tackle the challenge. We plan to use class-specific thresholds in pseudo-labeling. Specifically, in Equation (4), instead of using a hard threshold, $\mathcal{T}$, for all classes, we would rather use a class-specific one, $\mathcal{T}_k$, where we use the class-based statistic, e.g., sample frequency, to weight the confidence threshold for the action k . Such a strategy may help the model pay class-specific attention to optimization.

5.2. Real-World Application

One potential real-world application of the proposed method is to monitor patient movements in hospitals or elderly care institutions to detect abnormal activities. In such an environment, data distribution shifts can originate from the variability in data collection resulting from changing camera views, varying room configurations, and different subjects. The proposed method may help develop a robust action detection or recognition system without leveraging large amounts of data and labels, providing solutions for transfer learning in a variety of real-world environments.

5.3. Limitations

One limitation of the proposed approach is that the accuracy of pseudo-labels directly affects model performance. Low-quality pseudo-labels can propagate errors and degrade the optimization process. Techniques such as confidence-based filtering, adaptive thresholds, and ensemble predictions could be helpful to attenuate the impact of noisy labels.

6. Conclusions

In this paper, an original unsupervised domain adaptation (UDA) method is introduced to effectively adapt skeleton-based human action recognition to variable target environments without requiring labeled data from the latter. The proposed framework first leverages the benefits of contrastive prototype learning (CPL) to extract expressive action prototypes from a labeled source domain. Second, temporal adaptation modeling (TAM) refines a functional action recognition model to the target domain by associating and exploiting pseudo-labels. Experiments demonstrate the effectiveness of the proposed UDA strategy in comparison with fully supervised learning that requires voluminous labeled data from both the source and target domains. It also demonstrates the effectiveness of the proposed approach while bypassing the need for data augmentation in contrastive learning. Ablation studies suggest that the proposed CPL and TAM methods jointly contribute to the effectiveness of the overall model. Future work will study the use of a class-balanced pseudo-labeling scheme to ensure better representativity of training data from the target domain.

Author Contributions: Conceptualization, H.T. and P.P.; methodology, software, validation, formal analysis, H.T.; investigation, H.T. and P.P.; resources, P.P.; writing—original draft preparation, H.T.; writing—review and editing, P.P.; supervision, project administration, funding acquisition, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by MITACS Accelerate grant #IT29551 and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant #RGPIN-2020-04307.

Institutional Review Board Statement: This research only involved data from public resources, and no experiments on human subjects were conducted. Ethical review and approval were waived by the Office of Research Ethics and Integrity of the University of Ottawa because the project falls under Article 2.5 of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2) and therefore did not require Research Ethics Board review.

Informed Consent Statement: Not applicable.

Data Availability Statement: Information regarding public data sources is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kong, Y.; Fu, Y. Human action recognition and prediction survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [[CrossRef](#)]
2. Lange, B.; Chang, C.-Y.; Suma, E.; Newman, B.; Rizzo, A.S.; Bolas, M. Development and evaluation of low-cost game-based balance rehabilitation tool using the Microsoft Kinect sensor. In Proceedings of the 2011 annual international conference of the IEEE engineering in medicine and biology society, Boston, MA, USA, 30 August–3 September 2011; pp. 1831–1834.

3. Niu, W.; Long, J.D.; Han, D.; Wang, Y.-F. Human activity detection and recognition or video surveillance. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No. 04TH8763), Taipei, Taiwan, 27–30 June 2004; Volume 1, pp. 719–722.
4. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **2023**, *56*, 11. [\[CrossRef\]](#)
5. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
6. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
7. Chi, H.G.; Ha, M.H.; Chi, S.; Lee, S.W.; Huang, Q.; Ramani, K. InfoGCN: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196.
8. Shahroudy, A.; Jun, L.; Tian-Tsong, N.; Gang, W. NTU RGB+D: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
9. Liu, C.; Hu, Y.; Li, Y.; Song, S.; Liu, J. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, Mountain View, CA, USA, 23 October 2017; pp. 1–2.
10. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [\[CrossRef\]](#)
11. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020; Volume 33, pp. 18661–18673.
12. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
13. Choi, J.; Sharma, G.; Chandraker, M.; Huang, J.-B. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1717–1726.
14. Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; Ding, R. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 762–770.
15. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
16. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
17. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
18. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.-C. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656.
19. Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; Gong, Z. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
20. Lin, L.; Song, S.; Yang, W.; Liu, J. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2490–2498.
21. Thoker, F.M.; Doughty, H.; Snoek, C.G. Skeleton-contrastive 3D action representation learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual China, 20–24 October 2021; pp. 1655–1663.
22. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

4.2 Cross-Sensor Adaptation

4.2.1 Summary

Skeleton representations (e.g., skeleton resolution and graph configuration) are conditioned by the skeleton acquisition sensors used for data collection. However, the skeleton acquisition systems used may be different between the source and the target domains, which makes the distinct skeleton representations lead to graph domain shift that impacts the performance of the skeleton-based human action recognition model. In this context, **EGDA** approaches intrinsic skeleton graph discrepancy related to sensory particularities and contributes an unsupervised cross-sensor adaptation paradigm.

In implementation, an unsupervised learning method treats the skeletal action knowledge and domain graph configurations separately, made possible by the development of a generation space and a discrimination space. The interaction between the generation space and the discrimination space transfers learnt knowledge from a data-rich source domain and a data-scarce target domain, resulting in a joint model for both the source and the target domains.

4.2.2 Original contributions

The contributions of this component of **EGDA** are twofold:

- To the best of our knowledge, it provides the first formal investigation on skeletal graph discrepancy in human action recognition;
- The method introduces a novel paradigm that allows elaborating a single human action recognition model that functions effectively on different skeleton graphs, which improves the scalability of existing skeletal action recognition models when deployed in varying real-world environments.

My role in the development of this framework involved multiple responsibilities,

including conducting an in-depth investigation of the problem, designing and refining the methodology, and validating the framework through rigorous experimentation, performing comprehensive analysis of the results, while a fellow Ph.D. student, co-author on the paper, contributed to the implementation of the framework, helping to develop a robust and efficient PyTorch deployment.

4.2.3 Article

The research is reported in the paper “Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition” (H. Tian, J. Dickens, P. Payeur), published at the *21st Conference on Robots and Vision (CRV 2024)* [107]. The online version can be found at <https://crv.pubpub.org/pub/pb2qi4uo/release/1>. The paper, published under the Creative Commons Attribution (CC BY) license where authors retain the copyright, is included below as published. Permission to reprint the paper in this thesis was obtained from all co-authors.

Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition

Haitao Tian

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
htian026@uottawa.ca

James Dickens

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
jdick088@uottawa.ca

Pierre Payeur

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
ppayeur@uottawa.ca

Abstract— Recent research on human action recognition is largely facilitated by skeletal data, a compact graph representation composed of key joints of the human skeleton that is efficiently extracted by body tracking systems and that offers the merit of being robust to environmental variations. However, the skeleton resolution and joint connectivity of the extracted skeletons may vary with sensor devices, which results in different skeleton graph representations on collected data. This paper investigates a cross skeleton graph domain adaptation approach where a skeleton action recognition model is trained upon a source skeletal data domain but is expected to adapt onto a target domain configured with a different skeleton graph. It proposes an adversarial learning framework where a generation space is developed on which the model learns valid skeletal action knowledge from the source graph domain. Interaction with an embedded discrimination space is employed to extract heterogeneous graph features from the target domain. Optimization of the generation space and the discrimination space is realized alternatively under adversarial learning which guarantees action-aware and domain-agnostic skeletal knowledge, thus forming a joint human action recognition model effectively functioning on both graph domains. In experiments, the paper evaluates the proposed method by incorporating graph convolutional networks into two skeleton action recognition benchmarks, NTU-RGB+D and Northwestern-UCLA, where comparisons are conducted to demonstrate the effectiveness of the proposed approach. Code will be available at <https://github.com/tht106/CrossGraphDA>.

Keywords: domain adaptation; skeleton data; human action recognition; unsupervised learning

I. INTRODUCTION

Human skeletal data has been widely utilized to address a variety of human activity analysis tasks, such as action recognition [1], interactive gaming [2], and video surveillance [3], where human movements are interpreted by the trajectories of human skeleton joints captured by motion capture (MoCap) systems in a pre-configured environment. However, skeletal data representations may vary significantly in graph configurations, involving different skeleton resolution (i.e., number of joints) and joint connectivity (i.e., relative position of joints being tracked), depending on the MoCap systems used for data collection. For instance, the different generations of Microsoft Kinect rely on different

skeleton body tracking algorithms and sensing modalities [4], resulting in a high degree of skeletal representation discrepancy in joint permutations, numbers, and positions [5], as depicted in Fig. 1.

The discrepancy in skeleton graph leads to domain shift which remains a challenge for research on skeleton-based human action recognition. First, in many practical scenarios, the development of skeleton-based human action recognition models leverages copious training data available from a label-rich source environment equipped with a well-calibrated MoCap system, to train an optimal model transferrable to a target application domain in the real world. However, when the target domain involves utilizing a different MoCap system whose body tracking skeleton graph configuration differs from that of the source domain, the model transfer tends to achieve suboptimal performance and the model's scalability is precluded in practical scenarios. Second, a common practice to deal with model transfer is to retrain the model with sufficient data samples collected with the MoCap system in the target application domain. However, data collection and annotation are extremely labor-intensive and time-consuming. Moreover, the resulting retrained model becomes dependent to a single domain, leading to poor generalization performance.

This paper proposes an unsupervised domain adaptation framework, which allows a joint skeletal human action recognition model to learn from heterogeneous graph configurations. The unsupervised learning method treats the skeletal action knowledge and domain graph features separately, made possible by the development of a generation space and a discrimination space. In the generation space, advantage is taken of sufficient well-labeled data from the source domain to supervise a graph convolutional network (GCN) to learn *action aware semantics*. In the discrimination space, a small number of unlabeled skeletal data samples facilitate an adversarial learning paradigm in which the GCN integrates *domain-agnostic graph knowledge*. In general, the interaction between the generation space and the discrimination space transfers learnt knowledge from a data-rich source domain and a data-scarce target domain, resulting in a joint model for both the source and the target domains.

The contribution of the unsupervised domain adaptation approach is twofold. First, it provides the first investigation on skeletal graph discrepancy in human action recognition.

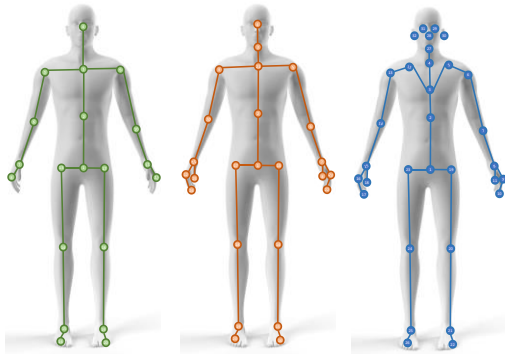


Figure 1. Skeletal joints detected by different generations of Microsoft Kinect. From left to right: Microsoft Kinect V1, Microsoft Kinect V2, and Microsoft Azure Kinect.

Second, it introduces a novel paradigm that allows to elaborate a single human action recognition model that functions effectively on different skeleton graphs, which will improve the scalability of existing skeletal action recognition models in varying real-world environments.

II. RELATED WORK

A. Human Action Recognition

Human action recognition [1] addresses the perception by computer vision of a variety of human actions. Recent research on human action recognition is largely facilitated by the encoding of human skeletons, a compact spatial-temporal sequence of 3D features extracted by processing sensors output while imaging human subjects and offering the merit of being robust to variations in the environment.

With the goal of further boosting state-of-the-art performance, GCNs have dominated recent research [6] [7] [8] on skeletal data-based human action recognition by involving skeletal data into end-to-end supervised learning. AS-GCN [6] introduces directed graph neural networks to model the temporal dependencies among skeletal joints. ST-GCN [7] employs spatial-temporal graph convolutional networks to capture both spatial and temporal information in skeletal data, achieving competitive results in action recognition. CTR-GCN [8] proposes a relations modeling framework for a GCN with the goal of interpreting relationships between joints.

Domain shift is observed in computer vision as a result of data distribution discrepancy emerging from domain changes [9][10]. In the context of skeletal data-based human action recognition, such a data representation shift leads to heterogeneous skeleton graph representations between the source domain and the target domain, in which a GCN model is prone to overfit to the source skeleton graph representation and suffers from the adaptation to a target domain. Consequently, the scalability of GCNs is significantly limited for practical human action recognition tasks.

B. Domain Adaptation

Domain adaptation (DA) [11] [12] has been successfully leveraged to approach the generalization issue of machine learning solutions in computer vision, by alleviating discrepancies between distinct domain distributions to improve the performance of generalization in unsupervised manners. For instance, DANN [13] introduces an effective domain adaptation procedure which aligns feature representations across domains by using a domain classifier to enforce domain invariance. Vu et al. [14] utilize entropy minimization to encourage unambiguous cluster assignments to maintain the target domain's clusters separable in image semantic segmentation.

However, research on domain adaptation in skeletal data-based human action recognition is quite limited. Mitsuzumi et al. [15] propose a data-efficient domain adaptation approach to learning cross-subject skeletal action recognition by utilizing phase randomization-based data augmentation. Skeleton-CutMix [16] proposes a simple and effective skeleton augmentation framework for supervised domain adaptation which hallucinates new skeleton representations by using pairs of skeletons from the source and target domains. But the existing research generally overlooks the cross-skeleton graph domain adaptation issue, and there is not an effective method proposed to address it.

C. Adversarial Learning

Recently, adversarial learning has been combined with domain adaptation motivated by the fact that generative adversarial networks (GAN) [17] excel at extracting essential semantic features that resemble a target distribution. UNIT [18] focuses on unsupervised image-to-image translation for domain adaptation. It employs a shared latent space and adversarial training to achieve domain-invariant representations. Hoffman et al. [19] embed a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. By imposing the DA operation, the segmentation network not only learns discriminative representations but also invariant encodings from different domains. Fukushi et al. [20] propose a few-shot generative model based upon cross-domain regularization and entropy regularization, which is effective for transferring the diversity of the motions contained in the source to a target domain.

III. METHODOLOGY

This section details the methodology for the proposed unsupervised domain adaptation in skeleton-based human action recognition. It begins with mathematical preliminaries about supervised skeletal action recognition, upon which it details the framework for unsupervised cross-graph domain adaptation.

A. Preliminaries

1) **Skeletal action recognition.** Human action movements can be composed of a sequence of skeletons, each of which consisting of a number of key body joints, and interpreting the trajectories of human movement over time. In the context of graph convolution, skeletons can be naturally treated as a

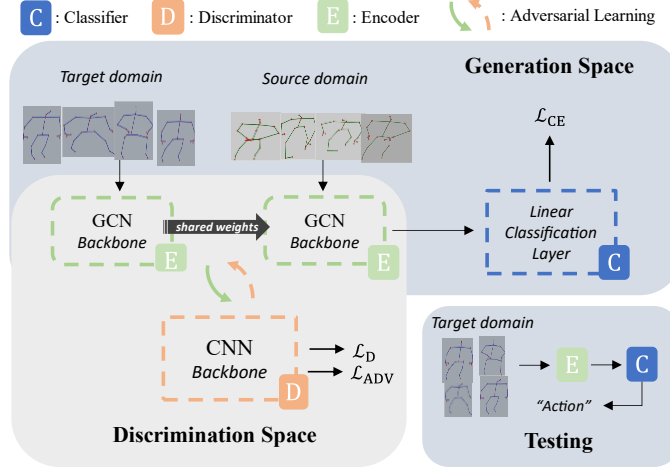


Figure 2. A conceptual overview of the proposed adaptation network for human action recognition.

graph-structured data form in which skeletal joints compose graph nodes, connected by skeletal bones, namely, node edges. A skeletal action recognition model considers a skeletal sequence input, $(\mathbf{X}, \mathbf{A}, \mathbf{y})$, where the $\mathbf{X} \in \mathbb{R}^{T \times V \times 3}$ is the sequence composed of T frames of skeletons, each of which is configured by the graph $\mathbf{A} \in \{0, 1\}^{V \times V}$ denoting the node connectivity of V skeletal joints; $\mathbf{y} \in \mathbb{R}^L$ is the corresponding action label of $\mathbf{X}$ from L action categories. In graph convolution, the GCN model aggregates action semantic representations by involving the skeletal sequence $\mathbf{X}$ into end-to-end supervised learning with the sequence-wise annotation $\mathbf{y}$. It employs an encoder $\mathbf{E}$ to perform graph convolution on the input which generates high-dimension feature encodings, $\mathbf{E}(\mathbf{X}, \mathbf{A}; \boldsymbol{\theta}_E)$, where $\boldsymbol{\theta}_E$ is the learnable network weights of $\mathbf{E}$. A classifier $\mathbf{C}$, parameterized by $\boldsymbol{\theta}_C$, then accepts the feature encodings and outputs action classifications on a soft-max layer. The optimization of the GCN, i.e., $\mathbf{C}(\mathbf{E}(\cdot))$ can be solved by typical neural network optimization schemes, such as stochastic gradient descent (SGD), with a cross-entropy loss:

$$L_{CE} = -\mathbf{y} \cdot \log[\mathbf{C}(\mathbf{E}(\mathbf{X}, \mathbf{A}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_C)] \quad (1)$$

2) **The cross-graph shift.** Despite the significant effectiveness of graph convolution on skeletal data, GCNs still demonstrate intrinsic limitations when learning from heterogenous skeleton graphs. Specifically, considering a labelled source domain, $\mathcal{D}_S = \{\mathbf{X}^s, \mathbf{A}^s, \mathbf{y}^s\}$, where $\mathbf{A}^s \in \mathbb{R}^{V_s \times V_s}$ and $\mathbf{y}^s \in \mathbb{R}^L$, and an unlabeled target domain, $\mathcal{D}_T = \{\mathbf{X}^t, \mathbf{A}^t\}$, where $\mathbf{A}^t \in \mathbb{R}^{V_t \times V_t} \neq \mathbf{A}^s$, the graph convolution of the encoder $\mathbf{E}$ on the source domain can be implemented by operating conventional 2D convolution on input $\mathbf{X}^s$, weighted by the adjacent matrix $\mathbf{A}^s$, which can be formulated as: $\mathbf{E}(\mathbf{X}^s, \mathbf{A}^s; \boldsymbol{\theta}_E) = \mathbf{E}_{\boldsymbol{\theta}_E}(\mathbf{X}^s) \cdot \mathbf{A}^s$, where $\mathbf{E}_{\boldsymbol{\theta}_E}$ is the optimal encoder parameterized by $\boldsymbol{\theta}_E$ obtained in (1), and $\cdot$ denotes the tensor product. Clearly, the knowledge distilled by $\mathbf{E}_{\boldsymbol{\theta}_E}$ on

the source domain resides on the graph $\mathbf{A}^s$. In this way, given a target domain sequence $\mathbf{X}^t$ associated with the graph $\mathbf{A}^t$, the cross graph discrepancy between $\mathbf{A}^s$ and $\mathbf{A}^t$ prevents the knowledge of $\mathbf{E}_{\boldsymbol{\theta}_E}$ from properly transferring to the target domain by formulating $\mathbf{E}_{\boldsymbol{\theta}_E}(\mathbf{X}^t) \cdot \mathbf{A}^t$.

B. Cross-graph Domain Adaptation

The goal of cross-graph domain adaptation is to learn a joint GCN model capable of functioning effectively on two skeleton domains, associated with the graphs, $\mathbf{A}^s$ and $\mathbf{A}^t$, respectively. However, given the absence of label statistics in the target domain, it is impossible to fine-tune an effective model with supervision in the target domain. To this end, an adversarial learning framework is proposed which incorporates a generation space to learn *action aware semantics* with a traditional GCN model while embedding a discrimination space to enforce the generation space to exploit *graph invariant representations* at the same time. The adversarial learning strategy is formulated in three successive stages, as detailed below. Fig. 2 depicts the overall concept.

