UNIVERSITY OF OTTAWA

Neural Architectures and Approaches for Person Re-Identification in Autonomous Surveillance System

By

Hamzah AlGhamdi

A thesis presented to the University of Ottawa in fulfillment of the thesis requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering

uOttawa

School of Electrical Engineering and Computer Science

Faculty of Engineering

Declaration of Authorship

I, Hamzah AlGhamdi, declare that this thesis, "Neural Architectures and Approaches for Person Re-Identification in Autonomous Surveillance System" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while a candidate for a research degree at the University of Ottawa.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Ottawa University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Hamzah Alghamdi

Date: 2026-05-28

Abstract

Video-based person re-identification is the task of recognizing the same person across different cameras and video clips. It is an important part of automated surveillance systems because a person may disappear from one camera, become partially blocked, or reappear later in another camera view. This task is difficult because people can look different when the camera angle, lighting, distance, background, or body pose changes.

This thesis studies how video-based person re-identification can be improved under three practical deployment conditions. First, it presents a resource-aware method for cases where only a small amount of labelled data is available. The method begins with one labelled video clip per person and gradually adds reliable, automatically labelled examples to improve training while keeping the model efficient. Second, it develops a fully supervised method for cases where labelled training data are available. This model uses information from the entire video sequence as well as local body-region details to improve recognition under occlusion, pose changes, and background clutter. Third, it introduces a transfer-based method for cases where a model trained on one camera network must be used in another network without new manual labels. This method helps the model recognize people more reliably when the camera setup, viewing angle, lighting, or background changes.

Experiments on several video-based person re-identification datasets show that the proposed methods improve performance under different levels of supervision, computational cost, and camera variation. Overall, the thesis provides a practical study of how video-based person re-identification systems can be designed for label-scarce, fully labelled, and cross-camera deployment settings.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Robert Laganiere. His invaluable guidance, continuous support, and insightful feedback have been instrumental in completing this thesis. His expertise and dedication have greatly enriched my research experience.

I also extend my sincere thanks to Prof. Abdulmotaleb Elsaddik, with whom I had the privilege to collaborate for a limited time. His insightful advice and encouragement have been incredibly beneficial, and I am grateful for the opportunity to work with him.

I am immensely grateful to my friends and colleagues at the VIVA and Discovery labs. Their unwavering support, stimulating discussions, and camaraderie have been a constant source of motivation and inspiration. Special thanks to Wassim Al-Ahmar, Yahya Alaa, Majed Alauidy, and Abdulrahman AlShariff for their constant help and encouragement, which have significantly contributed to the successful completion of this thesis.

My heartfelt thanks go to the University of Jouf for their assistance and support during my scholarship. I am especially grateful to Prof. Turki AlGhamdi, Prof. Omar AlRuily, and Prof. Madallah AlRuily for their guidance and support, which have ensured that my experience during my studies abroad was both enriching and fulfilling. Their dedication to my academic growth has been truly inspiring.

Additionally, I would like to acknowledge the administrative and technical staff at the University of Ottawa for their assistance throughout my studies. Their support has been crucial in navigating various logistical and technical challenges.

Lastly, I am profoundly thankful to my siblings for their steadfast support and guidance. Their encouragement has been a pillar of strength throughout

this journey. I owe a debt of gratitude to my parents for their sacrifices, which have allowed me to pursue my academic goals despite various challenges. This achievement is a testament to their love and dedication, and I hope it makes them proud.

To everyone who has contributed to my academic journey, I extend my deepest appreciation. Your support and encouragement have been invaluable, and I am eternally grateful.

Contents

List of Figures

List of Tables

List of Abbreviations

AP	**A**verage **P**recision
AFC	**A**ttention-aware **F**eature **C**omposition
CDS	**C**ombined **D**epth **S**pace
CFD	**C**omputational **F**luid **D**ynamics
CIR	**C**ross-**I**mage **R**epresentation
CRF	**C**onditional **R**andom **F**ield
COSAM	**C**o-segmentation-based **A**ttention **M**odule
CVSL	**C**ontrastive **V**iewpoint-aware **S**hape **L**earning
CMC	**C**umulative **M**atching **C**haracteristic
CNN	**C**onvolutional **N**eural **N**etwork
DCRN	**D**eep **C**onvolutional **R**esidual **N**etwork(s)
DRC-Net	**D**ynamic **R**esidual **C**onvolutional **Net**work(s)
DTL	**D**eep **T**ransductive **L**earning
DIMN	**D**omain-**I**nvariant **M**apping **N**etwork
ELU	**E**xponential **L**inear **U**nit
GELU	**G**aussian **E**rror **L**inear **U**nit
GLU	**G**ated **L**inear **U**nit(s)
FC	**F**ully **C**onnected
FLOP	**F**loating-point **O**perations per **S**econd
FEN	**F**eature **E**xtraction **N**etwork
FFN	**F**eature **F**usion **N**etwork
FUL	**F**ully **U**nsupervised **L**earning
GAP	**G**lobal **A**verage **P**ooling

GAN	**G**enerative **A**dversarial **N**etwork
GP	**G**eneralization **P**ercentile
GMM	**G**aussian-**M**ixture **M**odeling
GeM	**G**eneralized-**M**ean
GRL	**G**radient **R**eversal **L**ayer
GRU	**G**ated **R**ecurrent **U**nit
INP	**I**nverse **N**egative **P**enalty
IDE	**ID**-**D**iscriminative **E**mbedding
LN	**L**ayer **N**ormalization
LOMO	**L**ocal **M**aximal **O**ccurrence
LSFP	**L**ocalized **S**alience **F**eature **P**atch
LSTRL	**L**ong **S**hort-**T**erm **R**epresentation **L**earning
LSTM	**L**ong **S**hort-**T**erm **M**emory
MHA	**M**ulti-**H**ead **A**ttention
MLP	**M**ultilayer **P**erceptron
mINP	**M**ean **I**nverse **N**egative **P**enalty
mAP	**M**ean **A**verage **P**recision
MARS	**M**otion **A**nalysis and **R**e-identification **S**et
MSCAN	**M**ulti-**s**cale **C**ontext-**A**ware **N**etwork
MSA	**M**ultihead **S**elf-**A**ttention
MTMCT	**M**ulti-**T**arget **M**ulti-**C**amera **T**racking
NCC	**N**ormalized **C**ross **C**orrelation
NN	**N**earest **N**eighbor
NLA	**N**on-**L**ocal **A**ttention **B**locks
NLA-CNN	**C**onvolutional **N**eural **N**etwork with **N**on-**L**ocal **A**ttention **B**locks
NLP	**N**atural **L**anguage **P**rocessing
PCB	**P**art-based **C**onvolutional **B**aseline
PiT	**P**yramid **T**ransformer

PR	**P**recision-**R**ecall **C**urve
PPA	**P**ose-**G**uided **P**art **A**ttention
PDC	**P**ose-**D**riven **D**eep **C**onvolutional
R1	**R**ank-**1** **A**ccuracy
ReLU	**R**ectified **L**inear **U**nit
ResNet-50	50-layer **Res**idual **Net**work
RRS	**R**estricted **R**andom **S**ampling
RGB	**R**ed, **G**reen, and **B**lue
REET	**R**egion-**E**nhanced **T**okenization
RPP	**R**efined **P**art **P**ooling
RNN	**R**ecurrent **N**eural **N**etworks
Re-ID	**R**eidentification
SE	**S**queeze-and-**E**xcitation
SIR	**S**ingle-**I**mage **R**epresentation
SGD	**S**tochastic **G**radient **D**escent **O**ptimizer
SAN	**S**patiotemporal **A**ttention **N**etwork
StS	**S**et-to-**S**et
SVM	**S**upport **V**ector **M**achine
TCSS	**T**emporal **C**lip **S**hift and **S**huffle
TFS	**T**emporal **F**eature **S**hift
TAP	**T**emporal **A**verage **P**ooling **L**ayer
TKS	**T**emporal **K**ernel **S**election
TMT	**T**rigeminal **T**ransformers
UDA	**U**nsupervised **D**omain **A**daptation
ViT	**V**ision **T**ransformer
VPPF	**V**ideo **P**atch **P**art **F**eature
VAE	**V**ariational **A**utoencoders

Chapter 1

Introduction

The rapid development of surveillance systems has been crucial in meeting the growing public demand for public safety from government agencies and law enforcement authorities. Widespread camera networks are increasingly prevalent in urban areas, including the downtown core, transit hubs, and commercial districts, playing a vital role in maintaining community safety by discouraging criminal activity, aiding emergency response, and supporting public order. As human behaviour is recorded in surveillance systems, the need to track individuals across multiple locations becomes apparent, making it an essential next step in improving surveillance measures for people of interest.

Since most surveillance systems rely on multiple cameras, tracking individuals from one frame to the next requires identifying the same individual across many video tracklets (i.e., sequences of video frames that capture an individual's movements). Person re-identification (Re-ID) is crucial in this process. The goal of Re-ID is to accurately re-associate tracklets belonging to the same individual, including fragmented paths across cameras and reappearances within the same network. Re-ID must remain reliable under various conditions, including partial occlusion, scale variations, viewpoint changes, illumination variations, combined viewpoint and scale changes, and pose variations, as illustrated in Figure 1.1. Currently, most surveillance systems rely on human monitoring and analysis, a process that is time-consuming

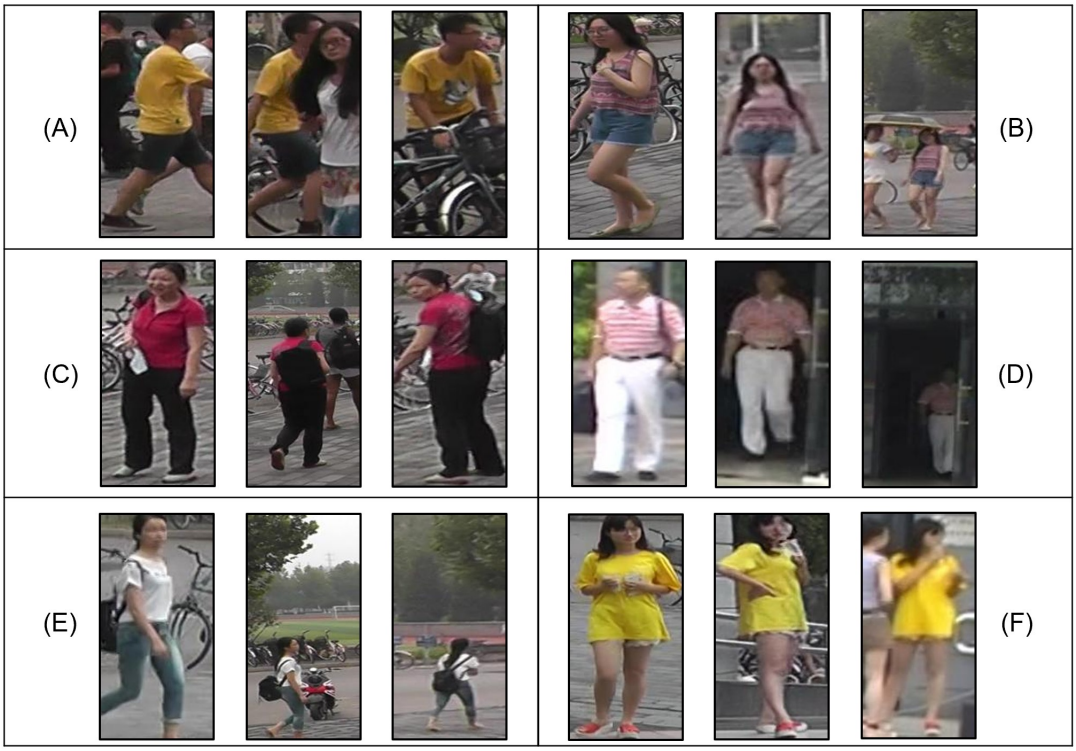


FIGURE 1.1: Examples of challenging conditions in person re-identification: (A) The same person under partial occlusion of different body regions. (B) The same person at different scales. (C) The same person from different viewpoints. (D) The same person under different illuminated conditions. (E) A combination of scale and viewpoint variations. (F) The same person in different body poses.

and may be prone to inaccuracies under these conditions. Hence, establishing autonomous Re-ID solutions can help minimize the impact of these variations and more accurately predict an individual's movements across all tracklets. The complexity of accurately identifying people repeatedly from various conditions has sparked substantial scientific interest. This ability is essential for comprehensive monitoring, enabling authorities to track an individual's movements across wide regions effortlessly, thereby improving public safety and security measures.