1) **In the generation space**, the framework takes the GCN model, i.e., $\mathbf{C}(\mathbf{E}(\cdot))$, as the base structure to learn action aware semantics. Optimization of the encoder parameters, $\boldsymbol{\theta}_E$ is realized by supervised learning with the labeled source domain data by solving a cross-entropy loss:

$$L_{CE}(\mathbf{C}, \mathbf{E}) = -\frac{1}{|\hat{\mathcal{X}}_S|} \sum_{(\mathbf{X}^s, \mathbf{y}^s) \in \hat{\mathcal{X}}_S} [\mathbf{y}^s \log \mathbf{C}(\mathbf{E}(\mathbf{X}^s, \mathbf{A}^s; \boldsymbol{\theta}_E); \boldsymbol{\theta}_C)] \quad (2)$$

where $\hat{\mathcal{X}}_S$ denotes a batch of data samples from the source domain $\mathcal{D}_S$.

2) **In the discrimination space**, an auxiliary network, discriminator $\mathbf{D}$, is employed to classify the graph feature encodings from different domains. Specifically, the discrimination space involves batch data from the two domains, $\hat{\mathcal{X}}_S$ and $\hat{\mathcal{X}}_T$, as a dual input. The discriminator $\mathbf{D}$, which is embedded on the end of the encoder of the base

network, accepts the feature encodings, $\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_s, \mathbf{A}^s)$ and $\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_t, \mathbf{A}^t)$, that are aggregated on the generation space (where $\mathbf{E}_{\theta_E}$ denotes the encoder $\mathbf{E}$ parameterized by θ_E according to (2)), and performs binary classification upon their original domain labels. The development of $\mathbf{D}$ effectively circumvents the absence of label information in the target domain while taking the advantage of unlabeled skeleton samples for domain adaptation. The optimization on $\mathbf{D}$ is solved by a binary cross-entropy loss:

$$\mathcal{L}_D(\mathbf{D}) = \log[1 - \mathbf{D}(\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_s, \mathbf{A}^s); \theta_D)] + \log[\mathbf{D}(\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_t, \mathbf{A}^t); \theta_D)] \quad (3)$$

where θ_D is the learnable weights of the discriminator. Note that the network weights of the encoder θ_E are fixed at this stage and only contribute to the forward propagation. The optimization based on (3) aims to endow the discriminator $\mathbf{D}$ with the ability to discriminate the heterogeneous graph representations upon different domains. The structure of $\mathbf{D}$ will be detailed in Figure 3 and Section IV. B.

3) **Adversarial optimization.** Lastly, the encoder $\mathbf{E}$ is optimized by the loss function in (4) with supervision of the discriminator $\mathbf{D}_{\theta_D}$, parameterized by θ_D , to produce graph-agnostic skeletal representations.

$$\mathcal{L}_{\text{ADV}}(\mathbf{E}) = \log[\mathbf{D}_{\theta_D}(\mathbf{E}(\hat{\mathcal{X}}_t, \mathbf{A}^t); \theta_E)] \quad (4)$$

In (4), the network weights of the discriminator θ_D are fixed and only contribute to the forward propagation.

The network components, classifier $\mathbf{C}$, encoder $\mathbf{E}$, and discriminator $\mathbf{D}$, are updated alternatively by minmax adversarial optimization formulated in (5):

$$\mathbf{C}_{\theta_C}, \mathbf{D}_{\theta_D}, \mathbf{E}_{\theta_E} = \arg \min_{\mathbf{C}, \mathbf{E}} \max_{\mathbf{D}} \mathcal{L}(\mathbf{C}, \mathbf{D}, \mathbf{E}) \quad (5)$$

$$\mathcal{L}(\mathbf{C}, \mathbf{D}, \mathbf{E}) = L_{\text{CE}}(\mathbf{C}, \mathbf{E}) + \mathcal{L}_D(\mathbf{D}) + \lambda \mathcal{L}_{\text{ADV}}(\mathbf{E}) \quad (6)$$

where λ denotes the hyperparameter adapting the weights of adversarial optimization.

After adaptation, the learnt network $\mathbf{C}_{\theta_C}(\mathbf{E}_{\theta_E}(\cdot))$ is utilized to infer testing samples from the target domain by formulating: $\mathbf{C}_{\theta_C}(\mathbf{E}_{\theta_E}(\mathbf{X}^t, \mathbf{A}^t))$.

IV. EXPERIMENTS

This section experimentally evaluates the effectiveness of the proposed unsupervised adaptation framework by investigating cross-graph skeletal action recognition tasks that leverage two public datasets.

A. Datasets

NTU RGB+D [21] is a large-scale human skeleton action dataset composed of 56,880 samples covering 60 human daily actions. The skeletal data set is recorded in indoor scenes with three *Microsoft Kinect V2* cameras mounted in different setups. Each camera provides a skeleton sequence encoded over 25 joints. The samples from the ‘‘cross subject’’ training set [21] are used for adaptation training.

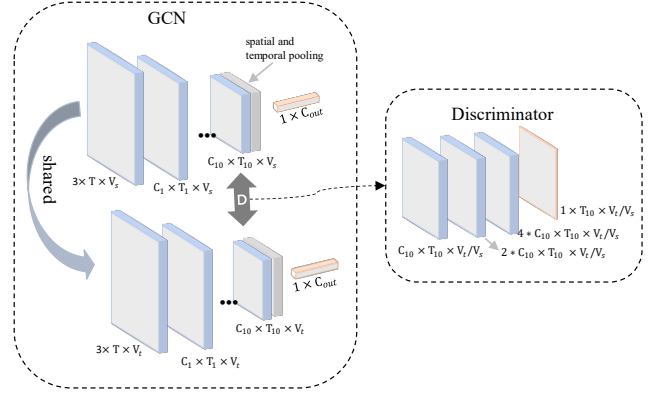


Figure 3. Illustration of network architectures for the proposed adaptation framework. First, the input layer of the GCN backbone accepts domain-specific skeletal data with the size of $3 \times T \times V_s$ and $3 \times T \times V_t$, followed by ten graph convolution blocks configured with the channel number $\{C_1, \dots, C_{10}\}$. $\mathbf{D}$ accepts high-dimensional feature encodings generated in the tenth block with the size of $C_{10} \times T_{10} \times V_s$ and $C_{10} \times T_{10} \times V_t$ and forward propagates through three normal convolutional layers with stride 2 and kernel size 4. Each of the first two layers is followed by a leaky ReLU. The channel number of $\mathbf{D}$ is $\{2C_{10}, 4C_{10}, 1\}$.

Northwestern-UCLA [22] is a relatively small skeletal dataset composed of 1494 short sequences of skeletons covering 10 human actions. This dataset is recorded by one *Microsoft Kinect V1* which provides skeleton representations encoded over 20 joints. The discrepancy on the number of joints and variations in camera setups with respect to **NTU RGB+D** acquisition configuration provides an effective evaluation scenario for domain adaptation. The skeleton coordinates of both datasets are normalized into $[-1, 1]$ during training and testing.

B. Implementation

The proposed network is deployed using PyTorch on a NVIDIA 3090 GPU. It uses CTR GCN [8] as the GCN backbone in the generation space, where encoder $\mathbf{E}$ is composed of ten blocks of graph convolution (refer to [8] for more details). Discriminator $\mathbf{D}$ is embedded on the final block of encoder $\mathbf{E}$, which is composed of a four-layer CNN as detailed in Fig. 3. The GCN backbone, CTR GCN, is optimized by using SGD, with momentum as 0.9, weight decay $4e-10$, and learning rate 0.1. The discriminator $\mathbf{D}$ is updated by the Adam optimizer with the learning rate as $1e-3$ and betas $\{0.9, 0.99\}$. The hyperparameter λ in (6) is set as 0.08. The performance of the proposed method is evaluated under the metric of Top-1 Accuracy [7].

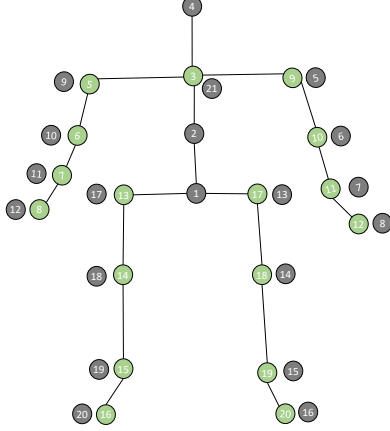


Figure 4. Joint remapping from Microsoft Kinect V2 to Microsoft Kinect V1. The joints connected by lines compose the remapped 20-joint skeleton under the same skeleton configuration (joint number and permutation) used by Microsoft Kinect V1. Each green joint on the skeleton is remapped from a nearby gray joint whose coordinates are indexed by the number indicated in the gray circle. For instance, the third joint on the remapped skeleton has the location information of the twenty-first joint of the source skeleton graph. Joints 1, 2 and 4 are the same under both mappings. Furthermore, five joints, 3-neck, 22-tip of the left hand, 23-left thumb, 24-tip of the right hand, and 25-right thumb, of the source skeleton graph are dropped after remapping.

C. Results

Two cross-graph skeletal action recognition tasks are considered to validate the performance of the proposed framework.

1) Test case on remapped skeleton representation

In the first cross-graph domain adaptation experiment, the original training dataset of **NTU RGB+D** (recorded with a *Microsoft Kinect V2*) is considered as the source domain. The skeletons from the **NTU RGB+D** are then remapped from 25 to 20 joints as per the matching rules detailed in Fig. 4, to mimic a target domain that is recorded by *Microsoft Kinect V1*. Experimental results on this cross-graph domain adaptation task are summarized in Table 1. First, a source-only model is trained with exclusive full supervision using the original training dataset composed of 25-joint skeletons with annotations to serve as a comparative. The source-only model reaches 83.3% on the test dataset supported by 25-joint encodings but achieves only 62.9% on the skeletons remapped to 20 joints. It demonstrates that a reduction on the skeleton resolution results in suboptimal performance for the GCN model. To overcome the issue, the model is adapted with the proposed adversarial learning scheme. In the second row of Table 1 it is observed that a significant performance improvement is achieved on the downsized 20-joint skeleton encodings (from 62.9% to 80.1%). At the same time, the adapted model exhibits stable performance when processing

Table 1. Comparative performance of the proposed method under the metric of Top-1 Accuracy.

NTU RGB+D 25 joints remapped to 20 joints		
Model	25 joints	20 joints
Source only	83.3%	62.9%
With adaptation	83.7%	80.1%

skeletons with 25 joints, as it was trained on. These results demonstrate that the proposed method successfully learns a joint model across skeleton graphs with different encodings and inherent resolution.

2) Test case on different skeleton encodings

The second experiment studies a more practical but challenging cross-graph action recognition scenario. Specifically, it considers **NTU RGB+D** with 25-joint skeleton encodings as the source domain and **Northwestern UCLA** with 20-joint encodings as the target domain. The skeleton sequences under the common 7 actions (i.e., "drop", "pickup", "throw", "sitting down", "standing up", "wear jacket", and "take off jacket") between the two domains are used for cross-graph domain adaptation training and testing. The overall experimental results over these 7 actions are detailed in Table 2. First, the "Source-only" model (trained on data (4672 samples) from **NTU RGB+D**) reaches 71.6% average Top-1 Accuracy on the 7 common actions when tested on 370 skeleton samples from **Northwestern UCLA** with fewer joints. After applying domain adaptation with 799 unlabeled training samples from the target domain, the proposed method gains an average performance boost of 14.7% on the target test dataset. Meanwhile, the proposed model with adaptation also outperforms the source-only model in a vast majority of action categories, with only very minor discrepancies in the weaker class of "sitting down". To explain the upper limit at 86.3% in Top-1 Accuracy, it is conjectured that the variations between **NTU RGB+D** and **Northwestern UCLA** in data collection settings, such as camera setups and environmental configurations, are significant enough and contribute complex data discrepancy in skeletal representations, presenting a significant challenge to any methods for skeleton-based human action recognition.

D. Ablation Study

1) Domain adaptation with fewer target domain samples

This work performs unsupervised learning for cross skeleton graph-based human action recognition by only using unlabeled target domain samples. Nevertheless, the development of the proposed domain adaptation framework still imposes data collection in the target domain. Further experimental investigation examines the principle that drives adaptation effectiveness with lower amounts of target domain samples to circumvent constraints on data usage. Specifically, it conducts an experiment where the GCN model is learnt with the same domain adaptation implementation detailed in the task "**NTU RGB+D to Northwestern UCLA**" but altering

Table 2. Experimental results of the adaptation task “NTU RGB+D to Northwestern UCLA”.

NTU-RGB+D to NW-UCLA	Drop	Pickup	Throw	Sitting Down	Standing Up	Wear Jacket	Take off Jacket	Top-1 Accuracy
Source only	24.4 %	74.2 %	33.3 %	97.8 %	95.1 %	83.7 %	92.7 %	71.6 %
With adaptation	66.7 %	82.8 %	73.0 %	97.6 %	99.8 %	92.3 %	94.4 %	86.3 %

Table 3. Performance variation related to different proportions of the target data samples used for domain adaptation. (100% denotes that 799 unlabeled samples from the target domain are used)

Percentage of data use	0% (source only)	5%	20%	30%	50%	70%	100%
Top-1 Accuracy	71.6 %	79.1%	82.5%	84.6%	85.1%	85.9%	86.3 %

Table 4. Experimental comparison of the discriminator implemented with different network structures. Memory denotes the allocated GPU memory resources during training upon the entire framework. Time/epoch denotes the time cost per epoch for training the framework on the source domain dataset.

Discriminator	Top-1 Accuracy	Memory	Time/epoch
CNN	86.3 %	4.1GB	55.7s
MLP	77.5%	3.5GB	49.4s

the ratios (varying from 0% to 100%) of samples used for training that were randomly selected from the target domain. Note that the number of test samples remains the same (i.e., 370) in the experiment. Overall experimental results are presented in Table 3. First, when no (0%) target domain samples are used, the model degrades into a source only model. Second, the model performance tends to increase monotonically along with the ratio of the target domain samples involved in adaptation. It is worth noting that even while using a very small number (e.g., 5%) of data samples from the target domain, the model will achieve a significant performance gain of 7.5% from the source only model (71.6% in Table 3), suggesting the capability of the proposed method for learning critical skeletal action semantics from heterogeneous domains.

2) Efficient implementation of the discriminator

The experimental results in Tables 1 and 2 suggest that the proposed method guarantees a performance boost to the target domain. Nevertheless, the introduction of a CNN-structured discriminator **D** inevitably imposes computational overhead to the conventional GCN based training pipeline. To circumvent the costs of the computational overhead, this research also examines an efficient implementation of **D** for domain adaptation. Specifically, it adopts a multi-layer perceptron (MLP) to fulfill the discrimination space. The MLP is composed of a spatial pooling layer and a temporal

Table 5. Performance variation related to hyperparameter λ .

λ	0	0.01	0.05	0.08	0.1	0.5
Top-1 Accuracy	71.6%	77.9%	85.8%	86.3%	86.0%	77.9%

pooling layer, followed by two linear layers equipped with a ReLU activation after the first linear layer. MLP aims to make binary classification upon the feature encodings generated by the encoder **E** while imposing a lower computational cost than a CNN network.

The experimental results are summarized in Table 4. First, the MLP-based discriminator reaches inferior performance compared to the CNN implementation in the task “NTU RGB+D to Northwestern UCLA”. It is conjectured that the dimension reduction imposed by the first two pooling layers could lead to information loss in spatial and temporal dimensions, compromising the eventual effectiveness of MLP in the discrimination space. On the other hand, even though the CNN-based discriminator, as proposed in Section III, preserves better adaptation performance, the convolution operations in the discrimination space involve an exponential number of network weights, leading to a heavier computation cost compared to MLP (e.g., extra 0.6GB memory use and 6.3s per epoch of training). Consequently, the experimental results demonstrated in Table 4 would suggest a practical trade-off between the model performance and computational cost in the implementation of the discriminator.

E. Parameter Learning

The adversarial interaction between the generation space and the discrimination space is crucial to the proposed method and is controlled by the hyperparameter λ , as formulated in Eq. (6). Setting an appropriate value of λ not only involves sufficient adversarial supervision for domain adaptation learning, but also alleviates the over-adaptation problem in the target domain. To this end, experimental investigation analyzes the choice of λ by conducting an experiment where it learns a GCN model with the same domain adaptation

implementation detailed in the task “**NTU RGB+D to Northwestern UCLA**” but altering λ over the range [0.01 0.5]. Experimental results are reported in Table 5. When $\lambda = 0$, the adversarial learning degrades into a traditional fully supervised GCN training upon the source domain where the resulting model is equivalent to a source-only model, which achieved 71.6% on the task “**NTU RGB+D to Northwestern UCLA**” in Table 2. Otherwise, the adaptation performance is clearly affected as λ varies with an eventual degradation in performance when the hyperparameter places too much weight on adversarial learning. Overall, it is observed that the best performance is achieved when λ is around 0.08.

CONCLUSIONS

This research investigates the scalability of existing human action recognition models as they are significantly limited by discrepancies in the skeletal representation that surface when the acquired data from a target domain is encoded in a different skeleton graph compared to that of the source training domain. This paper proposes the first domain adaptation framework for cross skeleton graph-based human action recognition. The proposed method utilizes adversarial learning to enable a model learning action-aware representations from a labeled source dataset, while extracting graph invariant representations from an unlabeled target domain. In its implementation, it employs a generation space and a discrimination space where a GCN model aggregates skeleton joint correlation from heterogeneous skeletal graphs while a discriminator predicts overall domain labels with respect to the respective domain aggregations encoded by the generation space. Experiments are conducted on two cross-graph domain adaptation tasks to demonstrate the validity of the proposed adaptation method and evaluate its effectiveness.

In future work, the aim will be to extend the proposed method to multi-variation domain adaptation. For instance, it will consider the skeletal data shift in both extrinsic (environment-specific) and intrinsic (sensor-specific) variations derived from different sources and devices, to attempt to transpose the proposed method for formulating more adaptive skeletal data-based human action recognition models.

ACKNOWLEDGMENTS

This research was supported in part by MITACS Accelerate and NSERC Discovery grants. The authors also acknowledge the collaboration of Spectronix Inc.

REFERENCES

- [1] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.
- [2] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, “Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor,” in *2011 annual international conference of the IEEE engineering in medicine and biology society*, 2011, pp. 1831–1834.
- [3] W. Niu, J. Long, D. Han, and Y.-F. Wang, “Human activity detection and recognition for video surveillance,” in *2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763)*, 2004, vol. 1, pp. 719–722.
- [4] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep learning-based human pose estimation: A survey,” *ACM Computing Surveys*, 56, no. 1 (2023): 1–37.
- [5] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, “Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect,” in *2015 international conference on healthcare informatics*, 2015, pp. 380–389.
- [6] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Actional-structural graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3595–3603.
- [7] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.
- [8] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise topology refinement graph convolution for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13359–13368.
- [9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Machine learning*, vol. 79, pp. 151–175, 2010.
- [10] S. Ma, Z. Yuan, Q. Wu, Y. Huang, X. Hu, C. H. Leung, D. Wang, and Z. Huang, “Deep into The Domain Shift: Transfer learning through dependence regularization,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” *Advances in neural information processing systems*, 19, 2006.
- [12] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
- [14] T.H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “Advent: adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [15] Y. Mitsuzumi, G. Irie, A. Kimura, and A. Nakazawa, “Phase Randomization: A data augmentation for domain adaptation in human action recognition,” *Pattern Recognition* 146 (2024): 110051.
- [16] H. Liu, Y. Liu, T.-J. Mu, X. Huang, and S.-M. Hu, “Skeleton-CutMix: Mixing Up Skeleton with Probabilistic Bone Exchange for Supervised Domain Adaptation,” *IEEE Transactions on Image Processing*, 2023.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [18] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” *Advances in neural information processing systems*, vol. 30, 2017.
- [19] J. Hoffman, D. Wang, F. Yu, and T. Darrell, “Fcns in the wild: Pixel-level adversarial and constraint-based adaptation,” *arXiv preprint arXiv:1612.02649*, 2016.
- [20] K. Fukushi, Y. Nozaki, K. Nishihara, and K. Nakahara, “Few-Shot Generative Model for Skeleton-Based Human Action Synthesis Using Cross-Domain Adversarial Learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3946–3955, 2024.
- [21] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, “Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.
- [22] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.

4.3 Cross-Sequence Adaptation

4.3.1 Summary

The scalability of existing skeletal action recognition models is significantly limited by the network architecture of GCNs that are tailored for graph convolutions on short, single-action (trimmed) sequences with fixed receptive field on the time axis. Consequently, the trained model is not deployable to a real-time system where the coming data consists of untrimmed sequences exhibiting concatenated actions. Cross-sequence adaptation refers to using **EGDA** to exploit skeleton action segmentation models that can be trained on short, trimmed skeleton videos, but that can run on longer untrimmed sequences.

In implementation, the approach is formulated in three steps: Stitch, Contrast, and Segment. First, “Stitch” proposes a temporal skeleton stitching scheme that treats trimmed skeleton videos as elementary human motions that compose a semantic space and can be sampled to generate multi-action stitched sequences. “Contrast” learns contrastive representations from stitched sequences with a novel discrimination pretext task that enables a skeleton encoder to learn meaningful action-temporal contexts to improve timewise action segmentation. Finally, “Segment” relates the proposed method to action segmentation by learning a segmentation layer while handling particular data availability.

4.3.2 Original contributions

The contributions of this component of **EGDA** consist of three aspects:

- It proposes an original formulation of temporal adaptation that allows for learning human action segmentation models from short, coarsely labeled sequences for a target application domain where only a small number of unlabeled long sequences are available for model refinement;
- It utilizes contrastive learning to learn fine-grained temporal segmentation embeddings from a source domain that can be represented by a skeleton encoder to transfer knowledge into a target domain for the task of action segmentation;
- The novel temporal cross-sequence adaptation framework transfers skeleton action knowledge into real-world action segmentation which leads to significant reduction of the data collection and annotation requirements in the target domain.

4.3.3 Article

This methodology and related research results are presented in “Stitch, Contrast, and Segment: Learning a Human Action Segmentation Model Using Trimmed Skeleton Videos” (H. Tian, P. Payeur), which was published at the *39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025)* [108]. The online version can be found at <https://ojs.aaai.org/index.php/AAAI/article/view/32792>. The published version of the paper is included below. Permission to reprint the paper in this thesis was obtained from the editor.

Stitch, Contrast, and Segment: Learning a Human Action Segmentation Model Using Trimmed Skeleton Videos

Haitao Tian*, Pierre Payeur

School of Electrical Engineering and Computer Science, University of Ottawa

htian026@uottawa.ca, ppayeur@uottawa.ca

Abstract

Existing skeleton-based human action classification models rely on well-trimmed action-specific skeleton videos for both training and testing, precluding their scalability to real-world applications where untrimmed videos exhibiting concatenated actions are predominant. To overcome this limitation, recently introduced skeleton action segmentation models involve untrimmed skeleton videos into end-to-end training. The model is optimized to provide frame-wise predictions for any length of testing videos, simultaneously realizing action localization and classification. Yet, achieving such an improvement imposes frame-wise annotated skeleton videos, which remains time-consuming in practice. This paper features a novel framework for skeleton-based action segmentation trained on short trimmed skeleton videos, but that can run on longer untrimmed videos. The approach is implemented in three steps: *Stitch*, *Contrast*, and *Segment*. First, *Stitch* proposes a temporal skeleton stitching scheme that treats trimmed skeleton videos as elementary human motions that compose a semantic space and can be sampled to generate multi-action stitched sequences. *Contrast* learns contrastive representations from stitched sequences with a novel discrimination pretext task that enables a skeleton encoder to learn meaningful action-temporal contexts to improve action segmentation. Finally, *Segment* relates the proposed method to action segmentation by learning a segmentation layer while handling particular data availability. Experiments involve a trimmed source dataset and an untrimmed target dataset in an adaptation formulation for real-world skeleton-based human action segmentation to evaluate the effectiveness of the proposed method.

Introduction

Skeleton videos represent the trajectories of human body key joints captured by skeleton acquisition systems that leverage 2D and 3D computer vision with specialized software to estimate the spatial configuration of a human body skeleton model in a pre-calibrated environment. Such sensing systems offer the merit of being robust to variations in the environment compared to other modalities such as RGB frames (Carreira and Zisserman 2017) and optical flows



Figure 1: (a): Two video sequences from PKU MMD v2 exhibiting multiple actions that can be segmented into several class-specific actions. Four actions (highlighted with yellow outlines) found in both sequences are permuted in different orders of occurrence (1,2,3,4 vs. 4,1,2,3). (b): A well-trained* skeleton-based action segmentation model may generate very different segmentation results on the two sequences. We identify this issue as the shortcuts learning problem which results from learning action permutations during training. [*The MS-GCN (Filtjens et al. 2022) model was trained over the small PKU MMD v2 skeleton video-based dataset using the official Cross-View protocol after removing the training sequences with action permutation order: “4,1,2,3”. See more details in “Ablation Studies”.]

(Feichtenhofer et al. 2016). Early research (Yan et al. 2018; Li et al. 2019; Chen et al. 2021) uses *trimmed* skeleton videos (i.e., class-specific action sequences as shown in Figure 1) to train human action recognition models, e.g., the end-to-end learning method (Yan et al. 2018) uses Graph Convolutional Networks (GCN) (Zhang et al. 2019) to parse human action dynamics from skeleton graphs, demonstrating high success rates in human-centered video analysis tasks (Niu et al. 2004). However, the scalability of existing human action recognition methods remains limited in the testing

phase when an *untrimmed* skeleton video covering multiple actions in a sequence is captured by a skeleton acquisition system while the trained recognition model is tuned for single action prediction. To overcome the issue, recent research (Ma et al. 2021; Filtjens et al. 2022; Li et al. 2023; Yang et al. 2023; Chai et al. 2024) develop a skeleton-based action segmentation model, providing frame annotations for *untrimmed* videos as an input. The difficulty of this paradigm lies in the exploitation of temporal structures on untrimmed videos (as illustrated in Figure 1) that imposes the collection and annotation of a large untrimmed dataset, which is time-consuming and labor-intensive.

This work attempts to answer the following research question: **Is it possible to use trimmed skeleton videos to learn an effective action segmentation model that works on untrimmed videos?** In general, trimmed skeleton videos are easier to acquire and annotate from a controlled domain, e.g., public datasets or laboratory environments, offering a copious data volume for learning a real-world downstream task, e.g., action segmentation. An ideal scenario involves training an off-the-shelf segmentation model only using trimmed skeleton videos and making it applicable to a target application domain that involves untrimmed and unlabeled skeleton videos. Alternatively, one would pre-train an initial model upon trimmed videos and customize an action segmentation model for the target domain with a fine-tuning phase using a small number of untrimmed videos. Either strategy transfers skeleton action knowledge from a domain of low temporal granularity to that of high temporal granularity, effectively avoiding learning from scratch and reducing the annotation effort of videos in the target domain.

In this context, this work proposes: *Stitch*, *Contrast*, and *Segment*, as a temporal domain adaptation approach for skeleton-based action segmentation. (1) *Stitch*. The framework treats a dataset of trimmed skeleton sequences as a semantic space of elementary motions and uses frame correspondence to align short skeleton sequences on spatial trajectories. Skeleton sequences being robust to variations in the environment backgrounds, frame correspondence can be registered upon the spatial configurations of human skeletons (Figure 2 provides an illustration) to generate multi-action stitched skeleton sequences. (2) *Contrast*. The framework then considers the resulting stitched videos as an informative volume of rich temporal structures and exploits meaningful skeleton representations to improve action segmentation via pre-training. It differs from previous skeleton contrastive learning works that use trimmed sequences to design pretext tasks (Thoker et al. 2021; Guo et al. 2022; Huang et al. 2023; Zhang et al. 2023; Lin et al. 2023) or that use sliding windows to adapt sequence-wise classification to action segmentation (Chen et al. 2022; Lin et al. 2023; Yang et al. 2023). (3) *Segment*. Temporal action segmentation is challenging as it treats each frame in a sequence as an instance, which spirals the dimensionality of a network

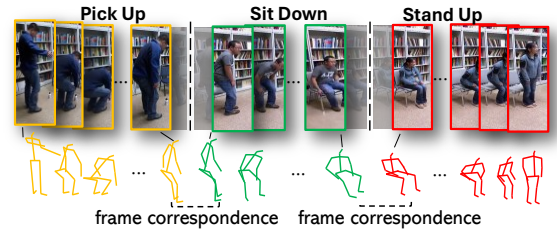


Figure 2: A concatenation of skeleton sequences resulting from stitching along the temporal axis. Upper row: RGB human motions involve strong dependency on environment backgrounds and significant variations in subject appearances. Lower row: skeleton motions represented by human key joints are agnostic to backgrounds, thus being easier to be manipulated with simple modifications (e.g., scaling and rotation). The proposed skeleton sequence stitching method involves using frame correspondence to align and concatenate short class-specific sequences.

structure. To this end, we opt to use a modular action segmentation structure (Filtjens et al. 2022), which allows to adopt different learning strategies (e.g., supervised adaptation and zero-shot adaptation) towards real-world human action segmentation where the labeled data is scarce.

The contributions of this work are threefold: *First*, it proposes a simple but effective temporal skeleton stitching method which can generate plausible untrimmed skeleton sequences where different actions are smoothly concatenated. *Second*, the proposed contrastive learning scheme demonstrates the effectiveness in exploiting temporal skeleton representations to improve action segmentation. *Lastly*, the novel temporal domain adaptation framework transfers skeleton action knowledge into real-world action segmentation which leads to significant reduction of the data collection and annotation requirements in the target domain. The project is available at: <https://htian026.github.io/GCL>, which contains supplementary material.

Related Work

Skeleton Action Recognition. Previous works (Yan et al. 2018; Li et al. 2019; Chen et al. 2021) regard each skeleton sequence as a spatial-temporal graph and parse human action dynamics by spatial and temporal graph convolutions from massive skeleton data samples, thus realizing sequence-wise action classification for a given testing video. However, action recognition considers a simplified situation where both training and testing samples are trimmed to cover the entire movements involved in only one specific human action from the beginning to the end.

Skeleton Action Segmentation. Considering that skeleton videos from the real world are rarely segmented and typ-

ically display human actions in continuity, action segmentation is the natural evolution of action recognition. For instance, MS-GCN (Filtjens et al. 2022) is jointly learned over a GCN and a Temporal Convolutional Network (TCN) (Carreira and Zisserman 2017)), where the former acts as a skeleton encoder to aggregate feature representations from untrimmed skeleton videos, while the latter models long-term action correlations from multi-scale temporal receptive fields. However, leveraging action segmentation escalates the difficulty of data acquisition as one needs to collect untrimmed long videos with frame-based annotation, which is labor intensive in practice.

Transfer Learning in Action Segmentation. Transfer learning algorithms help circumvent the shortage in data and have been used in RGB action video analysis. For example, I3D (Carreira and Zisserman 2017) is widely used in transfer learning for action segmentation by pre-training on short RGB videos (Piergiovanni and Ryoo 2018). In contrast, transfer learning in skeleton videos has not been extensively investigated. UNIK (Yang et al. 2021) involves pre-training a GCN on a large dataset and fine-tuning for action recognition in target domains with changing environments. To the best of our knowledge, the approach developed in the current paper presents the first method in transfer learning dedicated to skeleton action segmentation.

Skeleton Data-based Contrastive Learning. Contrastive learning has been utilized in skeleton representation learning (Thoker et al. 2021; Guo et al. 2022; Zhang et al. 2023; Lin et al. 2023). However, existing work focuses on action recognition and is difficult to apply to action segmentation. For instance, studies (Lin et al. 2023; Yang et al. 2023) utilize sliding windows to pre-segment an untrimmed sequence into equal-length sequence clips and apply representations extracted from trimmed skeleton videos to sequence clips. However, such a scheme essentially operates action recognition on untrimmed data, which inevitably introduces noisy boundaries on the segments and thus leads to compromised performance. In contrast, our work opts out of the use of sliding windows and formulates a contrastive learning framework specific for action segmentation.

Problem Statement

Skeleton-based Action Segmentation

A skeleton sequence, $L \in \mathbb{R}^{T \times V \times C_{in}}$, encodes the abstract dynamics of human movement over a period of time, T , depicting the trajectories of V skeleton joints in C_{in} (2 or 3) dimensions. When acquired with a skeleton acquisition system, a continuous (*untrimmed*) skeleton video tends to cover multiple actions concatenated in a single sequence.

Definition 1 (Action Permutation) Given an untrimmed skeleton video, $L \in \mathbb{R}^{T \times V \times C_{in}}$, we define the action permutation, $P = \{y_1, \dots, y_n, \dots, y_N\}$, as a sequence indicating the

order of occurrence of the actions within L , where y_n is the action label of the n -th action’s video segment $X_n \in \mathbb{R}^{T_n \times V \times C_{in}}$, with duration T_n which leads to $T = \sum_{n=1}^N T_n$. Herein, we have $L = X_1 \circ \dots \circ X_n \circ \dots \circ X_N$, where the symbol $\circ$ denotes the temporal concatenation operation.

Conventional skeleton *action recognition* models favor single-action prediction (Kong and Fu 2022), i.e., $|P| = 1$. Aiming to process a multi-action input sequence, where $|P| > 1$, an *action segmentation* model involves frame-wise action categorization, which identifies the change of state between successive actions by assigning different action labels to specific frames in a skeleton video sequence.

Temporal Domain Adaptation

Given that the annotation of untrimmed skeleton videos is extremely time-consuming and labor-intensive, this work focuses on a challenging alternative: learning skeleton-based action segmentation from trimmed skeleton videos. It assumes a large-scale source data domain, e.g., a laboratory dataset, containing well-trimmed and sequence-wise labeled short sequences X , and a small target data domain composed of untrimmed skeleton sequences L (labeled or not). Our goal is to utilize the source data X to help learn a robust and adaptive action segmentation model for the target domain associated with a few (if any) frame-wise labeled sequences within L . We call this process **temporal domain adaptation** as it transfers skeleton knowledge across different temporal granularities, e.g., from a coarse sequence-wise action recognition to finer frame-wise action segmentation. In this paper, two temporal domain adaptation objectives are pursued: supervised adaptation and zero-shot adaptation.

Methodology: Stitch, Contrast, and Segment

The proposed framework is structured in three steps. Initially, a skeleton temporal alignment and stitching strategy, *Skeleton-Stitch*, is proposed to generate stitched skeleton sequences. Second, it introduces Granular Contrastive Learning (*GCL*) to learn a skeleton encoder from stitched sequences. In the last step, it learns a segmentation layer upon the learnt skeleton encoder.

Skeleton-Stitch on Trimmed Sequences

Observing that human actions are naturally and smoothly linked one to another when performed in a sequence, two skeleton frames on a temporal boundary within an untrimmed skeleton video, representing the end of the first action and the start of the second, shall be continuous in the location of their spatial joints. In this paper, we assume that if two skeleton frames, sampled from two distinct sequences, are aligned in spatial joint configurations, they could form some correspondence for stitching successive sequence segments into a continuous sequence.

Specifically, we present a simple but efficient strategy to realize temporal stitching of trimmed sequences. The goal is to generate a stitched sequence, L , conditioned by the action permutation, $P_L = \{y_1, \dots, y_n, \dots, y_N\}$, based on the given data batch $\mathbb{B}$, using a *Skeleton-Stitch* operation $\mathcal{T}$:

$$L = \mathcal{T}(\mathbb{B}, P) \quad (1)$$

where $\mathbb{B}$ is a disordered collection of trimmed skeleton videos. In Eq. (1), $\mathcal{T}$ accepts $\mathbb{B}$ as a semantic space for sampling and P as a compositing surface for stitching. *Skeleton-Stitch* can be further decomposed as follows:

Data Preprocessing. We first apply a data pre-processing step to eliminate variation in views and scales. (1) We follow (Shi et al. 2019) to use a body coordinate system where the skeleton sequences are subtracted by the trajectory of the “spine” joint. (2) We follow (Liu et al. 2023) to apply a scaling operation that normalizes bone lengths on skeletons coming from different sequences.

Definition 2 (Frame Correspondence) Given two frames $p \in \mathbb{R}^{V \times C_{in}}$ and $q \in \mathbb{R}^{V \times C_{in}}$, sampled from two different skeleton sequences, p is a frame correspondence of q , if their spatial distance $d(p, q) \leq d^*$. The metric d is defined as $d(p, q) = \sum_{v=1}^V (\|p_v - q_v\|_2^2 + \max(1 - \frac{(p_v)^T q_v}{\|p_v\| \|q_v\|}, 0))$ which calculates the sum of the numerical differences on spatial locations and orientations at each skeleton joint. d^* is a threshold to indicate a valid correspondence.

Frame Correspondence Registration on $\mathbb{B}$. Given a template frame p sampled from a sequence of $\mathbb{B}$, we register its frame correspondence as $q = \mathbb{B}[i, t]$, where

$$d(p, \mathbb{B}[i, t]) \leq d^* \quad (2)$$

where t is the timestamp in the i -th sequence of $\mathbb{B}$. Eq. (2) encloses a loose constraint as p may correspond to more than one frame correspondence given the variable value of d^* . To solve the ambiguity, we determine the correspondence by introducing additional computation constraints.

Definition 3 (Class-specific Subset) Let $B \equiv \{1, \dots, |\mathbb{B}|\}$ be the indices of a batch of labeled data $\mathbb{B}$, we define the class-specific subset under the action category k as $B_k \equiv \{i \in B: y_i = k\}$, where y_i is the action label of i -th element in $\mathbb{B}$.

Multi-Action Stitches on P . *Skeleton-Stitch* $\mathcal{T}$ uses $P = \{y_1, \dots, y_n, \dots, y_N\}$ as a compositing surface to stitch multi-action sequences that are located at class-specific subsets: $B_{y_1}, \dots, B_{y_n}, \dots, B_{y_N}$. Specifically, $\mathcal{T}$ implements a N -step recursive search. At step $n \geq 2$, it stitches the sequence

$$\tilde{L}_{(n)} = \tilde{L}_{(n-1)} \circ X_n \quad (3)$$

where $\tilde{L}_{(n-1)}$ is obtained at step $n-1$ with $P_{\tilde{L}_{(n-1)}} = \{y_1, \dots, y_{n-1}\}$; $\tilde{L}_{(1)}$ is a single-action sequence randomly sampled from B_{y_1} . X_n is a single-action segment with $P_{X_n} = y_n$, and is obtained by

$$X_n = \mathbb{B}[i, t:] \quad (4)$$

$$\text{s.t. } i, t = \underset{\substack{i^* \in B_{y_n} \\ t^* \in \{1, \dots, T^*\}}}{\text{argmin}} d(p, \mathbb{B}[i^*, t^*]) \quad (5)$$

Eq. (4) denotes that X_n is a cropped sequence of $\mathbb{B}[i]$ by using the timestamps starting from t . Eq. (5) denotes that i and t are obtained by registering frame correspondence *w.r.t.* the template p (defined as the last frame of $\tilde{L}_{(n-1)}$), where i is searched from the class-specific subset B_{y_n} and t from the first T^* timestamps of X_i (where $T^* = |\mathbb{B}[i^*]|/\beta$, where β acts as a threshold for controlling the search range). We solve Eq. (4) as a *0-1 Integer Programming* problem (details are available in supplementary material). The recursive operation stops at step N and obtains a stitched sequence $L \equiv \tilde{L}_{(N)} = X_1 \cdots X_n \cdots X_N$.