Initially, Re-ID systems were predominantly image-based due to simpler acquisition settings and algorithms centered on static appearance matching. In image-based Re-ID, the task is typically formulated as retrieving all images of

a person of interest from a gallery using one still image as the query. While this formulation is simpler and more effective in controlled settings, it has inherent limitations because it only provides a single visual snapshot of a person. As a result, a person's appearance is heavily influenced by momentary factors such as viewpoint, pose, occlusion, blur, and illumination. For example, a person photographed from the front in one image may appear very different when observed later from behind, or from different angles, making direct matching less reliable if only a single still image is available.

In contrast, video-based Re-ID benefits from the temporal continuity of a tracklet, where the same person is observed across multiple consecutive frames. It provides a more detailed and reliable representation of a person because the different frames may capture complementary visual information. As a person moves, the sequence may reveal multiple body orientations, allowing the Re-ID model to observe characteristics that would not be visible in a still image, such as the transition from a back-facing to a front-facing view. Similarly, if one frame occludes a body part, another frame in the same sequence may show it. As a result, video-based Re-ID is better suited for surveillance applications because it can aggregate information over time, reduce reliance on a single outlier frame, and produce more reliable identity matching across multiple viewpoints, pose variations, and partial occlusions.

1.1 Study Context and Rationale

In recent years, video-based Re-ID has received considerable attention, but its roots may be traced back to multi-target multi-camera tracking (MTMCT), developed over a decade ago [1]. MTMCT is intended to detect whether a person is present in the field of view of various non-overlapping cameras (Figure 1.2). Although image-based Re-ID has historically been the preferred method, video-based Re-ID is currently attracting growing attention from

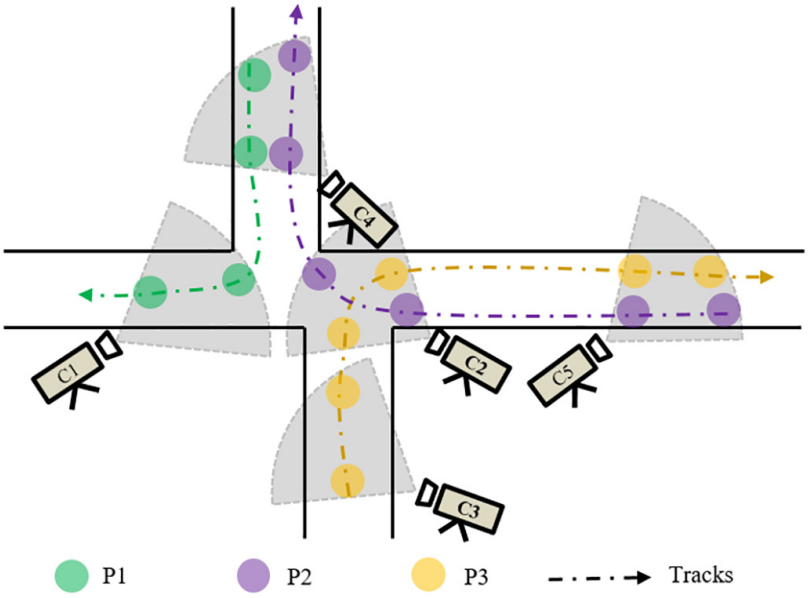


FIGURE 1.2: Multi-camera surveillance network illustration of person Re-ID.

the academic community. This shift is due to its ability to give rich temporal information, resulting in a more comprehensive portrayal of a person's movements over time. A single image cannot adequately represent a lengthy video sequence, especially given the improvements in tracking reliability. Therefore, using the full video sequence provides a richer and more reliable representation of an individual's appearance and movement over time.

Nevertheless, despite the tangible benefits of video-based Re-ID, it encounters other substantial obstacles. These issues can pertain to both general computer vision tasks and those specialized to the specific needs of Re-ID. Common difficulties include changes in illumination [2], occlusion [3], [4], background clutter [5], as well as scale and resolution variations [6], [7]. While background clutter refers to irrelevant objects that create visual interference, occlusion occurs when objects overlap, potentially resulting in errors. Moreover, the resolution variability between cameras can diminish the quality of an individual's features, while lighting alterations can modify a person's visual appearance, both of which make consistent identification more

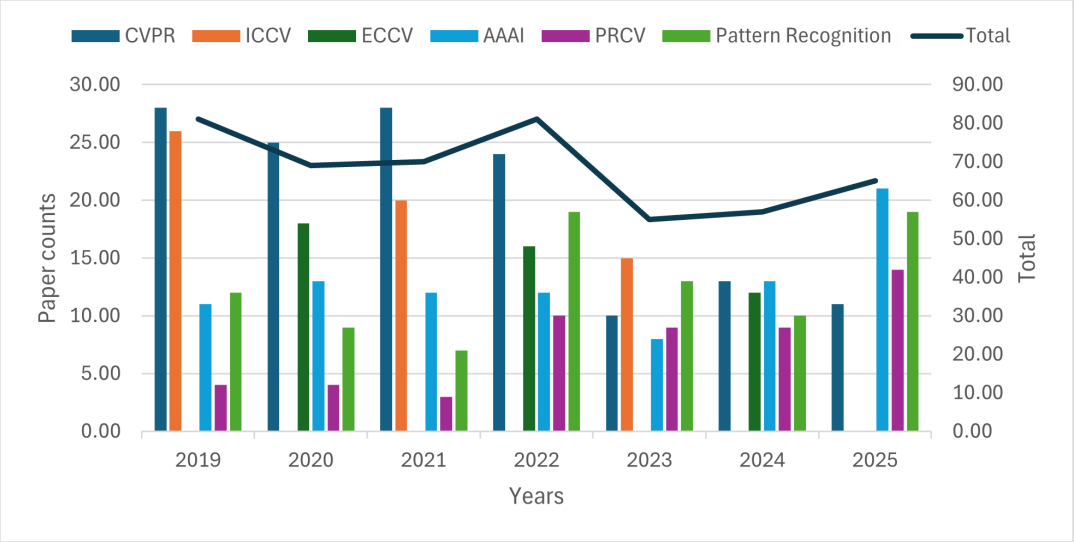


FIGURE 1.3: The number of person Re-ID papers on top conferences and journals over recent years.

challenging. In addition, misalignment of body parts, divergent viewpoints, and an individual's different poses are all video-based Re-ID-specific challenges that diminish model accuracy and lead to identification errors [8], [9]. Furthermore, achieving camera-viewpoint invariance remains a significant challenge, despite being partially addressed by current video-based Re-ID implementations.

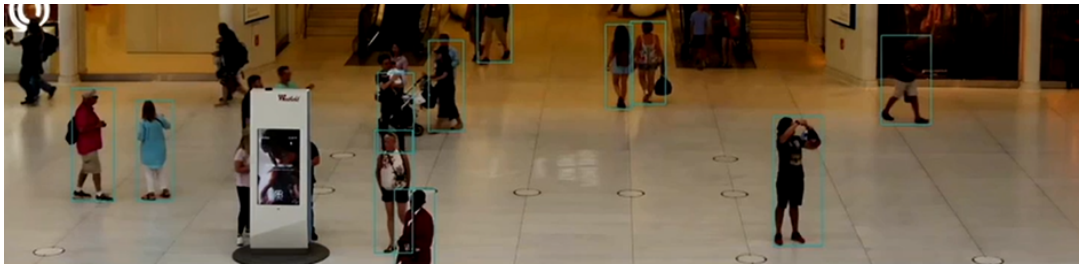
All of these challenges shape the robustness and universality of present Re-ID models. Such models may excel under controlled conditions where these challenges are minimized, but real-life scenarios inevitably present these issues. Therefore, for Re-ID systems to be genuinely effective, they must address these challenges. Overcoming them is essential to developing robust, reliable Re-ID systems that perform consistently in diverse, uncontrolled environments. Figure 1.3 shows that person Re-ID remains a well-established and active research area, with a steady stream of publications in major conferences and journals over recent years.

1.2 Research Objectives and Scope

This thesis studies person re-identification in an exclusively video-based context, with an emphasis on realistic deployment constraints in multi-camera surveillance networks. In practice, Re-ID is typically invoked as a re-association module when a tracker loses confidence due to occlusion, disappearance from the field of view, fragmented trajectories, or a person reappearing in another camera. This thesis is organized around two deployment axes: the severity of viewpoint and domain variations across cameras and the level of computational and annotation resources available. From this perspective, three gaps motivate this work. First, many recent approaches rely on increasingly large, resource-intensive models, while practical deployments often require lightweight, cost-sensitive solutions. Second, much of the literature still emphasizes older benchmarks, even though newer datasets are larger, more diverse, and more challenging in viewpoint and scale variations. Third, real deployments often have labelled data for one camera network but not for a new target network, creating a domain-shift problem due to differences in viewpoint, illumination, and camera setup.

Figure 1.4 clarifies the scope of this thesis within a complete surveillance system. Although the figure illustrates one example of identity recovery after tracking failure, the broader problem studied in this thesis is video-based person Re-ID within a camera network, including both fragmented trajectories within a single camera and the re-association of a person who reappears in another camera. Accordingly, tracking is assumed to be handled beforehand, and the contributions of Chapters 4–6 are studied as three operating regimes of the same broader video-based Re-ID problem rather than solutions to the full end-to-end surveillance pipeline.

Rather than treating Chapters 4–6 as three unrelated models, this thesis



(A) Object detection stage, where the system detects pedestrians using green bounding boxes.

(B) Tracking stage, where the tracker (denoted as the yellow box) assigns an identity to each detected person.