Granular Contrastive Learning on Stitched Sequences

In this section, we consider the stitched sequences generated by *Skeleton-Stitch* as an informative volume which can facilitate representation learning for action segmentation. Based on this, the framework proposes Granular Contrastive Learning (*GCL*), a novel contrastive learning framework compared to previous methods that use sliding windows.

Skeleton Encoder. We detail the proposed *GCL* framework in Fig. 3. It consists of a feature aggregation subnetwork, $\mathcal{M}: \mathbb{R}^{T \times V \times C_{in}} \rightarrow \mathbb{R}^{T \times C_{out}}$, parameterized by ST-GCN (Yan et al. 2018) which parses skeleton action encodings at the feature space, and a MLP-based feature projection layer, $\mathcal{G}: \mathbb{R}^{C_{out}} \rightarrow \mathbb{R}^{C_{prj}}$, providing one-dimensional vector of contrastive features at the output space.

Data augmentation. Previous contrastive learning paradigms utilize simple spatial and temporal skeleton modifications (Thoker et al. 2021; Guo et al. 2022; Huang et al. 2023; Zhang et al. 2023; Lin et al. 2023) as data augmentation tailored for independent single-action sequences, thus running short on exploiting contextual information between heterogenous actions. Aiming at representation learning for action segmentation, this work implements action-correlated data augmentations by means of *Skeleton-Stitch* $\mathcal{T}$.

Definition 4 (Granular Category Encodings) Given a stitched sequence, $L = X_1 \cdots X_n \cdots X_N$, and its activation at $\mathcal{M}$, i.e., $f = \mathcal{M}(L) \in \mathbb{R}^{T \times C_{out}}$, we treat f as a concatenated sequence of categorical activations, i.e., $f = f_1 \cdots f_n \cdots f_N$, where $f_n \in \mathbb{R}^{T_n \times C_{out}}$ is the corresponding activation of X_n whose action label is y_n and duration is T_n . We have $T = \sum_{n=1}^N T_n$. We define the granular category encoding of f_n as $h_n = \delta(f_n)$, where δ is a temporal pooling operation, i.e., $\delta: \mathbb{R}^{T \times C_{out}} \rightarrow \mathbb{R}^{C_{out}}$.

Let $L_i = \mathcal{T}(\mathbb{B}, P_i)$ and $L_j = \mathcal{T}(\mathbb{B}, P_j)$ be two stitched sequences presenting different action permutations, i.e., $P_i \neq P_j$, and k be one of the common action categories found in the two sequences, i.e., $k \in K \equiv (P_i \cap P_j)$, we use *granular*

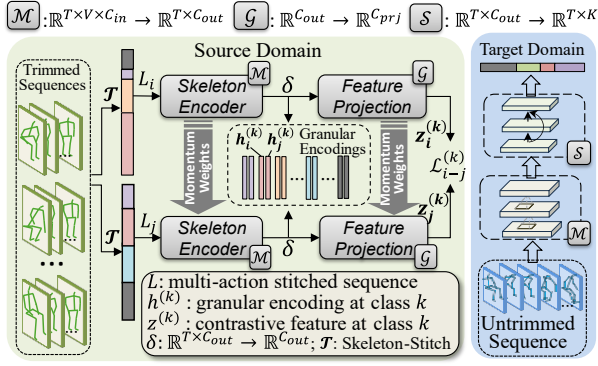


Figure 3: Conceptual overview of the proposed framework.

category encodings, $h_i^{(k)}$ and $h_j^{(k)}$, as a pair of *positives*, where $y_i = y_j = k$.

Stitched sequence-based pretext task. *GCL* aims to minimize the distance in between the positive pairs of all possible action category k at each batch. Specifically, let B_k be a class specific subset of granular category encodings of action category k collected from multiple stitched sequences, it uses the projection layer, G , to generate a pair of feature projections $z_i^{(k)} = G(h_i^{(k)})$ and $z_j^{(k)} = G(h_j^{(k)})$. The contrastive distance between $z_i^{(k)}$ and $z_j^{(k)}$ is calculated by the InfoNCE loss (Chen et al. 2020):

$$\mathcal{L}_{i-j}^{(k)} = -\log \frac{\exp(z_i^{(k)} \cdot z_j^{(k)} / \tau)}{\exp(z_i^{(k)} \cdot z_j^{(k)} / \tau) + \sum_{m \in \mathbb{M}} \exp(z_i^{(k)} \cdot m / \tau)} \quad (6)$$

where $i \in B_k$, $j \in J_k \equiv B_k \setminus \{i\}$, and $k \in K$. In Eq. (6), τ acts as the temperature parameter, and m denotes a *negative* counterpart sampled from a “memory bank”, $\mathbb{M}$, which is used to compensate for using a small batch size due to memory limitation. We follow the same implementation (Guo et al. 2022) on the memory bank. The optimization upon Eq. (6) encourages the subnetwork $\mathcal{M}$ to represent closely aligned granular embeddings at the action category k . The final objective function is the sum of all positive samples at all action categories:

$$\mathcal{L} = \frac{1}{|K|} \sum_{k \in K} \frac{1}{|B_k|} \sum_{i \in B_k} \frac{1}{|J_k|} \sum_{j \in J_k} \mathcal{L}_{i-j}^{(k)} \quad (7)$$

We evaluate the skeleton representations of a pretrained subnetwork $\mathcal{M}$ in multiple adaptation tasks for action segmentation, which will be detailed in the next section.

Skeleton Action Segmentation with Temporal Domain Adaptation

This section learns a segmentation (frame annotating) layer, $\mathcal{S}: \mathbb{R}^{T \times C_{out}} \rightarrow \mathbb{R}^{T \times K}$, which accepts hidden state at each frame from $\mathcal{M}$ as an instance and generates frame-wise softmax prediction for K action categories (Figure 3). For implementation, we build $\mathcal{S}$ upon a Temporal Convolutional

Network (TCN) which discerns action contexts (Lea et al. 2017) over adjacent frames by performing multi-scale dilated temporal convolutions (Farha and Gall 2019). We follow (Filtjens et al. 2022) for the implementation.

Our final action segmentation model is composed of the feature aggregation subnetwork, $\mathcal{M}$, and a segmentation layer, $\mathcal{S}$. Such a modularity allows us to relate the proposed *Skeleton-Stitch* and *Granular Contrastive Learning* to real-world action segmentation objectives while handling particular data availability. Particularly, it makes use of *source* trimmed sequences to pre-learn $\mathcal{M}$ via *GCL* and then realize multiple temporal domain adaptation strategies in the *target* domain. **In supervised adaptation**, it assumes that the *target* domain contains labeled untrimmed sequences, upon which we fine-tune $\mathcal{S}$ over the pre-trained $\mathcal{M}$. **In zero-shot adaptation**, as the *target* provides no training data, it uses *Skeleton-Stitch* to generate stitched data from the *source* domain for fine-tuning $\mathcal{S}$ over the pre-trained $\mathcal{M}$. More details on data generation are available in supplementary material.

Experiments

In experiments, *Stitch*, *Contrast*, and *Segment* (SCS) involves pairs of datasets, in each case one trimmed and the other untrimmed, for each temporal domain adaptation task considered, i.e., supervised adaptation or zero-shot adaptation. Moreover, experimental ablation studies examine best practices in this work, facilitating its reproduction.

Datasets

NTU RGB+D (Shahroudy et al. 2016) is a large-scale 3D human skeleton trimmed action dataset composed of 56,880 class-specific sequences. The dataset covers 60 human daily actions. The trimmed sequences are very short with low variation in action duration. **PKU-MMD** (Liu et al. 2017) is a popular 3D untrimmed skeleton dataset available in two versions. **PKU-MMD V1** covers 51 human daily actions. The dataset can be trimmed into 21,545 class-specific sequences. **PKU-MMD v2** contains 1,009 untrimmed video sequences covering 7,000 class-specific sequences. **Toyota Smarthome** (Das et al. 2019) is a large-scale 2D real-world video dataset for activities representing daily living. It covers up to 16,000 class-specific (trimmed) skeleton sequences with 31 action categories. **Toyota Smarthome Untrimmed** (Dai et al. 2022) is a real-world 2D dataset, containing 537 multi-action sequences covering 51 daily actions.

Implementation

In the experiments, we use multiple metrics to facilitate comparisons with related works. For instance, we use Top1 Accuracy (**Acc**) (Yan et al. 2018), mean Intersection over Union (**mIoU**) (Lin et al. 2023), mean Average Precision at IoU thresholds 0.1 (**mAP@0.1**) and 0.5 (**mAP@0.5**) (Yang

NTU to PKU1	Sliding Window	Evaluation Method	mAP% (0.1)	mAP% (0.5)
LAC (ICCV 23)	Y	Linear	61.8	-
SCS (s.a.)	N	Linear	80.9	76.4
CrosSCLR (CVPR 21)	Y	E2E	-	60.1
Hi-TRS (ECCV 22)	Y	E2E	-	66.6
LAC (ICCV 23)	Y	E2E	92.6	90.6
SCS (s.a.)	N	E2E	94.6	91.1

Table 1: Experimental results with supervised temporal adaptation from “NTU RGB+D to PKU MMD v1”.

NTU to PKU2	Sliding Window	Evaluation Method	mIoU%	Acc%
AimCLR (AAAI 22)	Y	Linear	15.7	39.8
ActCLR (CVPR 23)	Y	Linear	21.4	51.3
SCS (s.a.)	N	Linear	53.4	69.5
Hi-TRS (ECCV 22)	Y	E2E	-	55.0
SCS (s.a.)	N	E2E	60.9	81.1

Table 2: Experimental results with supervised temporal adaptation from “NTU RGB+D to PKU MMD v2”.

et al. 2023), and per-frame mAP (**mAP_f**) (Dai et al. 2022) in action segmentation evaluation. Unless otherwise specified, all models are tested on the testing set of the target domain under the Cross-Subject protocol where the dataset is split upon subject identifications. We conduct **Linear** evaluation (where we fix $\mathcal{M}$ while fine-tuning $\mathcal{S}$) and **End-to-end (E2E)** evaluation (where we fine-tune $\mathcal{M}$ and $\mathcal{S}$ jointly) to ensure fair comparison with previous contrastive models. More details are available in supplementary material.

Temporal Domain Adaptation Results

This section evaluates different learning strategies by amalgamating *Skeleton-Stitch* ($\mathcal{T}$) and *GCL* on multiple trimmed-untrimmed pairs of datasets. Our models are referred to as **SCS (s.a.)** in supervised adaptation and **SCS (z.a.)** in zero-shot adaptation.

NTU Trimmed to PKU Untrimmed. We consider a single source domain, NTU RGB+D, and two target domains, PKU-MMD v1 and PKU-MMD v2. **SCS (s.a.)** is compared to state-of-the-arts methods (Li et al. 2021; Thoker et al. 2021; Guo et al. 2022; Zhang et al. 2023; Lin et al. 2023; Yang et al. 2023) that utilize a sliding window to adapt contrastive skeleton representations from the trimmed source to the untrimmed target. Experimental results reported in Table 1 and Table 2 demonstrate that **SCS (s.a.)** significantly outperforms all state-of-the-art comparatives. We conclude that, by utilizing sliding window, existing work essentially operates action recognition (sequence-wise annotating) on the PKU-MMD domain from which the task granularity does not change across domains. In contrast, our model eliminates the use of a sliding window and learns a segmentation layer specific to the target domain, thereby leading to better performance. Second, **SCS (z.a.)** considers that the

Zero-Shot Adaptation	Evaluation Method	NTU to PKU1 mIoU%	NTU to PKU1 Acc%	NTU to PKU2 mIoU%	NTU to PKU2 Acc%
Baseline	E2E	10.5	18.9	5.4	11.6
SCS (z.a.)	Linear	45.1	52.9	21.2	31.5
SCS (z.a.)	E2E	53.6	61.3	20.8	31.9

Table 3: Experimental results with zero-shot adaptation.

target domain is totally unknown, as we directly train an action segmentation model on the source domain NTU RGB+D. To the best of our knowledge, our method is the first to propose a trimmed-to-untrimmed zero-shot learning approach. For comparison, we learn baseline models using the original trimmed sequences from the source domain (all frames of each trimmed sequence are assigned the original sequence-wise action label). Table 3 presents the testing results, demonstrating the substantial benefit of learning from stitched sequences using the proposed method.

PKU Trimmed to PKU Untrimmed. Next, we consider the trimmed PKU-MMD v1 as the source domain and the untrimmed PKU-MMD v2 as the target domain. The skeleton sequences under the common 41 single-person actions are used for temporal domain adaptation training and testing. For comparison, we learn baseline models using the source trimmed sequences in zero-shot adaptation and using the target untrimmed sequences in supervised adaptation. Experimental results are summarized in the “**PKU1 to PKU2**” columns of Table 4 and demonstrate that the proposed models, **SCS (s.a.)** and **SCS (z.a.)**, outperform the baseline comparatives by large margins.

TS Trimmed to TS Untrimmed. The experiment **TST to TSU** considers the trimmed Toyota Smarthome (TST) as the source domain and the Toyota Smarthome Untrimmed (TSU) as the target domain. For the untrimmed sequences in the target domain, we map the frame annotations of each sequence that are not in the label space of the source domain as a background. Experimental results with respect to zero-shot adaptation and supervised adaptation are summarized in the last two columns of Table 4, also demonstrating significant improvement over baseline comparatives.

Ablation Studies

Ablation experiments are conducted to better understand the effectiveness of each component introduced in this work.

Stitched Data vs. Original Data. *Skeleton-Stitch* uses frame correspondence to align heterogenous trimmed sequences and generate stitched sequences. In this section, we validate the effectiveness of the stitched sequences by reassembling datasets, e.g., PKU-MMD Untrimmed (v1 and v2) and Toyota Smarthome Untrimmed. Specifically, we first trim each sequence of the training set from an untrimmed dataset into class-specific sub-sequences using their associated data label and timestamps (we keep the original testing sets untouched for model evaluation). We shuffle the

Models	Eval. Method	Sou. Data	Tar. Data	GC L	$\mathcal{T}$	PKU1 to PKU2 mIoU%	TST to TSU Acc%	mAP _r %	Acc%
Baseline	E2E	-	✓	-	-	51.8	68.1	24.1	20.6
SCS (s.a.)	Linear	✓	✓	✓	✓	64.0	74.9	28.3	26.7
SCS (s.a.)	E2E	✓	✓	✓	✓	71.4	84.6	35.1	33.2
SCS (z.a.)	Linear	✓	-	✓	✓	24.9	36.7	20.5	19.4
SCS (z.a.)	E2E	✓	-	✓	✓	25.5	38.2	21.2	20.9

Table 4. Action segmentation results with temporal domain adaptation. Models are tested on the original testing set of the target domain under the Cross-Subject protocol.

trimmed sequences and re-stitch into two restitched datasets, **Dataset I** (where we randomly group and concatenate trimmed sequences into restitched sequences without registering frame correspondence) and the **Dataset II** (where we use $\mathcal{T}$ to stitch sequences upon frame correspondence). The restitched datasets have the same size (i.e., the number of data samples) and same semantic information (i.e., the action permutations of each sequence) as the original dataset counterpart. As demonstrated in Figure 4 (a), training over a randomly stitched dataset, **Dataset I**, leads to drastic performance decrease compared to the baseline model that uses the original datasets. In contrast, **Dataset II** achieves comparable performance to the baseline over all three untrimmed datasets (PKU1, PKU2, TSU), suggesting that the *Skeleton-Stitch* method is able to generate plausible sequences in line with the contents of original data.

Data Augmentation. The use of stitched data is also investigated as a source for data augmentation in both end-to-end supervised learning and contrastive learning. More specifically, it augments a small untrimmed dataset in an efficient way. **Expanded Dataset** in Figure 4 (a) demonstrates that after combining the original dataset with **Dataset II**, substantial improvement is obtained over the three untrimmed datasets. Second, we verify the effectiveness of stitched data in contrastive learning, i.e., the *GCL* model in Figure 4 (b). For comparison, we also use classical data augmentation strategies: spatial transformation and temporal scaling (Guo et al. 2022), to develop a baseline model. The experimental comparison between the baseline and *GCL* in Figure 4 (b) suggests that classical augmentation favoured by previous works may lead to a limited transferability in the adaptation towards action segmentation. However, classical augmentations are orthogonal to our method, e.g., the combination of two augmentations (*Classi. Aug.* + $\mathcal{T}$) effectively boosts the adaptation performance on all datasets.

Cross-Sequence Test. In this section, we extend the discussion initiated in Figure 1. PKU-MMD v2 is a small dataset covering 26 action permutations that are commonly shared by training and testing samples under the Cross-View protocol (where data is split based on camera views). We follow a custom Cross-Sequence protocol for the dataset where the training data covers 15 action permutations, and

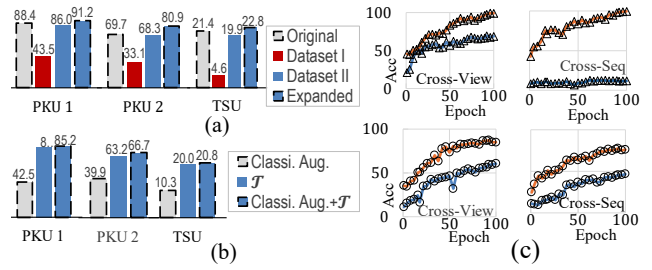


Figure 4: Effectiveness (Acc%) of Skeleton-Stitch in (a) end-to-end supervised learning, (b) GCL based pre-training; (c) Generalization performance without (Δ) and with ($\circ$) stitched sequences. **Orange** lines denote training accuracy across epochs and **blue** lines relate to test accuracy.

the testing data involves the remaining 11 permutations. We train models with two datasets: the original dataset and the expanded dataset using stitched sequences generated by *Skeleton-Stitch* (detailed in Supplementary Material). For comparison, we also provide results *w.r.t.* the Cross-View protocol. Figure 4 (c) presents the experimental results. It shows generalization discrepancy between action segmentation models evaluated under different evaluation protocols. **Upper row:** the baseline modes that are trained with original datasets demonstrate poor generalization (gap between training and testing accuracies) capability under the Cross-Sequence protocol. We conjecture that as the action permutation information is shared between the training data and the testing data under the Cross-View protocol, the baseline model can use observed action permutation as a shortcut to solve the segmentations on the testing data. Therefore, the model fails in the Cross-Sequence protocol which occludes the sharing of action permutation. **Lower row:** the enhanced models are trained with the expanded datasets and achieve relatively stable generalization performance in both the cross-view and cross-sequence protocols after being exposed to more action contexts via the use of *Skeleton-Stitch*.

Conclusion

Considering that learning segmentation models imposes expensive requirements for frame-wise data annotation, this paper introduces a temporal domain adaptation framework by treating trimmed skeleton videos as a research instance. It first proposes to generate spatially continuous and stitched skeleton sequences where different actions are smoothly concatenated. Second, it utilizes contrastive learning to learn fine-grained temporal segmentation embeddings from a source domain that can be represented by a skeleton encoder to transfer knowledge into a target domain for the task of action segmentation. The scalability of the proposed method is validated by comprehensive experimental results in supervised and zero-shot adaptation.

Acknowledgements

This research was supported in part by MITACS Accelerate and NSERC Discovery grants. The authors also acknowledge the collaboration of Spectronix Inc.

References

- Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.
- Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings IEEE Conference on Computer Vision Pattern Recognition*, 6299–6308.
- Chai, S.; Jain, R. K.; Liu, J.; Teng, S.; Tateyama, T.; Li, Y.; and Chen, Y.-W. 2024. A motion-aware and temporal-enhanced Spatial–Temporal Graph Convolutional Network for skeleton-based human action segmentation. *Neurocomputing*, 580, 127482.
- Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. *Proceedings IEEE/CVF International Conference on Computer Vision*, 13359–13368.
- Chen, Y.; Zhao, L.; Yuan, J.; Tian, Y.; Xia, Z.; Geng, S.; Han, L.; and Metaxas, D. N. 2022. Hierarchically self-supervised transformer for human skeleton representation learning. *European Conference Computer Vision*, 185–202.
- Dai, R.; Das, S.; Sharma, S.; Minciullo, L.; Garattoni, L.; Bremond, F.; and Francesca, G. 2022. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions Pattern Analysis Machine Intelligence*, 45(2), 2.
- Das, S.; Dai, R.; Koperski, M.; Minciullo, L.; Garattoni, L.; Bremond, F.; and Francesca, G. 2019. Toyota smarthome: Real-world activities of daily living. *Proceedings IEEE/CVF International Conference on Computer Vision*, 833–842.
- Farha, Y. A.; and Gall, J. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3584.
- Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 1933–1941.
- Filtjens, B.; Vanrumste, B.; and Slaets, P. 2022. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions Emerging Topics Computing*, 12(1), 1.
- Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; and Ding, R. 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. *Proceedings AAAI Conference on Artificial Intelligence*, 36(1), 1.
- Huang, X.; Zhou, H.; Wang, J.; Feng, H.; Han, J.; Ding, E.; Wang, J.; Wang, X.; Liu, W.; and Feng, B. 2023. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*.
- Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal Computer Vision*, 130(5), 5.
- Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. *Proceedings IEEE Conference on Computer Vision Pattern Recognition*, 156–165.
- Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3595–3603.
- Li, Y.; Li, Z.; Gao, S.; Wang, Q.; Hou, Q.; and Cheng, M.-M. 2023. A Decoupled Spatio-Temporal Framework for Skeleton-based Action Segmentation. *arXiv preprint arXiv:2312.05830*.
- Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021. 3d human action representation learning via cross-view consistency pursuit. *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4741–4750.
- Lin, L.; Zhang, J.; and Liu, J. 2023. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2363–2372.
- Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*.
- Liu, H.; Liu, Y.; Mu, T.-J.; Huang, X.; and Hu, S.-M. 2023. Skeleton-CutMix: Mixing Up Skeleton with Probabilistic Bone Exchange for Supervised Domain Adaptation. *IEEE Transactions Image Processing*.
- Ma, H.; Yang, Z.; and Liu, H. 2021. Fine-grained unsupervised temporal action segmentation and distributed representation for skeleton-based human motion analysis. *IEEE Transactions Cybernetics*, 52(12), 12.
- Niu, W.; Long, J.; Han, D.; and Wang, Y.-F. 2004. Human activity detection and recognition for video surveillance. *2004 IEEE International Conference Multimedia Expo (ICME) (IEEE Cat. No. 04TH8763)*, 1, 719–722.
- Piergiovanni, A.; and Ryoo, M. S. 2018. Learning latent super-events to detect multiple activities in videos. *Proceedings IEEE Conference on Computer Vision and Pattern*

Recognition, 5304–5313.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+d: A large scale dataset for 3d human activity analysis. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-contrastive 3D action representation learning. *Proceedings 29th ACM International Conference Multimedia*, 1655–1663.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings AAAI Conference on Artificial Intelligence*, 32(1), 1.

Yang, D.; Wang, Y.; Dantcheva, A.; Garattoni, L.; Francesca, G.; and Brémond, F. 2021. Unik: A unified framework for real-world skeleton-based action recognition. *BMVC 2021-The British Machine Vision Conference*.

Yang, D.; Wang, Y.; Dantcheva, A.; Kong, Q.; Garattoni, L.; Francesca, G.; and Bremond, F. 2023. Lac-latent action composition for skeleton-based action segmentation. *Proceedings IEEE/CVF International Conference on Computer Vision*, 13679–13690.

Zhang, J.; Lin, L.; and Liu, J. 2023. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. *Proceedings AAAI Conference on Artificial Intelligence*, 37(3), 3.

Zhang, S.; Tong, H.; Xu, J.; and Maciejewski, R. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1), 1.

4.4 Cross-Permutation Adaptation

4.4.1 Summary

The final enhancement brought in this section by **EGDA** introduces a dual-surrogate contrastive learning framework that exploits fine-grained extraction on action representations for skeleton data-based human action segmentation. It builds upon the fact that trimmed videos are easier to acquire and annotate (as they carry a single action), either from a public dataset or a controlled laboratory environment. While carrying essential dynamics of human motions, trimmed skeleton videos naturally offer good data representations for pre-training a feature extractor.

In implementation, the framework employs a novel data augmentation strategy, “Shuffle and Warp”, which generates diverse multi-action permutations to enrich the training data. Two surrogate tasks, Cross-Permutation Contrasting and Temporal Order Reasoning, are proposed to capture valid action representations across multiple temporal granularities. Experimental results highlight the effectiveness of this approach, showing that it outperforms state-of-the-art transfer learning methods in human action segmentation tasks.

4.4.2 Original contributions

The contributions of this component of **EGDA** consist of the following:

- A new data augmentation strategy, "Shuffle and Warp", that enables skeleton-based representation learning by utilizing trimmed sequences to create cross-permutation

action variations;

- The introduction of dual-surrogate contrastive learning for action segmentation. It develops two complementary surrogate tasks that target instance coherence and instance correlation of action representation;
- Extensive experiments on multiple action segmentation benchmarks, demonstrating state-of-the-art performance compared to existing methods.

4.4.3 Article

Research results are presented in “DuoCLR: Dual-Surrogate Contrastive Learning in Skeleton Data Based Human Action Segmentation” (H. Tian, P. Payeur), which was accepted to the *International Conference on Computer Vision (ICCV 2025)* at the time this thesis is finalized. The paper is provided below in the version that was accepted.

DuoCLR: Dual-Surrogate Contrastive Learning for Skeleton-based Human Action Segmentation

Anonymous Submission

Abstract

In this paper, a new contrastive representation learning framework is proposed to enhance action segmentation via pretraining using trimmed (single action) skeleton sequences. Unlike previous representation learning works that are tailored for action recognition and that develop isolated sequence-wise representations, the proposed framework focuses on exploiting multi-scale representations in conjunction with cross-sequence variations. More specifically, it proposes a novel data augmentation strategy, “Shuffle and Warp”, which exploits diverse multi-action permutations. The latter effectively assist two surrogate tasks that are introduced in contrastive learning: Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR). In optimization, CPC learns intra-class similarities by contrasting representations of the same action class across different permutations, while ROR reasons about inter-class contexts by predicting relative mapping between two permutations. Together, these tasks enable a Dual-Surrogate Contrastive Learning (DuoCLR) network to learn multi-scale feature representations optimized for action segmentation. In experiments, DuoCLR is pre-trained on a trimmed skeleton dataset and evaluated on an untrimmed dataset where it demonstrates a significant boost over state-the-art comparatives in both multi-class and multi-label action segmentation tasks. Lastly, ablation studies are conducted to evaluate the effectiveness of each component of the proposed approach.

1. Introduction

Contrastive learning [4, 5, 18] opens a new opportunity for taking advantage of the growing volume of unlabeled data for learning a rich latent feature representation aware of underlying data structures and semantics without human annotations. This approach effectively reduces the need for extensive labeled data in downstream task fine-tuning. In human-centric video understanding, human skeleton data [12, 29, 35] has become a widely used modality in contrastive learning as it depicts the trajectories of key body joints and provides a compact and robust depiction of hu-

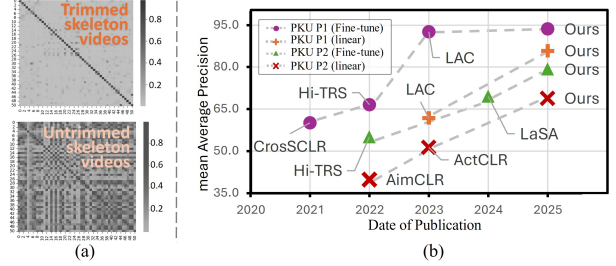


Figure 1. Limitation of traditional contrastive learning for skeleton-based action segmentation. (a): Distinct feature similarity matrices obtained from feature extraction on trimmed (upper) vs. untrimmed (lower) skeleton videos. Entry (i, j) represents the feature similarity (Eq. (1)) between action classes i and j . The feature extractor is learned on NTU-RGB+D using AimCLR [18]. Trimmed and untrimmed videos are both sampled from PKU P1. For untrimmed videos, action-specific features are extracted via post-segmentation in the feature space using frame annotations (detailed in Sec. 3.3). (b): Trend of contrastive learning performance in skeleton action segmentation by publication year. Please refer to Table 1 for detailed quantitative comparisons.

man motion that is resilient to environmental variations. Existing research primarily involves trimmed (single action) skeleton sequences into contrastive learning, where the model treats each sequence as an instance and learns intra-action temporal representations (i.e., *similarities*) through instance discrimination [6, 9, 45, 46]. The learned representations have been effectively transferred into trimmed-sequence-based downstream tasks such as *action recognition* [1, 9, 17, 20, 28, 38, 41, 47, 53] and *action retrieval* [13, 18, 24, 34, 48].

However, complex human-centric video understanding applications such as *action segmentation* often involve untrimmed long skeleton videos exhibiting multiple actions and complex action correlations. While focusing on sequence-wise instance discrimination tasks, traditional contrastive learning generally neglects the exploitation of inter-sequence temporal dependencies (i.e., *contexts*) that are critical for action segmentation. To highlight this limitation, Fig. 1a experimentally verifies that a state-of-the-art contrastive learning paradigm [18], which learns isolated sequence-wise representations, may lead to significantly

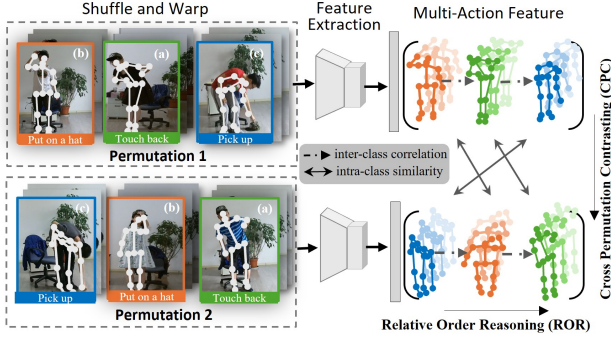


Figure 2. The proposed contrastive learning framework for action segmentation utilizes multi-action permutations. For example, six single-action skeleton sequences are randomly shuffled and warped into two sequences, Permutation 1 and Permutation 2, with three shared actions. Such a data augmentation introduces dynamic cross-sequence variations that enhance learning in two ways: i) CPC contrasts intra-class instance across different permutations, promoting permutation-invariant representations despite changing contexts (e.g., “Put on a hat” appears in both permutations but in different contexts); ii) ROR captures inter-class dependencies by mapping one permutation to another (e.g., learning the transformation from (b, a, c) to (c, b, a)).

different feature representations when applied to trimmed and untrimmed videos. We hypothesize that such a discrepancy results from the extraction of inconsistent action representations from dynamic action contexts permuted in untrimmed sequences, which leads to suboptimal transfer learning performance to action segmentation tasks (as illustrated in Fig. 1b).

In this work, a new contrastive learning framework is proposed to improve action segmentation on *untrimmed* videos by pretraining on *trimmed* skeleton videos. **First, the framework leverages multi-action permutations to fulfill cross-sequence data augmentation.** Previous single-action augmentations use simple sequence-wise modifications, e.g., spatial transformation [9, 18] and temporal scaling [25, 48], that generate low-level context-free data variations. This work rather introduces “Shuffle and Warp” to exploit high-level contextual variations: (1) “Shuffle” treats trimmed skeleton sequences as modular action elements that can be sampled and shuffled, introducing dynamic cross-sequence variations; and (2) “Warp” leverages geometrical transformations to concatenate the shuffled sequences into a view-consistent permutation of actions.

Second, two new surrogate tasks ensure multi-scale representation learning. To fully exploit the temporal variations present in multi-action permutations, this work introduces two contrastive learning surrogate tasks, Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR). As conceptually illustrated in Fig. 2, CPC learns *permutation-invariant intra-class similarities* via in-

stance discrimination where the representation of the same action class should be similar regardless of context variations arising from their neighboring segments. ROR focuses on the *permutation-aware inter-class contexts* by predicting the temporal mapping between action permutations, encouraging the model to learn relative positioning among actions, thereby adding a complementary regularization to the feature space.

Lastly, a hierarchical network facilitates sliding-window-free transfer learning. We amalgamate the key heuristics into a **Dual-Surrogate Contrastive LeaRning** (DuoCLR) network that can be trained end-to-end. Unlike previous works [7, 28, 47, 49] relying on a single visual encoder, DuoCLR employs a cascaded structure composed of a visual encoder and a temporal encoder. The additional temporal encoder enables a large temporal receptive field, allowing the model to capture long-range action dependencies without requiring sliding-window operations [7, 11, 47] on untrimmed skeleton videos. As a result, DuoCLR surpasses state-of-the-art methods in action segmentation (Fig. 1b). After pretraining, both encoders can be jointly fine-tuned as a feature extractor for downstream tasks.

The original contributions of the work are threefold. First, a new contrastive learning framework for skeleton data-based representation learning enables trimmed sequences to learn a fine-grained representation for action segmentation. Second, a new data augmentation strategy, “Shuffle and Warp”, for skeleton data introduces variations in between data sequences to improve feature learning. Third, the framework builds upon a dual-contrastive approach by combining CPC and ROR. It focuses on instance coherence and instance correlation, both critical for action segmentation.

2. Related work

Skeleton-based Action Segmentation. Compared to skeleton action recognition [6, 14, 25, 43, 45, 46] that aims at sequence-wise action classification, skeleton action segmentation [2, 3, 16, 26, 31] is inherently more challenging as the model must aggregate both action patterns and contexts among multiple actions permuted in a single sequence. To tackle the difficulty, recent research [16, 21, 21, 26] proposes to optimize an end-to-end model that incorporates frame-wise action classification into supervised training. Yet, training an action segmentation model from scratch relies heavily on frame-level annotations that are costly and time-consuming to generate [12, 29]. In this work, we circumvent the data annotation usage by leveraging trimmed skeleton videos for pretraining, where a contrastive learning network can leverage numerous action permutations to pre-learn a feature extractor relevant to action segmentation, and quick fine-tuning to untrimmed videos.

Skeleton Data Augmentation is primarily achieved by single-action skeleton modifications. For instance, studies [1, 7, 18, 20, 24, 28, 38, 41, 48, 52] delve into designing more effective single-action augmentations, e.g., “Shear”, “Crop”, “Spatial Flip”, “Spatial Rotate”, “Spatial Jitter”, “Axis Mask”, “Gaussian Noise”, and “Gaussian Blur”, demonstrating the effectiveness of carefully involving multiple augmentations in representation learning [18, 48]. Recent works also leverage multi-action skeleton augmentations, e.g., “Latent Action Composition” [47] and “Skeleton Actionlet” [28] are proposed to mix spatiotemporal statistics of two sequences; “Skeleton-Mix” [8, 19, 28, 30, 42, 50] is proposed to generate a synthetic sequence by merging both data and label from two sequences into a composable sequence. However, existing augmentations are tailored for action recognition and neglect exploiting high level contextual variations that are beneficial for action segmentation. To tackle the limitations, this work proposes utilizing multi-action permutations, a new data augmentation scheme which will be detailed in Section 3.2.

Skeleton-based Representation Learning [4, 5] is primarily achieved by learning action similarities from augmented skeleton views via effective surrogate tasks such as instance discrimination [1, 7–9, 13, 17–20, 24, 28, 41, 48, 52, 53], jigsaw puzzles [27, 37], adversarial learning [51], and language-video contrasting [21]. However, prior works mainly produce sequence-level representations, which are effective for simplified downstream tasks like action recognition but may be insufficient for action segmentation, where fine-grained feature representations are crucial. Recent efforts [7, 11, 47] attempt to mitigate this by applying sliding windows to generate fixed-length action clips from untrimmed sequences, enabling the use of standard action recognition models. However, this strategy introduces noisy boundaries, leading to degraded segmentation accuracy (experimentally verified in Section 4.3). These limitations highlight the need for a contrastive learning paradigm tailored for action segmentation.

3. Dual-Surrogate Contrastive Learning

3.1. Overview

Let $X \in \mathbb{R}^{T \times V \times C}$ be a trimmed skeleton video with T skeleton frames V skeleton joints, and C spatial dimensions, which covers the motion of a single action class. A contrastive learning paradigm aims to use trimmed skeleton videos to pretrain a feature extraction network $\mathcal{F}$ by which the features of the same action class should be similar. In previous studies [18, 20, 24], feature similarity between two sequences, X_i and X_j , is calculated as:

$$\text{sim}(\mathcal{F}(X_i), \mathcal{F}(X_j)) = \frac{\mathcal{F}(X_i) \mathcal{F}(X_j)}{|\mathcal{F}(X_i)| |\mathcal{F}(X_j)| \cdot \tau} \quad (1)$$

where τ acts as the temperature parameter. This work aims to pretrain a $\mathcal{F}$ using Dual-Surrogate Contrastive Learning (DuoCLR) and customize a task-specific (segmentation) layer $\mathcal{C}$ for a target domain composed of a small number of frame-wise labeled and untrimmed videos. Section 3.2 describes the Shuffle and Warp data augmentation technique for generating dynamic multi-action permutations. Section 3.3 details the network structure of $\mathcal{F}$. In Sections 3.4 and 3.5, DuoCLR introduces two surrogate tasks (as conceptually depicted in Fig. 3), Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR), respectively.

3.2. Multi-Action Permutations

Capturing temporal contexts is critical for action segmentation as it requires the model to discern the state of actions over a single untrimmed skeleton sequence. Since trimmed skeletal data lacks such context, Shuffle and Warp is introduced to generate dynamic multi-action permutations to support improved contrastive learning for action segmentation.

Temporal Shuffling was previously utilized to generate single-action augmentation where segments of a trimmed sequence are shuffled to solve a jigsaw puzzle task for action recognition [27, 33]. In this work, we rather involve multiple trimmed sequences at each augmentation of temporal shuffling to introduce cross-sequence variations. Specifically, let $\mathcal{X} = \{X^{(a)}, X^{(b)}, X^{(c)}, \dots\}$ be a set of trimmed skeleton sequences sampled from a data batch, we can generate a multi-action permutation as:

$$\text{Shuffle}(\mathcal{X}, \mathcal{P}) = \text{concat} \left(X^{\mathcal{P}[1]}, X^{\mathcal{P}[2]}, X^{\mathcal{P}[3]}, \dots \right) \quad (2)$$

where the function “*Shuffle*” first permutes the elements of $\mathcal{X}$ according to the permutation $\mathcal{P}$, e.g., $\mathcal{P} = (b, c, a, \dots)$, and then generates a single sequence using the tensor concatenation operation “*concat*”. The proposed data augmentation supports both labeled and unlabeled skeleton data where $a, b, c, \dots$ denotes action classes while labeled and sequence indices while unlabeled (discussed in Section 4.3).