(C) Tracking failure stage, where the tracker loses identities #31 and #33 and assigns them new identities #42 and #44.

(D) System state immediately before the Re-ID step is triggered.

(E) Re-identification stage, where the pedestrians #42 and #44 are matched back to #31 and #33 respectively.

FIGURE 1.4: End-to-end real-time person Re-ID pipeline showing detection, tracking, tracking failure, and identity recovery through Re-ID.

frames them as three operating regimes of the same broader video-based person Re-ID problem. The unifying perspective is that practical Re-ID design is governed mainly by two deployment axes: the severity of viewpoint and domain variations across cameras, and the level of computational and annotation resources available. From this perspective, the literature reveals three corresponding gaps. First, many recent methods rely on increasingly large, resource-intensive models, which motivates Chapter 4, where we study a cost-sensitive regime with minimal labelling, lightweight backbones, and reduced training complexity. Second, much of the literature still emphasizes older benchmarks, while newer datasets introduce a larger scale, stronger variability, and more challenging viewpoint conditions. This inspires Chapter 5, which studies the ideal, fully supervised regime and explores how far accuracy can be pushed when resource constraints are relaxed. Third, real deployments often have labelled data for one camera network but not for a new target network, creating a domain-shift problem due to differences in viewpoint, illumination, and camera setup. This leads to Chapter 6, where transfer-based adaptation is examined to reduce the need for new labelling. These three regimes, when combined, form a unified thesis: video-based person Re-ID should be designed as a series of solutions tailored to different combinations of viewpoint variation, resource availability, and deployment constraints rather than a single fixed solution.

Consequently, the scope of this thesis is limited to video-based Re-ID under these three operating regimes and their corresponding architectural parameters. It is assumed that tracking is performed beforehand and that an individual's appearance remains reasonably consistent throughout the surveillance period. To keep the Re-ID problem operationally tractable in practical deployments, this thesis also adopts a sliding-window gallery assumption. Under this assumption, the gallery cannot grow indefinitely; instead, identities are maintained only within an active temporal window and are removed once

TABLE 1.1: Summary of the three operating regimes studied in this thesis.

Regime	**Setting**	**Gap**	**Objective**
Cost-sensitive / one-shot	Minimal labels, limited resources.	Recent methods often rely on large and resource intensive models.	lightweight and efficient video ReID solution.
Ideal supervised	Full labels, relaxed resource constraints.	The literature still emphasizes older and less challenging datasets.	Maximize retrieval accuracy on newer and more challenging benchmarks.
Transfer / adaptation	Labelled source, unlabelled target, and severe domain and viewpoint shift.	Limited robustness to domain shift in new camera-network setups.	Reduce target-domain labelling requirements while improving cross-domain robustness.

they are no longer relevant to the current surveillance event.

Importantly, even within this active temporal window, the number of identities present in the system is not known in advance. New individuals may enter the camera network at any time, while others may leave, become occluded, or disappear from the active surveillance area. Therefore, the Re-ID system must operate under an open-set condition in which the active gallery size is bounded by time but the number of identities at any given moment remains variable and unknown.

This can be implemented, for example, by clearing identities at the end of an operational period or after a predefined inactivity timeout. Consequently, if the same person reappears after the window expires, the system treats that observation as a new identity instance rather than enforcing long-term identity persistence across arbitrarily long time spans. This assumption does not eliminate the open-set nature of person Re-ID, since new identities may still appear at any time within the active window, but it bounds the gallery size and clarifies that the thesis focuses on video-based Re-ID under temporally managed gallery growth rather than lifelong identity tracking. Given the complexity of Re-ID, this thesis does not attempt to address all challenges simultaneously; instead, it isolates and studies representative operating regimes to clarify the trade-offs between efficiency, accuracy, and cross-domain robustness.

1.3 Significance and Contributions

Problem statement. Given a set of video tracklets produced by trackers operating on a network of non-overlapping cameras, the central problem addressed in this thesis is to re-associate tracklets that belong to the same identity over time and across cameras. This re-association is essential in practical systems where tracking fragments occur due to occlusions, targets leaving and re-entering the field of view, or changes in viewpoint across cameras. While video tracklets provide richer temporal cues than single images, they also raise computational demands and expose additional generalization issues. In addition, this thesis considers Re-ID under a temporally bounded gallery setting, where new identities may enter the active surveillance window, but the system is not designed to maintain an unbounded identity gallery over arbitrarily long durations.

Research gap. As outlined in Section 1.2, the thesis is motivated by the mismatch between benchmark-oriented video Re-ID research and practical deployment requirements. In particular, the work focuses on how video-based Re-ID models can be designed under different levels of supervision, computational capacity, and cross-camera or cross-domain variation.

Contributions. This thesis makes several contributions toward bridging these gaps:

- **Resource-efficient one-shot video Re-ID (Chapter 4).** An unsupervised video-based Re-ID approach that leverages progressive labelling and learning to incorporate unlabeled data iteratively is proposed in this thesis. The design emphasizes practical deployment by supporting lightweight backbone options and assuming manual labelling is expensive. Also, a practical labelling mechanism consistent with entry-point monitoring is established, allowing a one-shot template per identity to

be acquired by assigning a new label when a person crosses a designated entry camera.

- **Fully supervised transformer-based video Re-ID (Chapter 5).** In this thesis, a supervised transformer-based model is presented to represent tracklets using spatiotemporal features derived from both global sequence-level cues and local body-region cues. Also, this chapter integrates a comprehensive matching pipeline, including re-ranking strategies drawn from literature, to strengthen retrieval performance when labelled data and computational resources are available.

- **Semi-supervised transfer for challenging domain and viewpoint shifts (Chapter 6).** This thesis introduces a semi-supervised architecture that combines transformers with adversarial learning to better handle intra- and inter-domain shifts. This setting is driven by newer, more challenging benchmarks, including large-scale, cross-platform evaluation protocols, where robustness to severe viewpoint change is a primary requirement.

- **Unified analysis across operating regimes.** Across Chapters 4–6, a consistent experimental and analytical perspective is provided, clarifying the manner in which design choices trade accuracy for efficiency and how adaptation mechanisms become necessary as viewpoint and domain shifts become more extreme.

Publication. Parts of this thesis have resulted in one peer-reviewed publication:

- H. Alghamdi, W. El Ahmar, and R. Laganière, "Deep Transductive Learning for Person Re-Identification," in *Image Analysis and Processing – ICIAP 2025*, Springer Nature Switzerland, 2026, pp. 512–524 [10].

1.4 Thesis Structure

This thesis is structured into seven chapters. Chapter 1 introduces person Re-ID, advocates a video-based setting, and outlines the problem statement, objectives, scope, and contributions. Chapter 2 reviews the relevant literature, starting from deep learning in computer vision and narrowing to image-based and video-based person Re-ID, with emphasis on the limitations that support this thesis. Chapter 3 presents the experimental methodology, including datasets, evaluation protocols, and metrics used throughout the thesis.

Chapter 4 presents the resource-efficient one-shot video Re-ID model designed for deployments where labelled data is unavailable and computational resources are constrained. Chapter 5 presents the fully supervised transformer-based model and focuses on maximizing accuracy under the assumption that labelled training data is available. Chapter 6 presents the semi-supervised transfer learning model designed to improve robustness under strong domain and viewpoint shifts, including newer, more challenging evaluation settings. Finally, Chapter 7 concludes the thesis by summarizing the main findings and outlining limitations and future research recommendations.

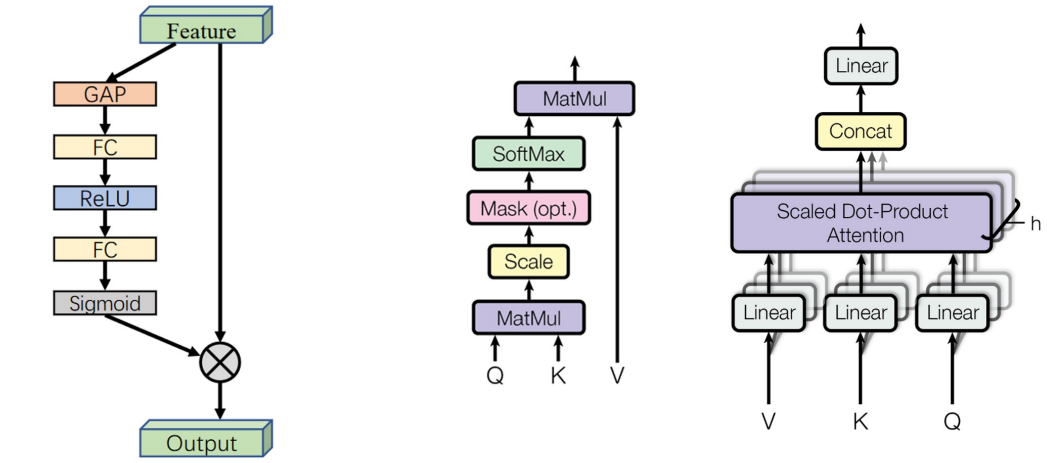
Chapter 2

Literature Review

This chapter reviews the literature most relevant to video-based person re-identification (Re-ID) and to the three learning regimes studied in this thesis. It begins with a concise discussion of background concepts that directly support later chapters, including efficient convolutional backbones, attention and transformer-based sequence modelling, and transfer-oriented learning under domain shift. The chapter then reviews person Re-ID more broadly, including feature extraction, metric learning, and the main challenges that affect identification reliability. Finally, it focuses on video-based person Re-ID, with particular emphasis on recent unsupervised, supervised, and transfer-based methods that motivate the methodological choices developed in Chapters 4, 5, and 6.

2.1 Background Concepts Relevant to Video-Based Person Re-ID

This section provides only the background concepts that are directly relevant to the methods developed later in this thesis. Rather than offering a general tutorial on deep learning, it focuses on the architectural and learning ideas that recur in modern video-based person re-identification (Re-ID), namely



(a) Squeeze and excitation (b) Self-Attention (c) Multi-Head Attention

FIGURE 2.1: Squeeze-and-excitation, self-attention, and multi-head self-attention blocks [12], [13].

convolutional backbones for efficient feature extraction, attention and transformer mechanisms for sequence modelling, and transfer-based learning under domain shift.

2.1.1 Backbone Architectures for Re-ID

Early deep person Re-ID methods were predominantly built on convolutional neural networks (CNNs), which remain attractive when computational efficiency is important. In particular, residual CNN backbones, such as ResNet, provide strong visual representations while remaining practical for lightweight or resource-aware deployments. This is especially relevant in video Re-ID settings, where multiple frames must be processed for each tracklet, where the overall cost of feature extraction becomes significant [11].

Recent work increasingly relies more on attention-based models (Figure 2.1). Attention mechanisms allow the network to emphasize informative regions or frames while suppressing irrelevant content such as background clutter, occlusion, or uninformative temporal segments [14], [15]. This is

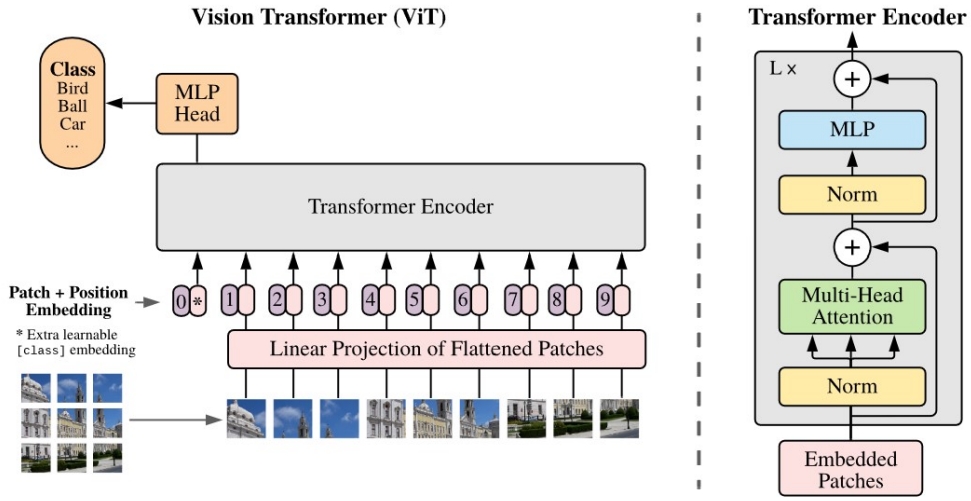


FIGURE 2.2: Vision Transformer overview and encoder structure [16].

particularly useful in Re-ID because identity evidence is often unevenly distributed across a person's body and across frames within a tracklet.

Transformers extend this idea by modelling long-range interactions through self-attention. After the introduction of the Vision Transformer (ViT) (Figure 2.2), transformer-based models became increasingly important in visual recognition because they represented images as sequences of tokens and captured global context with more flexibility than conventional CNNs [13], [16]. In video Re-ID, this makes transformer backbones especially suitable for integrating spatial and temporal cues across multiple observations of the same person. As a result, CNNs remain relevant in resource-constrained regimes, while transformer-based backbones become increasingly important in high-accuracy and domain-robust video Re-ID settings.

2.1.2 Temporal Modelling in Video Re-ID

A key distinction between image-based and video-based Re-ID is that video tracklets provide multiple observations of the same identity over time. This allows the model to exploit temporal continuity, pose variation, and viewpoint

Model	mAP	CMC-1	CMC-5	CMC-10	CMC-20
Image-based	74.1	81.3	92.6	94.8	96.7
Pooling	75.8	83.1	92.8	95.3	96.8
Attention	**76.7**	**83.3**	**93.8**	**96.0**	**97.4**
LSTM	72.0	80.4	92.7	94.9	96.9
GRU	70.5	79.7	91.5	93.8	95.3

TABLE 2.1: Video Re-ID sequence modelling comparison (after Gao et al. [19]).

changes across frames, rather than relying on a single still image. The challenge, however, is how to aggregate frame-level evidence into a stable and discriminative tracklet representation.

Early sequence modelling methods often relied on recurrent neural networks (RNNs), particularly LSTM and GRU variants, to encode temporal dependencies [17], [18]. While these models were useful for sequential reasoning, later studies in video Re-ID showed that simple pooling and, more importantly, attention-based aggregation, often outperformed recurrent alternatives because they focused more effectively on informative frames and were easier to optimize at scale. Table 2.1 represents an example in which attention-based aggregation produces stronger retrieval performance than LSTM and GRU sequence modelling [19].

This transition from recurrent modelling to attention-based aggregation is important for this thesis. It explains why later video Re-ID methods increasingly rely on spatiotemporal attention and transformer-based representations rather than classical sequence encoders.

2.1.3 Learning Regimes and Domain Adaptation

Video Re-ID has been studied under multiple learning regimes with varying amounts of supervision. In fully supervised settings, identity labels are available for all training tracklets, which allows the model to learn strong

identity-discriminative representations. In weakly supervised or one-shot settings, only limited identity information is available, forcing the acquisition of training signals through pseudo-labelling, iterative refinement, or conservative expansion of the labelled set. In transfer-based settings, labelled data may exist only in a source domain, while the target deployment domain remains unlabelled and differs in viewpoint, illumination, background, and camera characteristics.

This distinction is central to the structure of this thesis. Later chapters study three complementary deployment regimes: resource-aware learning with minimal supervision, supervised video Re-ID with stronger spatiotemporal modelling, and transfer-oriented learning under domain and viewpoint shifts. For this reason, only the learning paradigms that directly support these regimes are summarized here.

A common mechanism in label-scarce settings is self-training, in which a model trained on a small labelled set generates pseudo-labels for unlabelled samples and is retrained iteratively [20]. In Re-ID, such strategies are especially relevant because manual annotation of video tracklets is expensive and large camera networks naturally produce a large volume of unlabeled data.

Transfer learning, often referred to as "domain adaptation" in the Re-ID literature, addresses the case where a labelled source domain is available but the target deployment domain has no identity annotations. Figure 2.3 provides a schematic taxonomy of deep transfer learning adapted from Tan et al. [21]. The figure can be interpreted as four common mechanisms for knowledge transfer. In **instance-based transfer**, source samples are selectively reused or reweighted according to their relevance to the target domain. In **mapping-based transfer**, the source and target data are projected into a shared latent space to reduce the distribution gap between them. In **network-based transfer**, part of a source-trained model is reused and fine-tuned for the target task. In **adversarial-based transfer**, a feature extractor is trained to produce

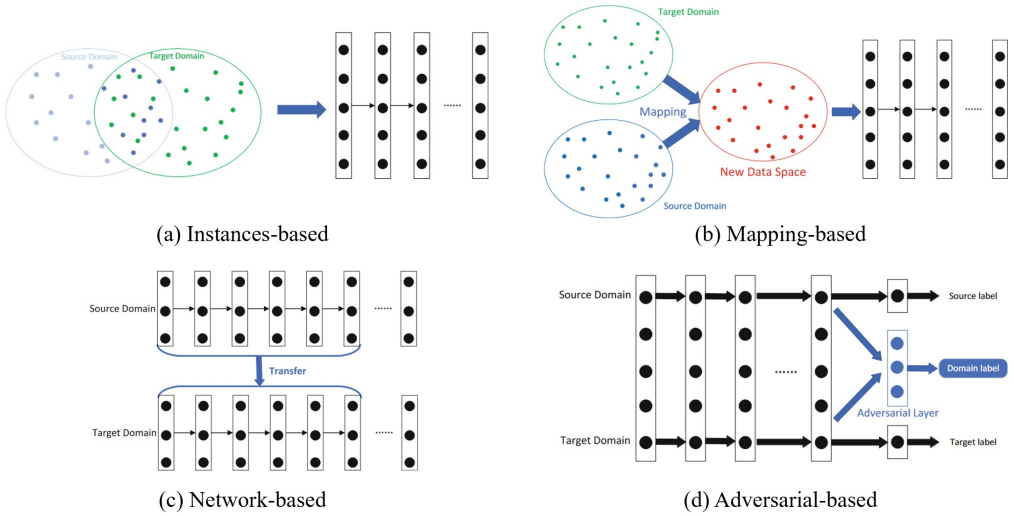


FIGURE 2.3: Schematic overview of four deep transfer learning mechanisms: instance-based sample reweighting, mapping-based shared-space alignment, network-based parameter transfer, and adversarial domain alignment [21].

representations that are useful for the main task while remaining difficult for a domain discriminator to separate [22], [23].

In adversarial transfer learning, typical strategies include gradient-reversal methods that encourage domain confusion, minimax formulations that alternate between feature learning and discriminator updates, and generative approaches that reduce appearance discrepancies by synthesizing source- or target-like samples [24].

Overall, these background concepts motivate the later literature review and the methodological choices developed in this thesis: efficient CNN-based representations remain relevant when resources are constrained, attention and transformers provide stronger sequence modelling for video Re-ID, and transfer-based learning becomes necessary when deployment conditions differ substantially from the source training domain.

2.2 Person Re-ID

In an image retrieval task, person Re-ID retrieves images from a dataset (the gallery) that match a query image (the probe). To achieve this, the system must learn discriminative features to distinguish individuals. Early representative techniques relied on handcrafted feature extraction and saliency modelling [25], [26], [27], [28]. Similarity-measure-based methods were also common, including large margin nearest neighbour with rejection, mirror representation, and constrained asymmetric multitask discriminant component analysis [29], [30], [31].

After the introduction of deep learning to Re-ID around 2014, Yi et al. and Li et al. were among the first to adopt deep models for the task [32], [33]. While Li et al. proposed a filter pairing neural network, Yi et al. used a CNN-based framework, which became a common backbone choice for subsequent methods.

At the system level, person Re-ID is used during inference after people have been detected in camera streams. As shown in Figure 2.4, the pipeline begins with one or more camera streams, followed by a person detection stage that extracts bounding boxes around visible pedestrians. These bounding boxes are not ideal or uniform person images: they may have different resolutions, include background pixels, and capture different visible body regions depending on viewpoint, occlusion, detector quality, and camera distance. In some cases, a bounding box may contain the full body, while in others it may contain only the torso or another partially visible region.

For video-based Re-ID, detections belonging to the same person within a camera are then associated over time to construct tracklets. A trained Re-ID model then extracts a descriptor from each tracklet, and these descriptors are compared against gallery descriptors to rank possible identity matches. The final stage associates the query tracklet with the best-matching identity or

treats it as a new identity when no reliable match is found.

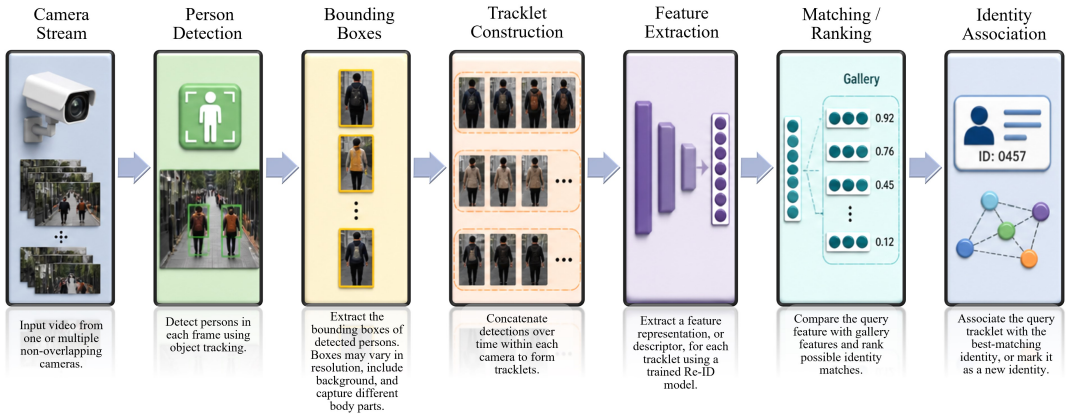


FIGURE 2.4: Person Re-ID inference pipeline. The tracker first extracts pedestrian bounding boxes from camera frames. These crops may vary in resolution, include background pixels, and contain different visible body regions. Detections are then concatenated over time to form tracklets. A trained Re-ID model extracts descriptors from the resulting tracklets, and matching is performed against the gallery to associate identities.