Temporal Warping. However, skeleton sequences collected under different imaging configurations may lead to unnecessary data variations in the skeleton scale and poses [35]. A representation learning network may use such intricacies to develop shortcuts to fulfill the surrogate task of contrastive learning [18, 38]. To tackle the issue, temporal warping is newly proposed to eliminate the unnecessary data variations by using a pairwise sequence concatenation operation. Specifically, given two heterogenous sequences $X_i \in \mathbb{R}^{T_i \times V \times C}$ and $X_j \in \mathbb{R}^{T_j \times V \times C}$, it uses a rotation $\mathcal{R} \in \mathbb{R}^{C \times C}$, a translation $\mathcal{T} \in \mathbb{R}^{C \times 1}$, and a scaling factor $s \in \mathbb{R}$, to map X_j into the camera configuration of X_i . In this way, “*Warp*” is defined as:

$$\text{Warp}(X_i, X_j) = \text{concat} \left(X_i, s(X_j * \mathcal{R} + \mathcal{T}) \right) \quad (3)$$

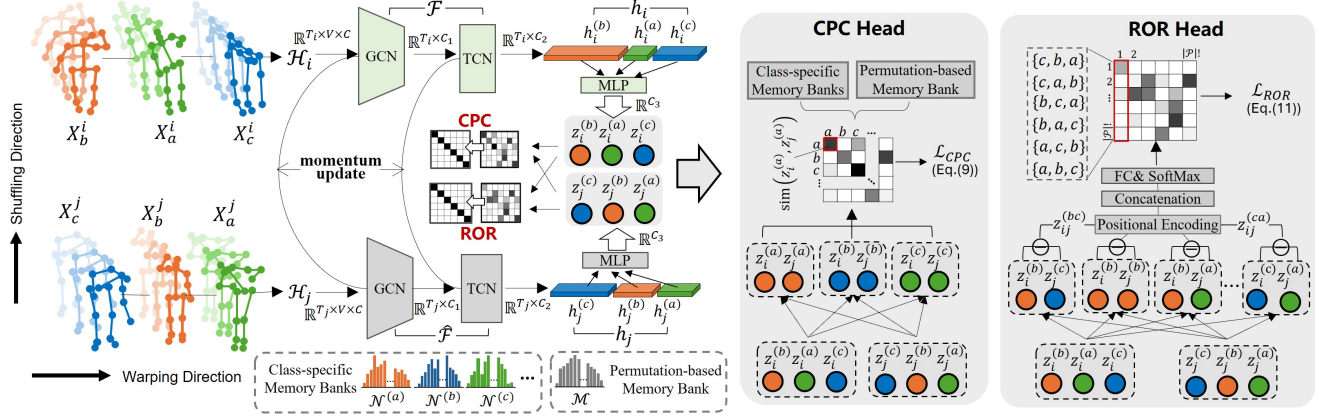


Figure 3. Computational workflow of the proposed DuoCLR approach. Trimmed single-action skeleton videos are augmented by using “Shuffle and Warp” on the input. In DuoCLR, the two surrogate tasks, “ROR” and “CPC”, share the feature encodings, $z_i^{(n)}$ and $z_j^{(n)}$, while fulfilling different pipelines for representation learning. By solving simultaneous CPC and ROR, the network learns a fine-grained action representation for action segmentation.

where $*$ denotes the batched dot product. Details regarding parameters estimation and sequence operation are available in supplementary material.

At last, we use “Shuffle and Warp” to generate the multi-action permutation $\mathcal{H}$ as follows

$$\mathcal{H} = \text{Warp}(\text{Shuffle}(\mathcal{X}, \mathcal{P})), \quad (4)$$

which denotes that $\mathcal{H}$ is generated by recursively operating the “Warp” operation on shuffled $\mathcal{X}$ under $\mathcal{P}$.

3.3. Feature Extraction

In this work, the feature extraction network $\mathcal{F}$ is composed of a visual encoder and a temporal encoder, as illustrated in Fig. 3.

Visual Encoder. It is common in the research community [13, 17, 20, 25, 28, 28] to use a graph convolutional network (GCN) [45] as a visual encoder to parse intermediate skeleton feature embeddings. In this work, we removed the temporal pooling layer of GCN to keep the feature’s temporal resolution, i.e., $\text{GCN}: \mathbb{R}^{T \times V \times C} \rightarrow \mathbb{R}^{T \times C_1}$. However, a vanilla GCN may be limited by their small temporal receptive fields [16], thus not exploiting contextual structures that are relevant to action segmentation.

Temporal Encoder. In order to effectively capture action relationships, we stack a temporal encoder onto the visual encoder to constructs $\mathcal{F}$. The temporal encoder adopts a temporal convolution network [15] that utilizes a series of dilated temporal convolutions to aggregate long-range temporal contexts of the hidden state from the visual encoder, i.e., $\text{TCN}: \mathbb{R}^{T \times C_1} \rightarrow \mathbb{R}^{T \times C_2}$.

At last, given a skeleton sequence input X , its hidden feature representation h is obtained as:

$$h = \underbrace{(\text{TCN} \circ \text{GCN})}_{\mathcal{F}}(X) \quad (5)$$

For feature extraction, DuoCLR differs from prior contrastive learning works that input single-action sequences. DuoCLR rather utilizes multi-action permutations for feature extraction prior to contrastive learning. As illustrated in Fig. 3, DuoCLR uses Shuffle and Warp (Eq. (4)) to generate a pair of multi-action permutations $\mathcal{H}_i$ and $\mathcal{H}_j$ of $\mathcal{X}$ with two random permutations $\mathcal{P}_i$ and $\mathcal{P}_j$. Using Eq. (5), it obtains multi-action representations $h_i = \mathcal{F}(\mathcal{H}_i)$ and $h_j = \mathcal{F}(\mathcal{H}_j)$, where $\mathcal{F}$ denotes a moving average of the feature extractor $\mathcal{F}$ and we follow the same implementation introduced in [18]. For notation simplicity, let’s assume the cardinality of $\mathcal{X}$ is 3, e.g., $\mathcal{X} = X^{(a)}, X^{(b)}, X^{(c)}$, and assume $\mathcal{P}_i = (b, a, c)$ and $\mathcal{P}_j = (c, b, a)$. In this way, DuoCLR generates multi-grained feature projections for contrastive learning by treating the feature representations, h_i and h_j , as concatenations of action-wise encodings: $h_i = \text{concat}(h_i^{(b)}, h_i^{(a)}, h_i^{(c)})$ and $h_j = \text{concat}(h_j^{(c)}, h_j^{(b)}, h_j^{(a)})$. Furthermore, an MLP-based projection head is used to generates *local* and *global* feature projections:

$$z_i^{(n)} = \text{MLP}[\phi(h_i^{(n)})]; z_i = \text{MLP}[\phi(h_i)] \quad (6)$$

where $n=a,b,c$ and ϕ denotes the temporal average pooling operation. Likewise, it generates $z_j^{(n)}$ and z_j .

3.4. Cross Permutation Contrasting (CPC)

While treating each action within a permutation as an instance, individual instances of the same action class can have different contextual neighbors (e.g., the action “Put on a hat” in Fig. 2 may appear next to different actions in each permutation). Traditional methods [18, 25, 28, 47] overlook such information and learn isolated sequence-wise representations, thereby achieving suboptimal transferability in action segmentation (as illustrated in Fig. 1). In

this work, DuoCLR utilizes multi-action permutations as data augmentation and employs Cross Permutation Contrasting (CPC) for self-supervised representation learning. Since the augmentation of action permutations expose the network to numerous action temporal contexts, CPC learns permutation-invariant temporal coherence where intra-class representations from the different permutations should be pulled together while inter-class representations should be pushed away.

Permutation Bank-based Instance Discrimination. In implementation, CPC employs local feature projections $z_i^{(n)} \in \mathbb{R}^{C_3}$ and $z_j^{(n)} \in \mathbb{R}^{C_3}$ as positive pairs and formulates single-action memory banks $\mathcal{N} = \{\mathcal{N}^{(m)}\}$ to store negative instances, each updated with $z_j^{(m)}$ where $m \neq n$. Afterwards, it learns short-term invariant representations for each instance regardless of surrounding context changes by optimizing a InfoNCE loss [5]:

$$\mathcal{L}_{\mathcal{N}}^{(n)} = -\log \frac{\text{sim}(z_i^{(n)}, z_j^{(n)})}{\text{sim}(z_i^{(n)}, z_j^{(n)}) + \sum_{z^- \in \mathcal{N}} \text{sim}(z_i^{(n)}, z^-)} \quad (7)$$

where z^- is a negative sample from the memory bank.

Second, CPC also involves a multi-action memory bank $\mathcal{M}$ in contrastive learning, which is updated by z_j . When $\mathcal{H}_i$ and $\mathcal{H}_j$ share the same permutation, i.e., $\mathcal{P}_i = \mathcal{P}_j$, it treats each permutation as an instance and encourages the network to learn long-term invariant representations of permutations. It calculates the similarity between permutation-wise positive pair, $z_i \in \mathbb{R}^{C_3}$ and $z_j \in \mathbb{R}^{C_3}$, and optimizes the second InfoNCE loss as:

$$\mathcal{L}_{\mathcal{M}} = -\log \frac{\text{sim}(z_i, z_j)}{\text{sim}(z_i, z_j) + \sum_{z^- \in \mathcal{M}} \text{sim}(z_i, z^-)} \quad (8)$$

Upon the formulation of Eq. (7) and Eq. (8), the final objective for CPC is formulated as:

$$\mathcal{L}_{CPC} = \lambda \sum_{n=a,b,c,\dots} \mathcal{L}_{\mathcal{N}}^{(n)} + (1 - \lambda) \mathcal{L}_{\mathcal{M}} \quad (9)$$

where λ is equals to 1 if $\mathcal{P}_i \neq \mathcal{P}_j$, and 0 otherwise.

3.5. Relative Order Reasoning (ROR)

Order reasoning of action videos is a fundamental task in video representation learning. Traditional approaches primarily focus on chronological order prediction within a single-action class, where models learn temporal coherence by reordering shuffled video frames [23] or clips [27, 33, 44]. However, in multi-action sequences, the order is inherently ambiguous, making traditional approaches not applicable. To address this, we introduce Relative Order Reasoning (ROR), a novel paradigm that leverages multi-action permutations. Instead of predicting a single "correct" order, ROR focuses on "relative" mapping between paired

action permutations and learns the capability of discerning the latent state of actions over long skeleton sequences, thus regularizing the feature space from a complementary direction to CPC.

In implementation, DuoCLR solves ROR via classification upon the mapping between $\mathcal{H}_i$ and $\mathcal{H}_j$. Specifically, since each pair of action permutations share the same actions, the relative mapping between the paired permutations is uniquely determined. For example, let $|\mathcal{P}_j|$ (where $\mathcal{P}_j = (c, b, a)$) be the action granularity of $\mathcal{H}_j$, there exist $|\mathcal{P}_j|! = 6$ factorial orders: $\{(c, b, a), (c, a, b), (b, c, a), (b, a, c), \dots\}$. The mapping from $\mathcal{H}_i$ to $\mathcal{H}_j$ is represented by the one-hot class label $\mathbf{1}(\mathcal{P}_i | \mathcal{P}_j) = [0, 0, 0, 1, 0, 0]^T$ while $\mathcal{P}_i = (b, a, c)$.

Positional Encodings . For optimization, ROR calculates Positional Encodings (PE) of $\mathcal{H}_i$ and $\mathcal{H}_j$. It uses local projections $z_i^{(n)} \in \mathbb{R}^{C_3}$ and $z_j^{(m)} \in \mathbb{R}^{C_3}$ (obtained via Eq. (6)) to calculate pair-wise feature comparison $z_{ij}^{(n,m)} = |z_i^{(n)} - z_j^{(m)}|$. Second, it generates the concatenation of $\{z_{ij}^{(n,m)} | n \in \mathcal{P}_i, m \in \mathcal{P}_j\}$ to obtain fine-grained position relations between h_i and h_j :

$$PE(h_i, h_j) = \text{concat}(z_{ij}^{(b,c)}, z_{ij}^{(b,b)}, z_{ij}^{(b,a)}, z_{ij}^{(a,c)}, \dots). \quad (10)$$

Such an operation allows the network to infer relative positioning of $\mathcal{H}_i$ and $\mathcal{H}_j$ by comparing their temporal structures in feature space, which is also illustrated in Fig 3. Afterwards, ROR utilizes a fully connected layer, $FC : \mathbb{R}^{C_3 \times |\mathcal{P}_j|^2} \rightarrow \mathbb{R}^{|\mathcal{P}_j|!}$, to optimize a Cross-Entropy loss:

$$L_{ROR} = -\mathbf{1}(\mathcal{P}_i | \mathcal{P}_j) \cdot \log(FC(PE(h_i, h_j))) \quad (11)$$

DuoCLR Loss. The final loss function for DuoCLR combines CPC and ROR losses, with a weighting factor α to balance their contributions:

$$L = \mathcal{L}_{CPC} + \alpha \mathcal{L}_{ROR} \quad (12)$$

The implementation of DuoCLR upon Eq. (12) leads to an expressive representation $\mathcal{F}$ that is fully transferable to action segmentation and will be comprehensively evaluated in Section 4.

4. Experiments

This section studies the application of DuoCLR in action segmentation. DuoCLR first involves a trimmed dataset for pretraining, followed by an evaluation phase where an untrimmed dataset is used to test the action segmentation performance. The study experimentally demonstrates the effectiveness of the proposed DuoCLR approach in comparison to state-of-the-art methods.

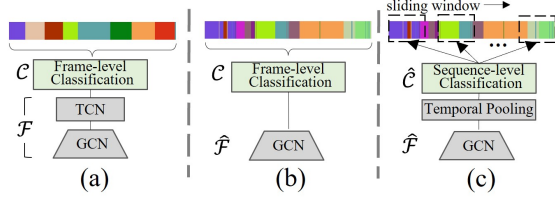


Figure 4. Transfer learning in skeleton-based human action segmentation. (a): Hierarchical structure introduced in Sec. 3.3. (b) and (c): Common structures used in existing works. See more details in supplementary material.

4.1. Implementations

- **Datasets.** Three **trimmed** skeleton datasets are considered for pretraining in DuoCLR: NTU RGB+D (NTU) [35], Toyota Smarthome (TS) [12] and Kinetics-400 (K4) [22]. Three **untrimmed** skeleton datasets are involved for evaluating the pre-trained models in the application of action segmentation, including PKU Multi-Modality Dataset (PKU) [29] (We use both versions: PKU1 and PKU2), Toyota Smarthome Untrimmed (TSU) [11], and Charades (CR) [36]. We report on mean Average Precision at IoU thresholds 0.1 ($\text{mAP}_{0.1}$) and 0.5 ($\text{mAP}_{0.5}$) [47], Top1 Accuracy (Acc) [45], mean Intersection over Union (mIoU) [28], and per-frame mAP (mAP) [11] for cross-view (CV) and cross-subject (CS) test cases of all untrimmed datasets.

- **Pretraining.** We adopt ST-GCN [45] as the skeleton visual encoder and MS-TCN [15] as the skeleton temporal encoder. In the pretraining, we use the entire (both training and testing) set of a trimmed dataset to learn the feature extractor $\mathcal{F}$. We adopt the same hyper-parameter values as in previous works [18, 20, 38], i.e., C_3 is set as 128 and τ in Eq. (1) as 0.07. The sizes of the two memory banks, $\mathcal{N}^{(m)}$ and $\mathcal{M}$, are set as 684 and 32,768, respectively. We set the granularity of each action permutation as 4 and α in Eq. (12) as 0.2. The framework is optimized with 50 epochs with a learning rate of 0.01. We implement Nearest Neighbors Mining [18] based on $\mathcal{N}^{(m)}$ after 40 epochs. More details are available in supplementary material.

- **Evaluation.** We consider two popular downstream tasks: multi-class action segmentation [16, 29] and multi-label (composite) action segmentation [11, 47], in each of which we learn a frame-wise classification (frame annotating) layer $\mathcal{C}$ over the pre-learned $\mathcal{F}$ (as illustrated in Fig. 4a). In multi-class segmentation, each skeleton frame is annotated with one action label. $\mathcal{C}$ is composed of a 1×1 convolution and SoftMax layer. It is optimized by the multi-class Cross-Entropy loss function with one-hot encoded frame annotations of the evaluation dataset. In composite action segmentation, each skeleton frame could be labeled with multiple action labels (e.g., “Cooking” and “Stirring”). $\mathcal{C}$ is composed of a 1×1 convolution and Sigmoid layer and is optimized using the binary Cross-Entropy loss

Models				Method Pretrain Eval.		CS%		CV%	
NTU → PKU1:						mAP _{0.1}	mAP _{0.5}	mAP _{0.1}	mAP _{0.5}
Baseline-I [45]	Linear	NTU	PKU1			56.0	50.3	56.7	54.9
LAC [47]	Linear	Poetics	PKU1			61.8	-	62.4	-
DuoCLR	Linear	NTU	PKU1			85.2	74.3	84.7	77.5
Baseline-II [16]	F-T	-	PKU1			89.6	86.4	92.1	88.9
MS2L [27]	F-T	NTU	PKU1			-	50.9	-	-
CMD [32]	F-T	NTU	PKU1			-	59.4	-	-
CrossSCLR [24]	F-T	NTU	PKU1			-	60.1	-	-
PCM3 [49]	F-T	NTU	PKU1			-	61.8	-	-
Hi-TRS [7]	F-T	NTU	PKU1			-	63.5	-	-
USDRL [40]	F-T	NTU	PKU1			-	75.7	-	-
LAC [47]	F-T	Poetics	PKU1			92.6	90.6	94.6	-
DuoCLR	F-T	NTU	PKU1			94.4	90.1	96.8	94.8
NTU → PKU2:						mIoU	Acc	mIoU	Acc
Baseline-I [45]	Linear	NTU	PKU2			30.7	47.2	26.4	44.7
AimCLR [18]	Linear	Poetics	PKU2			-	-	15.7	39.8
ActCLR [28]	Linear	NTU	PKU2			-	-	21.4	51.3
DuoCLR	Linear	NTU	PKU2			54.5	73.9	50.9	68.8
Baseline-II [16]	F-T	-	PKU2			50.9	69.3	45.6	66.0
Hi-TRS [7]	F-T	NTU	PKU2			-	-	-	55.0
LaSA [21]	F-T	-	PKU2			-	73.4	-	69.4
DuoCLR	F-T	NTU	PKU2			66.1	82.6	56.3	79.2

Table 1. Experimental comparison in the multi-class skeleton-based action segmentation task. Evaluation methods include Linear Evaluation and Fine-tune (F-T) Evaluation.

Models	Methods	Modality	TS → TSU		K4 → CR
			CS(%)	CV(%)	mAP(%)
MS-TCT [10]	F-T	RGB	33.7	-	25.4
Baseline-I [45]	Linear	Skeleton	18.2	11.4	6.1
LAC [47]	Linear	Skeleton	20.8	18.3	14.3
DuoCLR	Linear	Skeleton	24.1	19.8	19.1
Baseline-II [16]	F-T	Skeleton	28.3	20.3	21.8
LAC [47]	F-T	Skeleton	36.8	23.1	28.0
DuoCLR	F-T	Skeleton	35.1	26.3	33.9

Table 2. Experimental comparison in the composite skeleton-based action segmentation task.

with multi-hot encoded frame annotations of the evaluation dataset. In both applications, we use SGD to update $\mathcal{C}$ with Nesterov momentum 0.9 and learning rate 0.05.

4.2. Results

In the multi-class action segmentation task, we use NTU RGB+D for pretraining while using two datasets, PKU1 and PKU2, as the target for evaluation. A similar setting is considered in composite action segmentation where the trimmed TS and K4 are used as the pretraining domain while TSU and CR are used for evaluation. As with previous works, we use Linear Evaluation [24], where $\mathcal{F}$ is fixed while training $\mathcal{C}$, and Fine-tune (F-T) Evaluation [24], where $\mathcal{F}$ and $\mathcal{C}$ are trained jointly, to evaluate our approach. We report the experimental results in Table 1 and Table 2.

- **Comparison to State-of-the-art Methods.** We first compare the proposed DuoCLR to recent state-of-the-art methods [13, 14, 27, 46, 47] that involve a similar “trimmed-to-untrimmed” learning scenario. Experimental results demonstrate that DuoCLR outperforms most state-of-the-art comparatives across both multi-class (Table 1) and com-

Models	Methods	5% of PKU1		10% of PKU1	
		CS(%)	CV(%)	CS(%)	CV(%)
Baseline-I	Linear	33.9	36.0	41.6	42.7
DuoCLR	Linear	58.4	41.3	70.0	72.8
Baseline II	Fine-Tune	47.4	49.5	56.4	58.1
LAC[47]	Fine-Tune	73.9	75.4	79.8	81.1
DuoCLR	Fine-Tune	74.6	77.5	81.9	84.2

Table 3. Semi-supervised evaluation (mAP_{0.1}) with a low proportion of samples from the target domain.