This imperfect crop-generation process is one reason person Re-ID remains challenging in practical systems: the model must compare detections that may differ not only in identity-relevant appearance, but also in crop quality, scale, background content, and visible body extent.

2.2.1 Feature Extraction

This subsection discusses four feature extraction strategies. The first uses global features, which are straightforward but often less discriminative. The second uses local representations, which can be more fine-grained but also sensitive to occlusions and background clutter. The third uses attention-driven features, which suppress background and highlight discriminative regions, but may fail if attention is mislocalized. Finally, pose estimation introduces explicit body-part alignment cues but depends on the quality of pose detection.

Global Features

By extracting a feature vector for the whole image, we obtain a global representation, as illustrated in Figure 2.5(A). Gray and Tao presented an early method for viewpoint-invariant pedestrian recognition using AdaBoost [34]. Farenzena et al. proposed an appearance-based method using chromatic content, spatial arrangement, and recurrent motifs, weighted by symmetry and perceptual asymmetry [35]. Liao et al. proposed local maximal occurrence (LOMO) to extract handcrafted horizontal local features to tackle viewpoint changes [36].

In their earlier work, Zhao et al. proposed salience and patch matching and penalized inconsistent salience [26]. In later work, Zhao et al. proposed a view-invariant approach using discriminative learned mid-level filters [37].

While handcrafted methods dominated early global feature learning, CNN-based Re-ID methods are now prevalent. Wang et al. proposed a framework fusing single-image representation (SIR) and cross-image representation (CIR) [38]. Zheng et al. introduced the ID-discriminative embedding (IDE) descriptor by treating each identity as a class in a multi-class classification setup [39]. Qian et al. emphasized the value of multi-scale cues for discriminative representation learning in video Re-ID [40].

Local Features

Instead of extracting one global feature vector, an image can be divided into regions to obtain part-based descriptors. Sun et al. proposed PCB by partitioning feature maps into uniform horizontal stripes and using refined part pooling (RPP) to improve part consistency [41]. Song et al. extended this direction with domain-invariant mapping (DIMN) [42]. Sun and Zheng incorporated pedestrian rotation angle information to address viewpoint effects [43]. Varior et al. proposed a context-aware Siamese network with

LSTM-based temporal modelling, minimizing distances for positive pairs and maximizing distances for negative pairs [44].

Attention-driven Features

Global and local features may still treat background and foreground similarly. However, background often provides little value for matching and may introduce noise. Attention mechanisms aim to focus on identity-relevant regions while suppressing background, as illustrated in Figure 2.5(C) [45].

Li et al. proposed harmonious attention by combining pixel-level and channel-wise attentions [46]. Wang et al. introduced a fully attentional block generating both channel-wise and spatial-wise attention and combined triplet, focal, and attention losses [47]. Song et al. used background suppression in mask-guided contrastive attention, while Chen et al. used reinforcement learning signals to assess attention quality [5], [48].

Attention can also be leveraged across multiple images during training. Si et al. proposed DuATM using inter- and intra-sequence attention for feature alignment and sequence comparison [49]. Zheng et al. proposed CASN to enforce attention consistency across similar images [50]. Zhou et al. proposed a consistent attention regularizer to produce inter-class foreground masks [51]. Chen et al. used CRF-based group similarity, and Luo et al. used spectral/graph clustering to exploit inter-image relationships [52], [53].

Pose Estimation

Body part misalignment motivates pose-guided representations. Suh et al. fused appearance and part feature maps to align corresponding parts [54]. Zhao et al. detected discriminative body parts and aggregated part-level similarities [55].

Other approaches incorporate multi-channel or multi-scale cues. Cheng et al. learned global and local features jointly, and Li et al. proposed MSCAN

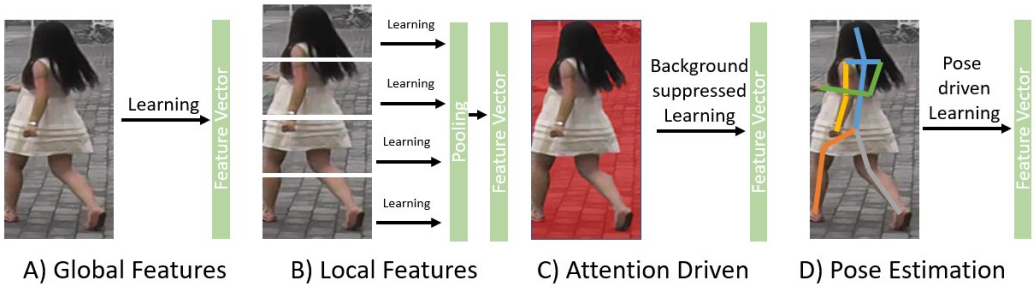


FIGURE 2.5: Person Re-ID feature extraction strategies.

using multiple receptive fields via different kernel sizes [56], [57]. Zhao et al. proposed SpindleNet to decompose and fuse features across stages [58].

Pose estimation can also reduce sensitivity to background clutter. Su et al. proposed PDC to learn representations from the whole image and local parts [59]. Xu et al. proposed AACN using pose-guided part attention and attention-aware feature composition [60]. Zhang et al. introduced DSAG-Stream to guide learning toward semantically aligned features, while Guo et al. learned dual part-aligned representations for human parts and non-human objects [61], [62].

2.2.2 Metric Learning

This subsection discusses two categories of metric learning in Re-ID. The first includes classical metric learning methods, which were common before deep embeddings became dominant. The second includes deep embedding-based metric learning objectives, which are now standard in modern Re-ID pipelines.

Classic Metric Learning

- **Mahalanobis Distance:** A generalized multi-dimensional form of Euclidean distance. The distance between two vectors X and Y uses the covariance matrix S, as shown in Equation 2.1:

$$D(X,Y) = (X-Y)^T \mathbf{S} (X-Y) \tag{2.1}$$

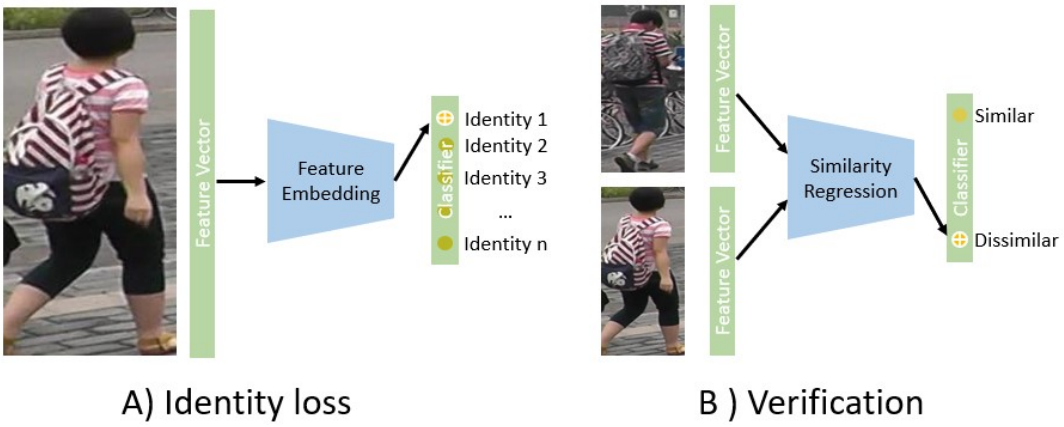


FIGURE 2.6: Identification loss versus verification loss.

Köstinger et al. and Zheng et al. used Mahalanobis-distance formulations (Equation 2.1) as a basis for metric learning [63], [64]. Köstinger et al. proposed KISSMe, which derives the metric from a log-likelihood test and remains a widely used baseline [63].

- **Support Vector Machine (SVM):** SVM-based learning is widely used for classification and ranking. Liu et al. used weak RankSVMs combined into a stronger ranker, and Prosser et al. used ensemble learning to optimize weights [65], [66]. Zhang et al. proposed LSSCDL to learn identity-specific mappings using SVMs [67].

- **Boosting:** Boosting combines weak classifiers into a stronger one and can learn nonlinear boundaries. Gray and Tao implemented AdaBoost to combine weak similarities into a robust similarity function [34].

Deep Metric Learning

Deep Re-ID pipelines commonly combine identification-style classification objectives with verification-style similarity objectives. Figure 2.6 contrasts these two learning signals.

- **Identification Loss:** When each identity is treated as a class, Re-ID can be optimized as a classification problem, as described by Zheng et al. [68].

Identification loss ℓ_{id} is often implemented as a cross-entropy objective, as shown in Equation 2.2:

$$\ell_{id}\left(x, C\right) = -\sum_{i} x_i \; log(\; C_i) \tag{2.2}$$

Representative examples of identification-loss-based training in Re-ID include prior work on adversarial learning, camera-aware learning, and embedding refinement [3], [50], [53], [69], [70], [71], [72], [73], [74], [75]. Wojke and Bewley used a cosine softmax variant, and Fan et al. proposed SphereReID with sphereLoss [76], [77]. Luo et al. used an additive margin softmax variant in a Re-ID setting [53]. Zheng et al. and Luo et al. adopted label smoothing, and Müller et al. discuss its effect on calibration and generalization [73], [78], [79].

- **Verification Loss:** Verification objectives treat Re-ID as a retrieval problem by pulling similar samples closer while pushing dissimilar samples apart, as discussed by Lecun et al. [80]. A contrastive-style verification loss can be written as Equation 2.3:

$$\ell_V\left(x, y\right) = \left(1 - \mu\right) D_{x,y}^2 + \mu \left(max\left\{0, \alpha - D_{x,y}\right\}\right)^2 \tag{2.3}$$

This loss can be used for similarity regression or binary verification, as in representative Siamese-style Re-ID formulations [33], [38], [44], [50], [68], [75].

Hermans et al. introduced triplet loss, which extends verification learning by using an anchor x, a positive y, and a negative z, as shown in Equation 2.4 [81]:

$$\ell_{Tri}\left(x, y, z\right) = max\left\{0, \alpha + D_{x,y} - D_{x,z}\right\} \tag{2.4}$$

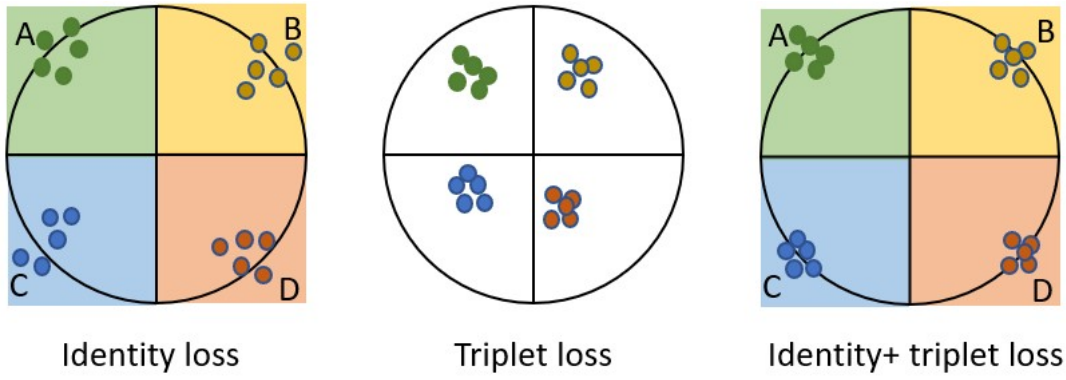


FIGURE 2.7: Embedding space intuition for identity separation.

Shi et al. and Hermans et al. discuss weighting and hard mining strategies to improve discriminability, and triplet loss is widely adopted in Re-ID pipelines [5], [7], [54], [81], [82].

- **Loss Fusion:** Since identification loss increases inter-class separation and verification-style losses reduce intra-class variation, combining both often yields stronger embeddings. Figure 2.7 illustrates this intuition. Representative examples combine identification with Siamese-style verification or with triplet loss. [42], [44], [47], [48], [51], [52], [61], [62], [68], [75].

2.2.3 Challenges

Re-ID challenges fall into two broad categories. The first includes common visual issues such as occlusion, background clutter, scale/resolution changes, illumination variations, and generalization. The second includes challenges that are especially critical for person Re-ID, such as viewpoint variations, body-part misalignment, and pose variations. Figure 2.8 summarizes these challenges visually.

Occlusion occurs when a target is partially obstructed by objects or other pedestrians. In retrieval settings, occlusion can cause extracted features to

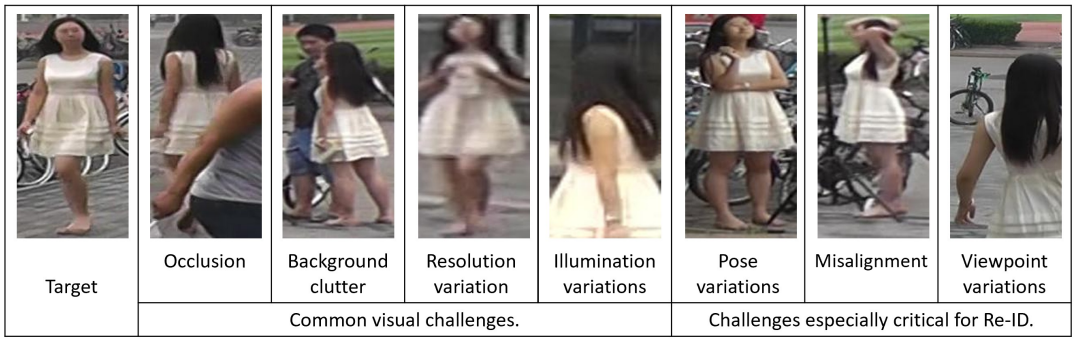


FIGURE 2.8: Examples of common visual challenges and challenges especially critical for person Re-ID, including occlusion, background clutter, resolution variation, illumination variations, pose variations, body-part misalignment, and viewpoint variations.

include irrelevant information, leading to errors if the model fails to focus on visible identity cues. Early approaches improved robustness by using body part-based matching and discarding [41], [83]. Miao et al. used attention, guided by pose landmarks, to focus on shared visible regions, and Hou et al. exploited spatio-temporal cues to recover occluded information using temporal attention and adjacent frames [4], [84].

Background clutter introduces irrelevant visual noise, complicating the separation of a person from the environment. This can be mitigated using part-based matching, foreground awareness, and background suppression [51], [85], [86], [87].

Variations in appearance due to illumination and resolution changes can significantly affect low-level cues such as texture and colour. These issues can be addressed via supervised learning, colour-invariant transforms, or self-supervised/generative augmentation to reduce domain gaps between degraded and clean observations [6], [7], [88], [89].

Body-part misalignment and pose variations further complicate matching. If body parts are not aligned between probe and gallery images, feature correspondence becomes unreliable. Cho et al. proposed pose-aware approaches that classify pose (front/back/side) to reduce ambiguity, while Xu

et al. proposed pose-guided attention models to learn alignment-invariant representations [60], [90].

Finally, the camera viewpoint challenge is a core difficulty in Re-ID: the model must operate across non-overlapping cameras with different angles, scales, and background statistics. While complete viewpoint invariance remains difficult, many modern approaches incorporate mechanisms to reduce viewpoint sensitivity [91], [92], [93], [94].

2.3 Unsupervised Video-Based Person Re-ID

While the field of person Re-ID publishes dozens of research papers every year, the number of studies specific to video Re-ID remains relatively low. According to Wu et al. [72], most early unsupervised person Re-ID works were image-based and relied on dictionary learning, graph matching, or metric learning. Although the field has since evolved toward pseudo-labelling, clustering, and domain-adaptive learning, this progress should not be interpreted as a complete solution. Many unsupervised video Re-ID methods remain highly sensitive to pseudo-label noise, depend on repeated clustering or refinement stages, and are often evaluated mainly on standard benchmarks, such as MARS and Duke. As a result, their reported gains do not always translate into robust performance under stronger viewpoint shifts, larger domain gaps, or resource-constrained deployment.

More broadly, unsupervised video Re-ID methods often face a three-way tension between label reliability, temporal modelling, and computational practicality. Methods that aggressively expand pseudo-labels may improve coverage but also amplify early errors. Methods that introduce more sophisticated refinement or domain adaptation mechanisms can reduce some of this noise, but they often become more complex, less stable, and harder to deploy. These limitations foster the need for approaches that remain effective

under weak supervision while also being more explicit about efficiency and generalization.

Early unsupervised video Re-ID methods were largely built around two broad strategies: association-based learning and clustering-based learning. Association-based approaches progressively expand supervision by linking unlabelled tracklets to reliable anchors or previously estimated matches, as in stepwise learning, EUG, ProLearn, and RACE [72], [95], [96], [97]. Clustering-based approaches partition the embedding space and use the resulting clusters as surrogate identities, as seen in DGM, DBC, SSLR, TCPL, SCL, and subsequent variants [92], [98], [99], [100], [101]. More recent methods continue this direction while focusing more explicitly on hard matches, frame quality, and label reliability, for example, through attention-guided refinement or improved sampling strategies [102], [103].

Despite their historical importance, both techniques have clear limitations. Association-based methods depend heavily on the reliability of early matches; if the initial anchors are weak, identity errors can propagate through later iterations. Clustering-based methods reduce manual supervision more aggressively, but they are sensitive to feature quality, cluster purity, and the assumed number or distinctive identities. In video Re-ID, these weaknesses are further amplified by noisy frames, fragmented tracklets, and viewpoint variations across cameras. Consequently, many early methods are effective mainly when the benchmark structure is relatively favourable but become less reliable when pseudo-label noise and cross-domain shift are more severe.

2.3.1 Addressing Pseudo-Label Noise and Domain Gaps

To offset performance degradation induced by noisy pseudo-labels and domain shifts, recent unsupervised and domain-adaptive Re-ID research has converged on two complementary priorities: enhancing the reliability of

pseudo supervision and narrowing cross-domain differences. Several research studies actively purify or refine clustering-based labels on the pseudo-label side by using teacher-student distillation, multi-view label generation, or outlier-aware filtering. Such guidelines include distillation-driven purification [104], multi-view pseudo-labelling with inter-/intra-cluster gap refinement [105], and loss-based identification of out-of-distribution samples by temporal ensembling and GMM modelling [106]. Other methods further increase robustness by enhancing foreground cues and supplementing informative signals under distracting conditions [107], while multi-granularity collaboration reduces errors on hard instances by producing labels across multiple embedding spaces [108].

In parallel, UDA-oriented strategies focus on bridging domain gaps under increasingly realistic limitations. Source-free adversarial adaptation aims for discriminative generalization without using source data or deals with appearance-style mismatches [109]. Alternatively, online UDA frameworks keep source-guided similarity to reduce loss during streaming adaptation [110]. Efficiency has also been a focus; prototype-based flexible learning can quickly create coarse supervision and reduce dependence on expensive clustering [111]. Taken together, these limitations motivate the resource-aware and domain-robust designs developed later in this thesis. In particular, they highlight the need for methods that remain stable under noisy pseudo-labels while also reducing the impact of cross-domain mismatch. Accordingly, the approach studied later in this thesis emphasizes robustness to pseudo-label noise through more reliable supervision signals and improved cross-domain generalization through reduced feature discrepancy between domains.

Lastly, Liu et al. [112] introduce a weakly supervised tracklet association learning approach using video labels for person re-identification. They address the limitations of existing methods that rely heavily on expensive manual labels or perform poorly in unsupervised settings. Their method

focuses on weakly supervised learning, where tracklets are grouped into bags and identified with video labels. They propose a Cross-Bag Tracklet Association (CTAL) term to explore tracklet associations between bags by mining reliable positive tracklet pairs and hard negative pairs. This approach aims to balance the need for labelled data with recognition performance, showing effectiveness on weakly labelled MARS and Duke datasets [112].

2.3.2 Transfer Learning and Domain Adaptation in Person Re-ID

Transfer learning, often referred to as "domain adaptation" in the Re-ID literature, addresses the setting where a labelled source dataset is available, but the target deployment domain has no identity annotations. This setting is particularly relevant in surveillance applications because collecting fully labelled target-domain Re-ID data is expensive, while appearance variations across cameras, sites, and capture conditions often prevent direct generalization from one dataset to another.

A common strategy is to first train a model on a labelled source dataset and then adapt the learned representation to an unlabelled target dataset. Examples of this approach include camera-invariant and mutual-learning-based adaptation strategies, as well as methods that transfer similarity structures or pseudo-label information across domains [74], [113], [114], [115]. These methods demonstrate that transfer learning can reduce labelling requirements substantially; however, their performance often degrades when the source and target datasets differ strongly in viewpoint, scene layout, illumination, or camera characteristics. This limitation is especially important in person Re-ID, where domain gaps are often severe and generalization remains difficult. As noted in prior work, transfer-based methods can perform well when the

source and target domains are relatively similar, but their robustness declines as the domain shift increases [116].

Compared with image-based Re-ID, transfer learning for video-based person Re-ID remains relatively underexplored. Zahra et al. [117] report that only a small portion of surveyed Re-ID projects explicitly address model generalization or domain adaptation. Early studies were motivated in part by the limited scale and diversity of existing datasets. After the introduction of MARS [118], Xu et al. [119] explored domain adaptation in a reduced camera setting using limited temporal information, while Meng et al. [120] extended this direction by exploiting full tracklets from the source camera. These works demonstrated the feasibility of transfer learning for video Re-ID, but they remained relatively constrained in their adaptation settings.

One-shot or weakly supervised adaptation is a related approach. EUG [72] was an influential early example of using minimal supervision to introduce pseudo-labels into unlabelled data. Subsequent works extended this idea through few-shot adaptation, iterative refinement, and cluster-based pseudo-labelling [91], [103], [121], [122]. However, in many of these settings, the source and target data are still drawn from the same dataset and differ mainly by camera view or split, rather than by a true cross-dataset domain shift.