Model	PKU1	PKU2	TSU	Charade
DuoCLR	63%	59%	55%	42%

Table 4. The ratio of labeled data is saved while reaching the same performance with the vanilla model that was trained on 100% data.

posite (Table 2) action segmentation tasks. We conclude that, since prior methods are limited in exploiting sequence-independent representations, the resultant feature space preserves a low transferability to action segmentation. In contrast, our model benefits from two efficient surrogate tasks, leading to better performance.

• **Linear and Fine-tune Evaluation.** A vanilla supervised pretraining paradigm involves learning a feature space $\mathcal{F}$ via a conventional end-to-end action recognition model [45] using trimmed skeleton videos without data augmentation. We treat such a paradigm as a baseline model (Baseline-I in Table 1 and Table 2) while making comparison with DuoCLR in Linear Evaluation. Experimental results demonstrate that our model significantly outperforms Baseline-I, which validates the effectiveness of the proposed DuoCLR approach in learning expressive action representations for the action segmentation tasks. We also compare DuoCLR with Baseline-II, an action segmentation model [16] that can be trained end-to-end on untrimmed datasets. DuoCLR, when fine-tuned, achieves substantial performance improvements over Baseline-II, as shown in Table 1 and Table 2 for both multi-class and composite tasks. This outcome highlights the advantage of incorporating DuoCLR’s pre-trained features, which enhance the model’s adaptability and performance in segmentation.

• **Semi-supervised Evaluation** The last experiment examines the transferability of our model while using fewer data samples in evaluation. Specifically, the experiment varies the ratios (5% and 10%) of samples from the target domain involved in transfer learning (i.e., Linear and Fine-tune Evaluation). Experimental results in Table 3 demonstrate that the model’s performance tends to increase monotonically along with the ratio of samples from the target domain. Table 4 suggests that DuoCLR can save up to 63% labeled data to reach the compatible performance compared to full-supervised learning.

4.3. Ablation studies

• **Effectiveness of Surrogate Tasks.** To assess the independent and combined effectiveness of these tasks, we conduct

NTU to PKU1	Method	CS (%)	CV (%)
Baseline-I	Linear	56.0	60.3
CPC	Linear	83.9	85.1
ROR	Linear	76.7	79.5
CPC+ROR (DuoCLR)	Linear	85.2	84.7

Table 5. Ablation study of two surrogate tasks: CPC (Cross Permutation Contrasting), ROR (Relative Order Reasoning), and combined DuoCLR. Results are reported as mAP_{0.1}.

PKU1 to PKU2	Methods	Pretraining	CS(%)	CV(%)
Baseline-II	Fine-Tune	-	69.3	66.0
DuoCLR	Linear	Trimmed PKU1	78.1	76.0
DuoCLR	Fine-Tune	Trimmed PKU1	78.0	77.6

Table 6. Experimental results (Acc) of pretraining using untrimmed datasets.

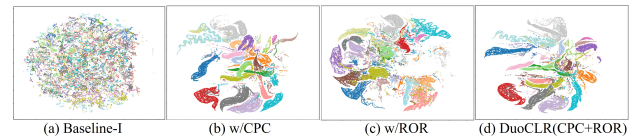


Figure 5. Feature clusters visualization by t-SNE [39]

experiments using each task individually and in combination, pretraining on NTU and evaluating on PKU1. First, either surrogate task demonstrates a significant performance improvement compared to the vanilla supervised pretraining paradigm (Baseline-I) as summarized in Table 5. In Fig. 5b and Fig. 5c, CPC and ROR demonstrate their independent effectiveness at learning more separate feature clusters compared to Baseline-I (Fig. 5a). Second, CPC achieves superior effectiveness over ROR likely because it directly enforces instance-level similarity, which is critical for clustering. Last, DuoCLR achieves its best learning result when combining the two tasks, i.e. 85.2% and 84.7% on the benchmarks of CS and CV, meanwhile demonstrating the best separability in feature clustering (Fig. 5d).

• **Pretraining with Untrimmed Datasets.** It is also effective to apply DuoCLR to an untrimmed dataset for representation learning. In this experiment, we consider PKU1 as the pretraining dataset where we trim each sequence into separate single-action sequences using their associated frame annotations. By applying “Shuffle and Warp” to these trimmed sequences, DuoCLR learns a feature extractor, $\mathcal{F}$, which is then evaluated on the original untrimmed PKU2 dataset. As shown in Table 6, DuoCLR outperforms the vanilla end-to-end supervised model (Baseline-II) in both linear and fine-tune evaluation protocols, showcasing the benefit of pretraining.

• **Effectiveness of Data Augmentation.** We learn a new model (DuoCLR-I) by incorporating two classical skeleton augmentations, “Shear” and “Crop” [24], with NTU used for pretraining and PKU1 for evaluation. As shown in Table 7, DuoCLR-I achieves an accuracy of 58.3%, sug-

NTU to PKU1	Shear, Crop	Shuffle	Warp	Supervised	CS%
DuoCLR		✓	✓	✓	85.8
DuoCLR-I	✓			✓	58.3
DuoCLR-II		✓		✓	73.8
DuoCLR-unsup	✓	✓	✓		81.6
DuoCLR-III	✓	✓	✓	✓	87.1

Table 7. Experimental comparison of DuoCLR and its variants using different skeleton data augmentation strategies.

Encoder	Classifier	Method	Params	Previous	Ours
GCN	Sequence-wise $\hat{\mathcal{C}}$	F-T	263K	63.5 [7]	77.6
GCN	Frame-wise $\mathcal{C}$	Linear	263K	61.8 [47]	81.9
GCN+TCN	Frame-wise $\mathcal{C}$	Linear	405K	-	85.2

Table 8. Experimental comparison between two types of classifiers. For $\hat{\mathcal{C}}$, the sliding window size is set as 300.

gesting that traditional augmentations alone provide limited transferability for action segmentation tasks. Our "Shuffle and Warp" method, however, is scalable to previous augmentations as we observe an orthogonal performance increase (1.3% over regular DuoCLR in PKU1) with the model DuoCLR-III after combining two types of augmentation. Additionally, we perform an ablation study on the roles of "Shuffle" and "Warp" individually. The DuoCLR-II variant, which uses only "Shuffle" (Eq. (2)), achieves an accuracy of 73.8%. When compared to DuoCLR, which uses both "Shuffle" and "Warp", the results demonstrate the importance of "Warp" which reduces irrelevant variations.

- **Pretraining with Unlabeled Datasets.** DuoCLR also supports unsupervised pretraining. For this approach, labeled proxies are created using sequence subscripts within each data batch. To learn an unsupervised DuoCLR model (DuoCLR-unsup, Table 7), we complement the augmentation with "Shear" and "Crop" to compensate for the lack of explicit labels. As seen in Table 7, regular DuoCLR outperforms the unsupervised variant (81.6% vs. 85.8% accuracy). This performance gap is likely due to the challenge of accurately identifying negative pairs in an unsupervised setting, which can lead to noise to the learning process.

- **DuoCLR is general to different network structures.** Fig. 4 illustrates two network structures commonly used in skeleton action segmentation: 1) Sliding-window-free structure (Fig. 4b) composed of a GCN encoder $\hat{\mathcal{F}}$ and a frame-wise classifier $\mathcal{C}$; 2) sliding window based structure (Fig. 4c) composed of a $\hat{\mathcal{F}}$ and a sequence-wise classifier $\hat{\mathcal{C}}$ while using sliding windows to pre-segment input sequences. Fig. 4a illustrates the network proposed in Sec. 3.3. For evaluation, we train three DuoCLR models with different structures and conclude the experimental results in Table 8). It demonstrates that DuoCLR is effective for different structures while surpassing previous works with clear margins.

- **Effectiveness of Action Granularity.** In this study, we treat the action granularity ($\mathcal{G}$) of action permutations as a hyperparameter and learn how it affects optimizations

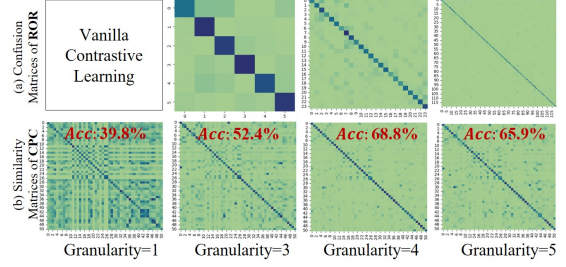


Figure 6. The confusion matrices obtained by ROR on different action granularities.

of DuoCLR. In practice, setting a high value of $\mathcal{G}$ will introduce diverse action contexts, however, it can also lead to high computational complexity, e.g., $\mathcal{O}(\mathcal{G})!$ for ROR while the size of the FC layer grows. In experiments, we increase $\mathcal{G}$ from 1 to 5 while preserving other implementations parameters as defined in Table 1. We illustrate confusion matrices (showing how ROR is good at mapping permutations) and similarity matrices (showing how CPC is good at learning consistent features) in Fig. 6, where Acc values evaluate the transferability of the resultant (CPC+ROR) network on PKU 2. While $\mathcal{G} = 1$, it degrades to a vanilla contrastive learning model [18]. While $\mathcal{G}$ increases, each model is capable of learning at different action granularity and the best model is achieved when $\mathcal{G}$ is around 4.

5. Conclusions

This paper introduces a dual-surrogate contrastive learning (DuoCLR) framework while exploiting fine-grained extraction on action representations for skeleton data-based human action segmentation. It applies "Shuffle and Warp" to generate various multi-action permutations where two surrogate tasks, Cross Permutation Contrasting and Relative Order Reasoning, learn permutation-invariant intra-class similarities and permutation-aware inter-class contexts. The learned action representations can be represented by a feature extractor which can be entirely transferred into action segmentation. Experimental results demonstrate that 1) DuoCLR enables training with both unlabeled and labeled trimmed skeleton videos; 2) DuoCLR outperforms state-of-the-art methods in multiple tasks; 3) DuoCLR can save 42%-63% frame-level labeled data to reach the compatible fully-supervised performance in action segmentation; 4) DuoCLR supports learning with RGB videos that can lead to better action segmentation performance compared to previous RGB based approaches.

References

- [1] Mohamed Abdelfattah, Mariam Hassan, and Alexandre Alahi. Maskclr: Attention-guided contrastive learning for robust action representation learning. In *Proceedings of*

- the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18678–18687, 2024. 1, 3
- [2] Shurong Chai, Rahul Kumar Jain, Jiaqing Liu, Shiyu Teng, Tomoko Tateyama, Yinhao Li, and Yen-Wei Chen. A motion-aware and temporal-enhanced spatial-temporal graph convolutional network for skeleton-based human action segmentation. *Neurocomputing*, 580:127482, 2024. 2
- [3] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan Al-Regib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 2
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3
- [5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3, 5
- [6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13359–13368, 2021. 1, 2
- [7] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. Hierarchically self-supervised transformer for human skeleton representation learning. In *European Conference on Computer Vision*, pages 185–202. Springer, 2022. 2, 3, 6, 8
- [8] Zhan Chen, Hong Liu, Tianyu Guo, Zhengyan Chen, Pinhao Song, and Hao Tang. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. *arXiv preprint arXiv:2207.03065*, 2022. 3
- [9] Seunggeun Chi, Hyung-gun Chi, Qixing Huang, and Karthik Ramani. Infogcn++: Learning representation by predicting the future for online skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 3
- [10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and François Brémond. Ms-tct: Multi-scale temporal con-transformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20051, 2022. 6
- [11] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2533–2550, 2022. 2, 3, 6
- [12] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 833–842, 2019. 1, 2, 6
- [13] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 525–533, 2023. 1, 3, 4, 6
- [14] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. Skeletr: Towards skeleton-based action recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13634–13644, 2023. 2, 6
- [15] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 4, 6
- [16] Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, 12(1):202–212, 2022. 2, 4, 6, 7
- [17] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv preprint arXiv:2303.06242*, 2023. 1, 3, 4
- [18] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 1, 2, 3, 4, 6, 8
- [19] Jinhua Hu, Yonghong Hou, Zihui Guo, and Jiajun Gao. Global and local contrastive learning for self-supervised skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3
- [20] Xiaohu Huang, Hao Zhou, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang, Xinggong Wang, Wenyu Liu, and Bin Feng. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*, 2023. 1, 3, 4, 6
- [21] Haoyu Ji, Bowen Chen, Xinglong Xu, Weihong Ren, Zhiyong Wang, and Honghai Liu. Language-assisted skeleton action understanding for skeleton-based temporal action segmentation. In *European Conference on Computer Vision*, pages 400–417. Springer, 2025. 2, 3, 6
- [22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6
- [23] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017. 5
- [24] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4741–4750, 2021. 1, 3, 6, 7
- [25] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional

- networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 2, 4
- [26] Yunheng Li, Zhongyu Li, Shanghua Gao, Qilong Wang, Qibin Hou, and Ming-Ming Cheng. A decoupled spatio-temporal framework for skeleton-based action segmentation. *arXiv preprint arXiv:2312.05830*, 2023. 2
- [27] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2490–2498, 2020. 3, 5, 6
- [28] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2363–2372, 2023. 1, 2, 3, 4, 6
- [29] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 1, 2, 6
- [30] Hanchao Liu, Yuhe Liu, Tai-Jiang Mu, Xiaolei Huang, and Shi-Min Hu. Skeleton-cutmix: Mixing up skeleton with probabilistic bone exchange for supervised domain adaptation. *IEEE Transactions on Image Processing*, 2023. 3
- [31] Hao Ma, Zaiyue Yang, and Haoyang Liu. Fine-grained unsupervised temporal action segmentation and distributed representation for skeleton-based human motion analysis. *IEEE Transactions on Cybernetics*, 52(12):13411–13424, 2021. 2
- [32] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, pages 734–752. Springer, 2022. 6
- [33] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016. 3, 5
- [34] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18846–18856, 2023. 1
- [35] Amir Shahrourdy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1, 3, 6
- [36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 6
- [37] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13328–13338, 2021. 3
- [38] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1655–1663, 2021. 1, 3, 6
- [39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7
- [40] Wanjiang Weng, Hongsong Wang, Junbo He, Lei He, and Guosen Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. *arXiv preprint arXiv:2412.09220*, 2024. 6
- [41] Cong Wu, Xiao-Jun Wu, Josef Kittler, Tianyang Xu, Sara Ahmed, Muhammad Awais, and Zhenhua Feng. Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5949–5957, 2024. 1, 3
- [42] Linhua Xiang and Zengfu Wang. Joint mixing data augmentation for skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3
- [43] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10276–10285, 2023. 2
- [44] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10334–10343, 2019. 5
- [45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2, 4, 6, 7
- [46] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580*, 2021. 1, 2, 6
- [47] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Lac-latent action composition for skeleton-based action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13679–13690, 2023. 1, 2, 3, 4, 6, 7, 8
- [48] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3427–3435, 2023. 1, 2, 3
- [49] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Prompted contrast with masked motion modeling: Towards versatile 3d ac-

- tion representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7175–7183, 2023. [2](#), [6](#)
- [50] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Shap-mix: shapley value guided mixing for long-tailed skeleton based action recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1688–1696, 2024. [3](#)
- [51] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. [3](#)
- [52] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3825–3833, 2023. [3](#)
- [53] Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13913–13922, 2023. [1](#), [3](#)

4.5 Supplementary Information

This section provides additional details about the experiments conducted in Chapter 4.

4.5.1 *Additional implementation details*

The proposed components of the **EGDA** framework were implemented using PyTorch and deployed on an NVIDIA GTX 3090 GPU with 24 GB of memory. For the contrastive learning implementation (Section 4.3), Stochastic Gradient Descent (SGD) was used, with a weight decay of 0.0001 and a learning rate of 0.01. For supervised learning (Section 4.1, Section 4.3, and Section 4.4), the framework utilized SGD with Nesterov momentum (0.9), a weight decay of 0.0005, and a learning rate of 0.05. These configurations were chosen to ensure stable optimization and effective training for the respective tasks.

4.5.2 *Computational cost*

This section also evaluates the computational cost of each model proposed in the **EGDA** framework. Specifically, Figure 4.4.1 illustrates the relationship between computational costs and model accuracy for different **EGDA** implementations. The horizontal axis represents the training time (sec/iteration), while the vertical axis displays accuracy based on the Acc Top-1 metric [41] for human activity recognition.

- **EGDA-a** and **EGDA-b** focus on cross-view and cross-sensor adaptation for human action classification, as detailed in Sections 4.1 and 4.2. They consider a single source domain, NTU RGB+D, and three datasets, PKU-MMD part I, part II, and NW-UCLA,

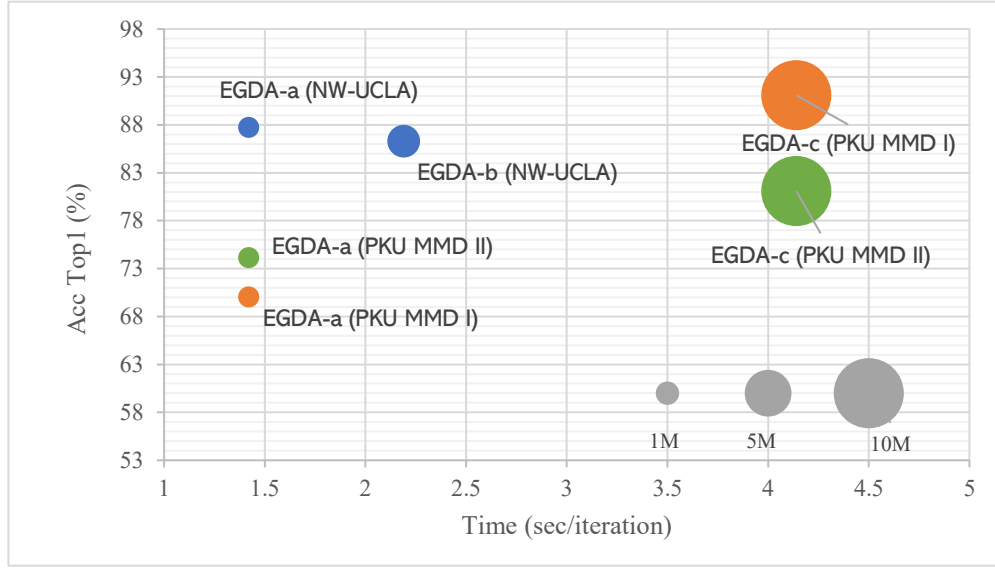


Figure 4.5.1. Comparison of the computational costs for different components implemented as part of the proposed **EGDA** framework. The comparison includes models trained on two datasets: PKU MMD [74] (Parts I and II, represented by orange and green circles, respectively) and NW-UCLA [73] (represented by blue circles). The bubble size corresponds to the number of parameters in the model, with scales for 1M, 5M, and 10M parameters displayed as a reference.

as the target domains. Results are reported under the “cross-subject” protocol in Figure 4.5.1. It is observed that **EGDA-a** supports quicker learning iterations than **EGDA-b**, making the approach particularly suitable for scenarios where computational efficiency is critical. Both models also remain fairly compact.

- **EGDA-c** is proposed for cross-sequence adaptation as formulated in Section 4.3. Unlike **EGDA-a** and **EGDA-b** that rely on trimmed videos for both training and testing,

EGDA-c is trained on trimmed videos, but tested on longer untrimmed videos in human action segmentation. Figure 4.5.1 compares the model's performance on two target datasets (PKU-MMD part I and part II). First, the method leads to larger models that require longer training while the base action segmentation network imposes an additional temporal segmentation layer compared to an action recognition network (as detailed in Section 4.3). Moreover, **EGDA-c** achieves competitive accuracy in comparison with its counterpart adaptation frameworks, though accuracy varies with the complexity of the target dataset.

The overall results provide effective trade-offs between computational efficiency and model accuracy, highlighting the scalability and performance of the proposed **EGDA** framework across different tasks and datasets.