Only more recently have studies addressed transfer between genuinely different video Re-ID datasets. Mekhazni et al. proposed CAWCL, which combines a feature bank with a camera-discriminator network to reduce cross-domain discrepancy [123]. Zhang et al. proposed DCCAL, which incorporates segmentation-attentive learning to improve alignment between source and target datasets [124]. These works represent important progress toward realistic video-based domain adaptation, but they remain relatively few compared with the broader supervised literature. An important remaining gap is that transformer-based backbones, despite their strong performance in supervised video Re-ID, are still underexplored in domain-adaptive video

Re-ID. Recent supervised studies have shown the benefit of transformer-based spatiotemporal modelling [125], [126], yet most transfer learning methods for video Re-ID still rely on CNN-dominant or non-transformer adaptation pipelines. This gap inspires the later parts of this thesis, particularly the chapter on transductive video Re-ID, where domain-invariant representation learning becomes a primary objective rather than a secondary refinement step.

2.3.3 Challenges and Emerging Trends

The task-specific nature of unsupervised Re-ID remains problematic and requires tailored solutions. For example, Dong et al. [127] solve occluded-person Re-ID by integrating multi-view information with a propagation technique to manage missing or occluded body parts. Khaldi et al. [128] target the unique problems of UAV-collected images—extreme viewpoints, low resolution, and limited annotations—by proposing a three-stage unsupervised pipeline (generative $\rightarrow$ contrastive $\rightarrow$ clustering) that learns view-invariant features without labels and outperforms prior methods on drone datasets.

Recent design trends stress the importance of both robustness and practicality. Nguyen et al. [129] propose Contrastive Viewpoint-aware Shape Learning (CVSL), which uses 2D pose-derived shape cues and contrastive losses to build texture-invariant, viewpoint-robust shape representations for long-term Re-ID. Recent surveys and reviews show the field is broadening: Ma et al. [130] analyze video Re-ID methods across data, algorithms, compute, and real-world application challenges, highlighting issues such as scarce datasets, constrained feature learning, and the need to scale to open-world scenarios. In addition to this, more recent works [131], [132] examine multilinear subspace learning to better preserve high-order structure in human representations, hinting toward richer, more expressive feature spaces.

To address specific challenges and emerging trends, several novel approaches have been proposed. Wang et al. [133] propose a Relation-Preserving Feature Embedding (RPE) model that leverages structural relations among samples to enhance unsupervised person Re-ID performance without requiring annotations. Their approach integrates sample content and neighbourhood structure relations by combining autoencoders and graph autoencoders. A key component is the relation and content information fusion (RCIF) module, which dynamically merges information from both perspectives. To address the lack of identity labels, they employ an adaptive optimization strategy to update affinity relations among samples iteratively, rather than reconstructing the entire affinity matrix. Rigorous experiments on widely used benchmarks demonstrate the superiority of RPE over state-of-the-art unsupervised methods [133].

Zhang et al. [134] introduce a Dual Representation Modelling and Progressive Contrastive Learning (DRMPCL) method for unsupervised video person re-identification (USL-VReID). This method addresses the diversity of frames within tracklets and the uncertainty of tracklet representation quality. It introduces a Dual Representation Modelling (DRM) module to obtain easy and hard frames for reliable tracklet representations, a Discrepant Prototype Contrastive Learning (DPCL) module to preserve intrinsic variety in tracklet features, and a Multi-level Progressive Learning (MPL) strategy to promote feature embedding from easy to hard at the frame level and from commonality to diversity at the tracklet level. Their extensive experiments demonstrate superior performance against state-of-the-art methods on Duke and MARS datasets [134].

Overall, progress in pseudo-label purification, domain adaptation, and domain-specific design is progressively narrowing the gap between unsupervised and supervised Re-ID. Such advances make unsupervised methods more practical in the real world, especially in niche scenarios like aerial Re-ID,

Method	Year	MARS				Duke				cite
		mAP	Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20	
OIM	2017	13.50	33.70	48.10	-	43.80	51.10	70.50	-	[135]
Stepwise	2017	19.65	41.21	55.55	66.76	46.76	56.26	70.37	79.20	[95]
RACE	2018	24.50	43.20	57.10	67.60	-	-	-	-	[97]
DAL	2018	23.00	49.30	65.90	77.90	-	-	-	-	[136]
EUG	2018	42.45	62.67	74.94	82.57	63.23	72.79	84.18	91.45	[72]
PGOR	2018	46.51	66.50	76.38	84.69	68.42	72.47	86.38	94.45	[52]
PL	2019	42.60	62.80	75.20	83.80	63.30	72.90	84.30	91.40	[96]
DGM	2019	16.87	36.81	54.01	68.51	33.62	42.36	57.92	69.31	[98]
DBC	2019	43.80	64.30	79.20	87.80	66.10	75.20	87.00	-	[99]
TCPL	2020	43.60	65.20	77.50	-	67.90	76.80	87.80	-	[100]
SSLR	2020	43.60	62.80	77.20	-	69.30	76.40	88.70	-	[92]
SCL	2022	46.60	66.60	77.00	-	78.40	82.20	93.20	-	[101]
SRC	2022	40.50	62.70	-	-	76.50	83.00	-	-	[103]
TASTC	2023	47.20	65.60	-	-	68.20	76.80	-	-	[137]
MPC	2023	71.40	81.60	-	-	87.30	89.30	-	-	[108]
CAWCL	2023	44.80	62.20	-	-	-	-	-	-	[123]
RPE	2024	40.40	63.30	75.40	80.60	71.50	77.80	89.30	91.70	[133]
Weakly	2024	75.2	83.9	93.6	95.5	87.4	89.7	97.2	98.0	[112]
DCCAL	2024	74.50	84.00	-	-	91.40	93.00	-	-	[124]
DRMPCL	2025	71.6	82.0	91.1	93.4	92.7	94.0	98.6	99.4	[134]

TABLE 2.2: Unsupervised video Re-ID methods on MARS and Duke (2017–2025).

where there isn't enough labelled data yet [128], [130].

Table 2.2 show the performance trend of the two most frequently reported video Re-ID benchmarks, but it also reveals an important limitation of the literature itself. First, most methods are still compared mainly on MARS and Duke, which means the field has a much clearer ranking on these datasets than on newer and more challenging benchmarks. Second, several entries report incomplete metrics or omit one of the datasets entirely, which makes direct comparison less reliable. Third, strong improvements on MARS do not necessarily imply robustness under larger viewpoint shifts, newer camera settings, or cross-platform scenarios. For this reason, the table should be read as evidence of progress on standard benchmarks, not as proof that the underlying challenges of generalization, scalability, and deployment have been solved.

2.4 Supervised Video-Based Person Re-ID

Recent advances in deep learning have led to strong progress in supervised person Re-ID, and video-based approaches have benefited in particular from the availability of multiple observations of the same identity across time. However, this literature also has important limitations. Many supervised video Re-ID methods assume abundant identity labels, stable benchmark conditions, and sufficient computational resources for multi-branch feature extraction, temporal aggregation, or transformer-based modelling. As a result, strong benchmark performance does not necessarily imply suitability for practical deployment, especially when annotation is scarce, viewpoint shifts are severe, or compute budgets are limited.

In broad terms, supervised video Re-ID methods can be viewed through two phases. The first phase, roughly before 2023, is dominated by CNN-based pipelines that improved performance through stronger training recipes, attention modules, and temporal aggregation. These methods were effective on standard benchmarks, but many of their gains were incremental and depended on increasingly specialized architectural components. The second phase, after 2023, is characterized by transformers and foundation-style backbones, which improved global spatiotemporal modelling but also increased model complexity, training cost, and dependence on large-scale pretraining. This progression improved accuracy, yet it also widened the gap between benchmark-oriented performance and resource-aware deployment.

With attention mechanisms gaining popularity, many CNN-based video Re-ID works incorporate explicit spatial-temporal saliency modelling. COSAM [138] proposed a co-segmentation-inspired attention module that leverages both spatial and temporal information to compute regional attention, emphasizing identity-consistent regions across frames and capturing individuals along with their accessories. BiCnet-TKS [139] further

adopted a dual-attention design, where one branch captures local attention and the other captures global attention, and introduced a temporal kernel selection (TKS) block to model temporal relations within the tracklet. A broader discussion of the evolution and limitations of deep Re-ID pipelines in this period is also presented in the survey and outlook by Ye et al. [140]. Table 2.3 summarizes representative supervised video Re-ID methods and helps illustrate both the progress and the limitations of the literature. In particular, it shows that benchmark gains are often accompanied by increasing architectural complexity, while evaluation remains concentrated on a small number of standard datasets.

2.4.1 Emerging trends in supervised video Re-ID

The post-2023 period is increasingly defined by two interacting changes. The first is the architectural shift toward transformers [16], [141] and foundation priors [142], [143], [144], motivated by global token interactions and content-adaptive aggregation. The second is the shift in problem settings, where cross-platform deployment and extreme viewpoint and scale changes expose limitations of near-ground-level optimized pipelines. These developments are coherent with the structure of this thesis: Chapter 5 focuses on transformer-based spatiotemporal representation learning and long-range temporal reasoning, while Chapter 6 focuses on generalization when viewpoint shifts and when domain-invariant learning becomes a requirement.

Following the introduction of transformers and vision transformers, several studies demonstrated that transformer backbones are appropriate for retrieval tasks because they can learn transferable representations and model long-range interactions through attention [16], [141], [145], [146]. These developments motivated video Re-ID designs that treat spatiotemporal encoding and temporal evidence fusion as primary components rather than add-on

modules. Two representative transformer works that tackle video Re-ID challenges are Zang et al. [147] and Alsehaim and Breckon [125]. Zang et al. propose a multidirectional, multi-scale pyramid transformer (PiT) that employs vertical and horizontal patch divisions and pyramid representations to preserve fine-grained cues while retaining global context, which helps when discriminative evidence appears at different spatial granularities across frames. Alsehaim and Breckon emphasize that temporal robustness is affected not only by the backbone but also by clip organization and part representation. Their temporal clip shift and shuffle (TCSS) mechanism disrupts clip ordering during training to reduce over-reliance on a single temporal configuration, and their video patch part feature (VPPF) module promotes part-aware token features that remain stable under pose changes and occlusions. More recently (after Alsehaim 2023), Ma et al. have proposed T²MEA [126], a dual-branch transformer that explicitly breaks down temporal motion (dynamic) features and spatial appearance (static) features before integrating them. A content branch uses a Spatial–Temporal Aggregation (STA) module to capture coarse-grained global spatio-temporal structure, while a fovea branch extracts fine-grained cues using a zero-parameter token channel shift interaction (TCSI) module and a Spatial Patches Shift Enhancing (SPSE) module to capture temporal dynamics and enhance appearance under occlusion/illumination changes. The branches are fused by a Cross-Attention Aggregation (CAA) module to create robust, enriched video representations.

Similarly, Trigeminal Transformers (TMT) explicitly exploit the multi-view nature of videos by projecting tracklets into spatial, temporal, and spatial-temporal views via view-wise projectors and then applying self-view and cross-view transformers to both enhance and aggregate view-specific cues, yielding more comprehensive representations for video Re-ID [148]. TMT complements approaches such as TCSS/VPPF and T²MEA by systematically capturing diverse observations (three views) and by explicitly modelling

cross-view interactions, which can be particularly helpful when discriminative evidence is distributed across different view modalities.