Chapter 5. Conclusion

To emphasize the main contributions brought by this research to the area of domain adaptation for computer vision applications, this chapter presents general conclusions on the proposed frameworks detailed in Chapter 3 and Chapter 4. Section 5.1 summaries the benefits carried by the two original domain adaptation frameworks, **ESDA** and **EGDA**, respectively, while Section 5.2 addresses some of the common features and observed limitations of the proposed approaches. The latter lead to opportunities discussed in Section 5.3 for future works to further refine the concepts reported in this thesis.

5.1 Summary of Contributions

This thesis introduces original methodologies and frameworks to enhance domain adaptation performance when dealing with real-world deep learning-based computer vision applications. The research is motivated by the fact that the fundamental effectiveness of deep learning models trained on large-scale benchmarks can be substantially compromised when transposed to a real-world target domain whose data distribution is dynamic and diverse. In this thesis, two novel domain adaptation frameworks are proposed, named Enhanced Scene Domain Adaptation (**ESDA**) and Enhanced Graph Domain Adaptation (**EGDA**) respectively. They are instantiated in two computer vision applications: image semantic segmentation and human activity analysis. The overall objective is to support

practical deployment of deep learning solutions to challenging engineering applications where data collection and data annotation are expensive and sometimes intractable.

*5.1.1 Contributions from **ESDA***

The proposed methods under **ESDA** aim to tackle scene domain shifts in outdoor street scenes while achieving reliable deployment of image semantic segmentation models in dynamic and open real-world settings. First, **the pixel-wise weighted adaptation framework** (Section 3.1) contributes an effective fine inter-domain discrepancy regularization mechanism for cross-domain semantic segmentation. Compared to previous works, it approaches the over-adaptation issue without placing any constraints on the conventional deep learning semantic segmentation framework. Second, **the prototypical knowledge-oriented adaptation framework** (Section 3.2) integrates pixel-wise adaptation into a soft adversarial adaptation loss which allows semantic segmentation models to treat target domain pixels particularly and precisely. The introduction of prototypical knowledge and soft adversarial adaptation loss are both new to the domain adaptation research community. Third, **the target-dependent classifier adaptation** (Section 3.3) introduces a new method for effective adaptation to the target domain clusters even given a certain degree of failure on domain alignment. The model can readily be applied to existing adaptation models and leads to state-of-the-art adaptation performance. Extensive experimental validation demonstrates the effectiveness of the three frameworks under

ESDA across various and challenging real-world semantic segmentation tasks while achieving significant improvements in performance.

*5.1.2 Contributions from **EGDA***

The proposed approaches under **EGDA** target graph domain shifts in skeleton-based human action recognition, aiming at improving the adaptability of human action recognition models in uncontrolled, real-world scenarios, e.g., varying camera perspectives, different sensor configurations, untrimmed skeleton videos, and various actions ordering. For instance, **the cross-view adaptation framework** (Section 4.1) innovatively leverages the benefits of Contrastive Prototype Learning to extract expressive action prototypes from a labeled source domain and Temporal Adaptation Modeling to refine a functional action recognition model to the target domain by associating and exploiting pseudo-labels. **The cross-sensor adaptation framework** (Section 4.2) presents novel research on learning an adaptive skeletal human action recognition model that effectively functions with heterogenous skeleton graph configurations. The formulation of **cross-sequence adaptation** (Section 4.3) introduces a new approach for learning human action segmentation models from short, coarsely labeled sequences for a target application domain where only a small number of unlabeled long sequences are available for model refinement. Finally, **cross-permutation adaptation** (Section 4.4) contributes new ways to improve on skeleton-based action segmentation via pre-training on changing action permutations composed of more widely available trimmed skeleton videos. Extensive experimental

results validate the practicality and adaptability of the proposed frameworks contributing to **EGDA** for addressing various graph domain shifts and achieving competitive performance on human action recognition accuracy.

5.2 Discussion

5.2.1 Common characteristics of *ESDA* and *EGDA*

Although designed to specialize on distinct types of domain shift and targeting different real-world computer vision applications, the two original overarching frameworks, **ESDA** and **EGDA**, share multiple characteristics and contribute jointly to the vast problem of domain adaptation.

- **Model-Based Adaptation.** Both **ESDA** and **EGDA** are model-based adaptation methods that function as external components to their base network (FCN or GCN respectively) and do not require modification to the architecture of the base network. In this way, the proposed methods are compatible with different state-of-the-art deep learning architectures. Second, during testing, the resulting models optimized either by **ESDA** or **EGDA** will retain their learned adaptability in the target domain while ensuring that no additional adaptation steps are required at test time. That ensures that the adaptation based on **ESDA** and **EGDA** does not affect the inference speed of the original model.

- Data Assumption.** In this thesis, both frameworks use copious data from the source domain to support the transfer learning of a model into a label-scarce target domain. For instance, **ESDA** leverages well-annotated source data to learn an optimal model while utilizing unlabeled target data to measure pixel-wise inter-domain discrepancy and obtain an adaptive model. Also, **EGDA** proposes a contrastive learning framework to pre-learn action semantic patterns from plenty of labeled skeletal data samples from the source domain while a successive phase can fine-tune an optimal model for the target domain using minimal data. Both frameworks contribute to circumvent the dependency on data in the target domain.
- Consistent Methodology.** In this thesis, both frameworks adopt a consistent methodology aimed at developing domain-invariant features, which are critical for addressing domain shifts. In **ESDA**, adversarial learning trains a domain classifier to distinguish between the source and target domains, guiding the network to generate indistinguishable domain features. In contrast, **EGDA** employs contrastive learning, focusing on maximizing the similarity between paired samples to learn robust feature representations. Both approaches promote the generation of domain-invariant features, a strategy which effectively mitigates scene and graph domain shifts.

In conclusion, these common characteristics reflect a unified philosophy in the design of the domain adaptation frameworks to tackle domain shifts across diverse computer vision tasks.

5.2.2 Limitations

This subsection identifies potential limitations that may limit the proposed frameworks' robustness and applicability in real-world applications and remain beyond the scope of this thesis.

- Each proposed framework builds upon a certain type of learning paradigm. For instance, **ESDA** utilizes pixel-wise adversarial learning to regularize the training of adaptation networks at a finer level. **EGDA** leverages supervised contrastive learning to discover high-level human action semantic patterns from plenty of unlabeled skeletal data samples from the source domain. An integrated framework that amalgamates multiple learning techniques was not investigated.
- The proposed frameworks are evaluated using specific types of network backbones. For instance, DeepLab-v2-ResNet101 [64] is used in **ESDA** as the semantic segmentation backbone while **EGDA** utilizes ST-GCN as the action analysis backbone. The effectiveness of other popular backbone architectures, e.g., FCN8s-VGG16 [60], ResNet-38 [110], DRN-D-105 [111], AS-GCN [40], and 2s-GCN [112], was not investigated.

- In practical applications, computer vision systems often require adaptation frameworks capable of processing multiple modalities while addressing different domain shifts, e.g., a human behavior-based autonomous driving solution would need to combine scene domain adaptation and graph domain adaptation for street image semantic segmentation and pedestrian action recognition at the same time. However, either **ESDA** or **EGDA** is tailored as a custom framework for a certain type of data modality, i.e., RGB image or skeleton video. They are not inherently designed to transpose easily to multi-modal adaptation scenarios.
- Finally, this work focuses on single-source-single-target domain adaptation. Multi-source domain or multi-target domain adaptations can be challenging to the proposed methodologies. Moreover, each DA framework assumes that the label space (object category or action classes) between the source and the target domain is the same. However, such an assumption could not hold generally in real-world applications.

5.3 Future work

Aiming at dealing with the aforementioned limitations, the thesis provides a strong basis for several key directions for future research.

- **Unified Feature Alignment.** Future efforts should incorporate multiple learning strategies, e.g., adversarial learning and contrastive learning within a single

framework, e.g., **ESDA** or **EGDA**. Such a study may significantly enhance adaptation performance for individual data modalities by exploiting complementary learning outcomes.

- **Hybrid Multi-Modal Adaptation** . The current adaptation frameworks are limited to two specific modalities, i.e., image and skeleton. However, many real-world systems naturally collect multi-modal data streams in which developing hybrid domain adaptation frameworks that simultaneously adapt across multiple modalities will be crucial. Future research should explore hybrid domain adaptation frameworks that integrate both scene and graph knowledge, e.g., a hybrid system that can simultaneously input scene image data and graph skeleton videos, leading to an advanced DA technique for real-world and more complex computer vision tasks.
- **Network Backbones**. The current adaptation frameworks rely on a single backbone architecture (FCN or GCN). Future work should expand the scope to incorporate advanced network backbones. For instance, it is possible to accommodate Vision-Language Models (VLMs) such as CLIP in domain adaptation, which will integrate language guidance to enhance model generalization by injecting high-level semantic context that is often hard to learn solely from visual patterns. Such extensions would enhance the scalability of the frameworks, enabling their application to a broader range of real-world applications.

- **Multi-Source Domain Adaptation.** The proposed frameworks rely heavily on the quality and diversity of the source domain data samples. However, a single source domain would be insufficiently representative which will hinder generalization to the target domain. Future research should explore aggregating different source domains to leverage the diverse knowledge shared across domains, making the frameworks more generalizable for deployment in data-scarce environments.
- **More applications.** Future work can expand the developed domain adaptation frameworks to a broader range of computer vision tasks, such as point cloud segmentation, 3D object detection, and depth estimation for scene parsing, while addressing domain variations across different geometric and structural properties.

Additional References

The list below provides references to additional sources from the literature that are not already included in the seven publications that form this thesis.

- [1] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, 1999, vol. 2, pp. 1033–1038.
- [2] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Computer Vision—ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*, 2000, pp. 751–767.
- [3] P. Orbanz and J. M. Buhmann, "Nonparametric Bayesian image segmentation," *International Journal of Computer Vision*, vol. 77, pp. 25–45, 2008.
- [4] J. Tighe and S. Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," *International Journal of Computer Vision*, vol. 101, pp. 329–349, 2013.
- [5] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
- [6] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3*, 1994, pp. 151–158.
- [7] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*, 2010, pp. 708–721.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [9] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.
- [10] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, pp. 261–318, 2020.

- [11] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.
- [12] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [13] Y. LeCun and others, “LeNet-5, convolutional neural networks,” URL: <http://yann.lecun.com/exdb/lenet>, vol. 20, no. 5, p. 14, 2015.
- [14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [15] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.
- [16] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, “Graph convolutional networks: a comprehensive review,” *Computational Social Networks*, vol. 6, no. 1, pp. 1–23, 2019.
- [17] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [18] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, “Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 673–682, 2016.
- [19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: a large video database for human motion recognition,” in *2011 International conference on computer vision*, 2011, pp. 2556–2563.
- [20] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, “Segmentation and recognition using structure from motion point clouds,” in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part I 10*, 2008, pp. 44–57.
- [21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
- [22] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International journal of computer vision*, vol. 111, pp. 98–136, 2015.
- [23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

- [24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, pp. 740–755.
- [25] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kotschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.
- [26] K. Muhammad, T. Hussain, H. Ullah, J. Del Ser, M. Rezaei, N. Kumar, M. Hijji, P. Bellavista, and V. H. C. de Albuquerque, "Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [27] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.
- [28] S. Ma, Z. Yuan, Q. Wu, Y. Huang, X. Hu, C. H. Leung, D. Wang, and Z. Huang, "Deep into The Domain Shift: Transfer Learning through Dependence Regularization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [30] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 128–145, 2021.
- [31] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
- [32] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha, "Deep learning algorithm for autonomous driving using googlenet," in *2017 IEEE intelligent vehicles symposium (IV)*, 2017, pp. 89–96.
- [33] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.
- [34] J.-W. Hsieh, Y.-T. Hsu, H.-Y. M. Liao, and C.-C. Chen, "Video-based human movement analysis and its application to surveillance systems," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 372–384, 2008.
- [35] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763)*, 2004, vol. 1, pp. 719–722.
- [36] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcms in the wild: Pixel-level adversarial

- and constraint-based adaptation,” *arXiv preprint arXiv:1612.02649*, 2016.
- [37] B. Filtjens, B. Vanrumste, and P. Slaets, “Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks,” *IEEE Transactions on Emerging Topics in Computing*, 2022.
 - [38] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, “Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor,” in *2011 annual international conference of the IEEE engineering in medicine and biology society*, 2011, pp. 1831–1834.
 - [39] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise topology refinement graph convolution for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13359–13368.
 - [40] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Actional-structural graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3595–3603.
 - [41] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.
 - [42] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE transactions on neural networks and learning systems*, 2021.
 - [43] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, “The history began from alexnet: A comprehensive survey on deep learning approaches,” *arXiv preprint arXiv:1803.01164*, 2018.
 - [44] L. Wang, S. Guo, W. Huang, and Y. Qiao, “Places205-vggnet models for scene recognition,” *arXiv preprint arXiv:1508.01667*, 2015.
 - [45] S. Targ, D. Almeida, and K. Lyman, “Resnet in resnet: Generalizing residual architectures,” *arXiv preprint arXiv:1603.08029*, 2016.
 - [46] Y. Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images,” in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, 2001, vol. 1, pp. 105–112.
 - [47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
 - [48] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR ’05)*, 2005, vol. 1, pp. 886–893.
 - [49] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human

- action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.
- [50] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, “A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 793–802.
 - [51] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, “Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7287–7296.
 - [52] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
 - [53] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [54] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3d human pose in the wild using imus and a moving camera,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 601–617.
 - [55] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, “Sparse inertial poser: Automatic 3d human pose estimation from sparse imus,” in *Computer graphics forum*, 2017, vol. 36, no. 2, pp. 349–360.
 - [56] B. Chen, H. Ji, Z. Wang, B. Filtjens, C. Wang, W. Ren, B. Vanrumste, and H. Liu, “Language-Assisted Human Part Motion Learning for Skeleton-Based Temporal Action Segmentation,” *arXiv preprint arXiv:2410.06353*, 2024.
 - [57] H. Ji, B. Chen, X. Xu, W. Ren, Z. Wang, and H. Liu, “Language-Assisted Skeleton Action Understanding for Skeleton-Based Temporal Action Segmentation,” in *European Conference on Computer Vision*, 2025, pp. 400–417.
 - [58] Y.-H. Li, K.-Y. Liu, S.-L. Liu, L. Feng, and H. Qiao, “Involving distinguished temporal graph convolutional networks for skeleton-based temporal action segmentation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 647–660, 2023.
 - [59] X. Tian, Y. Jin, Z. Zhang, P. Liu, and X. Tang, “Spatial-temporal graph transformer network for skeleton-based temporal action segmentation,” *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44273–44297, 2024.
 - [60] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, “Learning to adapt structured output space for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

- [61] L. Xu, Q. Wang, X. Lin, and L. Yuan, "An efficient framework for few-shot skeleton-based temporal action segmentation," *Computer Vision and Image Understanding*, vol. 232, p. 103707, 2023.
- [62] D. Yang, Y. Wang, A. Dantcheva, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond, "Lac-latent action composition for skeleton-based action segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13679–13690.
- [63] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1992–2001.
- [64] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [65] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 2016, pp. 102–118.
- [66] M. Wrenninge and J. Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," *arXiv preprint arXiv:1810.08705*, 2018.
- [67] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.
- [68] M. Antico, N. Balletti, G. Laudato, A. Lazich, M. Notarantonio, R. Oliveto, S. Ricciardi, S. Scalabrino, and J. Simeone, "Postural control assessment via Microsoft Azure Kinect DK: An evaluation study," *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106324, 2021.
- [69] L. Caruso, R. Russo, and S. Savino, "Microsoft Kinect V2 vision system in a manufacturing application," *Robotics and Computer-Integrated Manufacturing*, vol. 48, pp. 174–181, 2017.
- [70] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, 2012, pp. 20–27.
- [71] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, 2012, pp. 28–35.
- [72] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International journal of robotics research*, vol.

- 32, no. 8, pp. 951–970, 2013.
- [73] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.
 - [74] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, “Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding,” *arXiv preprint arXiv:1703.07475*, 2017.
 - [75] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, “Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.
 - [76] P. Merriaux, Y. Dupuis, R. Bouteau, P. Vasseur, and X. Savatier, “A study of vicon system positioning performance,” *Sensors*, vol. 17, no. 7, p. 1591, 2017.
 - [77] G. Nagymáté and R. M. Kiss, “Application of OptiTrack motion capture systems in human movement analysis: A systematic literature review,” *Recent Innovations in Mechatronics*, vol. 5, no. 1., pp. 1–9, 2018.
 - [78] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
 - [79] M. Hein and M. Andriushchenko, “Formal guarantees on the robustness of a classifier against adversarial manipulation,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [80] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.
 - [81] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
 - [82] A. Piergiovanni and M. S. Ryoo, “Learning latent super-events to detect multiple activities in videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5304–5313.
 - [83] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [84] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, “MSeg: A composite dataset for multi-domain semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2879–2888.
 - [85] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Moment matching for multi-source domain adaptation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.
 - [86] S. Sun, H. Shi, and Y. Wu, “A survey of multi-source domain adaptation,” *Information Fusion*, vol. 24, pp. 84–92, 2015.
 - [87] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A.

- Bharath, “Generative adversarial networks: An overview,” *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.
- [88] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
 - [89] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
 - [90] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, 2020, pp. 1597–1607.
 - [91] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang, “Hierarchical contrast for unsupervised skeleton-based action representation learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 1, pp. 525–533.
 - [92] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, “Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 1, pp. 762–770.
 - [93] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, “Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition,” *Information Sciences*, vol. 569, pp. 90–109, 2021.
 - [94] F. M. Thoker, H. Doughty, and C. G. Snoek, “Skeleton-contrastive 3D action representation learning,” in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 1655–1663.
 - [95] B. Xu, X. Shu, and Y. Song, “X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition,” *IEEE Transactions on Image Processing*, vol. 31, pp. 3852–3867, 2022.
 - [96] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.
 - [97] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*, 2017, pp. 1126–1135.
 - [98] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.
 - [99] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, “Learning to generalize: Meta-learning for domain generalization,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.
 - [100] D. Li and T. Hospedales, “Online meta-learning for multi-source and semi-supervised domain adaptation,” in *European Conference on Computer Vision*, 2020,

- pp. 382–403.
- [101] X. Guo, C. Yang, B. Li, and Y. Yuan, “Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3927–3936.
 - [102] H. Tian, S. Qu, and P. Payeur, “Unsupervised Pixel-Wise Weighted Adversarial Domain Adaptation,” in *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part I*, 2021, pp. 586–600.
 - [103] H. Tian, S. Qu, and P. Payeur, “A prototypical knowledge oriented adaptation framework for semantic segmentation,” *IEEE Transactions on Image Processing*, vol. 31, pp. 149–163, 2021.
 - [104] H. Tian, S. Qu, and P. Payeur, “Learning a target-dependent classifier for cross-domain semantic segmentation: Fine-tuning versus meta-learning,” *Pattern Recognition*, vol. 147, p. 110091, 2024.
 - [105] M. Chen, H. Xue, and D. Cai, “Domain Adaptation for Semantic Segmentation With Maximum Squares Loss,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
 - [106] H. Tian and P. Payeur, “Unsupervised Temporal Adaptation in Skeleton-Based Human Action Recognition,” *Algorithms*, vol. 17, no. 12, p. 581, 2024.
 - [107] H. Tian, J. Dickens, and P. Payeur, “Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition,” in *Proceedings of the Conference on Robots and Vision*, 2024.
 - [108] H. Tian and P. Payeur, “Stitch Contrast and Segment_Learning a Human Action Segmentation Model Using Trimmed Skeleton Videos,” in *Proceedings of the AAAI conference on artificial intelligence*, 2025, vol. 39, no. 7.
 - [109] H. Tian and P. Payeur, “Semi-Supervised Adaptation for Skeletal-Data-Based Human Action Recognition,” *Engineering Proceedings*, vol. 58, no. 1, p. 25, 2023.
 - [110] L. Du, J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye, and X. Zhang, “Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 982–991.
 - [111] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, “Confidence regularized self-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5982–5991.
 - [112] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12026–12035.