Complementing the transformer-centric lines, Long Short-Term Representation Learning (LSTRL) advocates explicit decomposition of long-term multi-granularity appearance and short-term motion representations via a Multi-granularity Appearance Extractor (MAE) and a Bi-directional Motion Estimator (BME). LSTRL shows that combining robust, multi-scale appearance aggregation across frames with efficient, reciprocal short-term motion estimation improves resistance to noisy frames, spatial misalignment, occlusions, and scale variations; moreover, MAE and BME are plug-and-play and can be integrated into existing architectures to provide multi-stage long–short temporal features [149]. LSTRL thus offers a complementary inductive bias to transformer methods: while transformers focus on global token interactions and cross-view fusion, LSTRL emphasizes structured long/short decomposition and efficient motion estimation that can be adopted alongside transformer backbones or hybrid CNN–transformer systems.

More recent studies increasingly combine convolutional inductive biases with transformer global reasoning to improve stability and efficiency on tracklets. DCCT [150] deeply couples convolutional and transformer branches and introduces complementary spatial attention and hierarchical temporal aggregation, supported by self-distillation to transfer spatiotemporal knowledge into compact representations. TCViT [151] focuses more directly on temporal dependency modelling by emphasizing temporal correlation learning within a transformer framework. Collectively, these methods point to a consistent design space that includes tokenization granularity, part-aware modelling, clip/frame mixing strategies, and temporal attention span. This direction aligns with Chapter 5, where temporal modelling decisions and spatiotemporal feature fusion are treated as central design objectives. At the same time, the growing complexity of these models raises an important

question for the literature review: whether the reported gains reflect stronger video Re-ID reasoning itself or simply the increasing scale of the architecture and training recipe.

Although these transformer-based and hybrid approaches improve spatiotemporal reasoning, they also introduce new limitations. Many rely on increasingly elaborate tokenization schemes, multi-branch fusion, auxiliary modules, or large pretrained backbones, which makes it difficult to separate genuine modelling gains from gains due to scale and complexity. In addition, several methods are still validated primarily on standard benchmarks, especially MARS, with less consistent reporting on newer datasets or on settings where viewpoint and domain shift are more severe. Thus, the main limitation of this literature is no longer the absence of strong models, but rather the limited clarity about which improvements are architecturally essential, which are benchmark-specific, and which remain practical under realistic deployment constraints.

In parallel, supervised video Re-ID has increasingly shifted toward more challenging, realistic deployment settings where domain gaps are severe. A prominent example is aerial-ground (UAV-to-CCTV/wearable) video Re-ID, where the domain gap includes extreme viewpoint differences (from near top-down to ground), drastic scale and resolution variations, and temporal discontinuities. The AG-VPReID benchmark [152] and the associated AG-VPReID challenge [153] emphasize these issues at scale, and the reported top solutions commonly rely on multi-stream processing and transformer-based temporal reasoning to aggregate sparse identity evidence over long tracklets while reducing cross-view discrepancy. This line of work shifts the emphasis from marginal within-domain improvements to robust temporal reasoning under cross-platform constraints, which motivates the generalization focus adopted later in this thesis.

Another notable trend is the shift from training from scratch to leveraging

large-scale pretraining and foundation priors, which makes the downstream contribution increasingly about adaptation and temporal aggregation. Fu et al. [154] propose large-scale pretraining for person Re-ID under noisy labels, where the training strategy combines supervised-style objectives with prototype/contrastive components to improve representation quality while reducing the impact of label noise. In addition, recent work adapts vision-language representations to video Re-ID without explicit text prompts. For example, [155] learns a text-free CLIP adaptation for video-based person Re-ID and shows that CLIP-derived priors can be transferred to tracklet-level matching when coupled with video-appropriate temporal fusion. These foundation-driven approaches typically improve invariance to appearance changes (illumination, background, style), but they also increase the importance of temporal aggregation design and loss geometry, since the backbone is no longer the only source of performance gains.

Generative augmentation has also expanded from earlier GAN-based synthesis toward diffusion-based controllable diversity, particularly to address rare poses, occlusions, and viewpoint bias. Pose-dIVE [156] uses diffusion-based augmentation conditioned on pose/view to diversify underrepresented configurations, while Diffusion-Re-ID [157] proposes a generation-and-filtering paradigm and introduces large-scale synthetic data to strengthen Re-ID pretraining. These methods complement the cross-platform and cross-domain motivation by enlarging the effective training distribution while attempting to control noise through filtering.

Beyond representation learning, practical robustness has received renewed attention through explicit input-quality control. Mamedov et al. [158] proposed a filter module that pre-filters problematic inputs before Re-ID inference, motivated by real-world cases where low-quality detections and corrupted observations degrade matching reliability. Finally, broader vision trends increasingly influence how complex visual tasks are framed. Ke et al. [159]

survey compositional visual reasoning and emphasize multi-step reasoning with intermediate concepts, while Lin [160] documents rapid growth in VLM and diffusion-driven paradigms. While these surveys are not Re-ID methods, they reinforce the direction that modern systems increasingly move beyond monolithic feature matching toward richer representations, stronger priors, and more structured reasoning.

Table 2.3 summarizes representative supervised video Re-ID methods, but it should be interpreted with caution. The literature provides a much clearer ranking on MARS than on other datasets because MARS is the most consistently reported benchmark. By contrast, reporting on Duke, iLIDS-VID, LS-VID, and more recent cross-platform settings is less complete and less uniform. This imbalance makes it difficult to determine whether improvements reflect general progress in video Re-ID or optimization toward a small subset of standard benchmarks. It also reveals a broader limitation of the literature: many methods improve retrieval accuracy by adding architectural complexity, stronger pretraining, or specialized modules, but the field is less clear about which of these gains remain robust under stronger viewpoint variations, larger domain shifts, or more practical deployment constraints.

In summary, supervised video Re-ID evolved from CNN optimization and attention-based temporal aggregation to transformer-based and foundation-driven spatiotemporal reasoning under more complex deployment constraints in the last decade. This evolution motivates the thesis focus in Chapter 5 on transformer-based temporal modelling and evidence fusion, and it supports the generalization and robustness analyses in Chapter 6 when domain gaps are introduced.

No.	year	Model	Duke		Mars		iLIDS-VID		cite
			mAP	Rank1	mAP	Rank1	mAP	Rank1	
1	2020	Resnet50 + Multi-Granular Hypergraphs + mutual information minimization	-	-	85.8	90.0	-	85.6	[161]
2	2020	Resnet50 + Multi-Granularity Attention	-	-	85.9	88.8	-	88.6	[162]
3	2020	Resnet50 + Spatial GCN + Temporal GCN	95.7	97.29	83.7	89.95	-	-	[163]
4	2020	Resnet50 + Adversarial Feature Augmentation (AFA)	95.4	97.2	82.9	90.2	-	88.5	[164]
5	2020	Resnet50 + Temporal Saliency Erasing + Temporal Saliency Boosting (TSB)	96.2	96.9	85.1	89.8	-	86.6	[165]
6	2020	Resnet50 + Attribute-aware Identity-hard Triplet Loss + Attribute-driven Spatio-Temporal Attention	95.3	95.4	84.4	88.2	-	-	[166]
7	2021	Resnet50 + global correlation estimation + temporal reciprocating learning	-	-	84.8	91.0	-	90.4	[167]
8	2021	Resnet50 + bilateral complementary network	96.1	96.3	86.0	90.2	-	-	[139]
9	2021	Resnet50 + key-points estimator + 3D graph layers	-	-	86.7	91.4	-	89.7	[168]
10	2021	Resnet50 + factorized attention	96.4	97.4	86.1	90.3	-	89.3	[169]
11	2021	Resnet50 + transformer	97.1	97.6	87.0	90.8	-	92.0	[170]
12	2021	Resnet50 + spatial-temporal aggregation	97.4	98.3	85.8	91.5	-	91.5	[171]
13	2021	Resnet50 + Warmup + Random Erasing + non-local attention + Label Smoothing + BNNeck + Center Loss + TP	95.4	94.9	87.6	83.0	83.2	98.3	[140]
14	2022	Contextual Alignment Vision Transformer + residual position and multi-shape patch embedding	-	-	87.2	90.8	93.3	98.0	[172]
15	2022	Resnet50 + Salient-to-Broad Module	-	-	86.2	91.0	92.5	97.6	[173]
16	2022	multidirection and multiscale Pyramid Transformer	-	-	86.8	90.22	-	96.5	[147]
17	2022	ViT + Region-Enhanced Tokenization	-	-	90.68	78.61	-	74.16	[174]
18	2022	ViT + local and Global Features	-	-	86.36	94.67	-	96.63	[125]
19	2023	ViT + Resnet50	-	-	87.5	92.3	91.7	98.4	[150]
20	2023	Resnet50 + Multigranularity Appearance Extractor	-	-	86.8	91.6	92.2	98.6	[149]
21	2024	ViT-B/16 + CLIP encoder	-	-	88.1	91.7	-	-	[155]
22	2024	Resnet50 + Crossview Transformer	97.5	97.8	85.8	91.2	91.3	98.6	[148]
23	2024	ViT + Temporal Correlation	-	-	87.6	91.7	94.3	99.3	[151]
24	2025	ViT + Cross-Attention Aggregation	-	-	89.9	95.92	-	-	[126]
25	2025	ViT-L/14 + CLIP encoder	-	-	91.5	93.2	96.3	98.5	[152]

TABLE 2.3: Supervised video Re-ID methods (2020–2025).

Chapter 3

Methodology Overview

This chapter outlines the overarching methodology for investigating person Re-ID and focuses on the experimental setup, including the datasets and evaluation metrics that form the foundation for the Re-ID models developed and evaluated in subsequent chapters. It introduces the datasets used in this thesis, summarizes their sources and key attributes, and presents the common preprocessing conventions applied before training and evaluation. It also reviews the evaluation metrics used to assess model performance, thereby establishing a consistent and rigorous assessment framework. This overview sets the stage for the specific methodologies discussed in Chapters 4, 5, and 6, which focus on the implementation of unsupervised, supervised, and semi-supervised models, respectively.

3.1 Datasets

We will focus only on the general large-scale video datasets for deep learning methods. However, to be comprehensive, we will include all widely used Re-ID datasets in this section.

This thesis notes known ethical and privacy concerns surrounding certain legacy benchmarks. In particular, the DukeMTMC-ReID dataset and its derived subsets were later retracted/taken down due to ethical and privacy concerns. Accordingly, they are discussed in this thesis only for historical

comparison with prior person Re-ID literature, rather than as an endorsement of their continued use.

3.1.1 Image Datasets

The Market-1501 dataset

The Market-1501 dataset is one of the most prominent person Re-ID datasets, which contains 32,668 gallery images and 3,368 query images captured by six cameras. It also includes 500,000 irrelevant images as a distractor set, which makes the dataset even more challenging. This dataset is captured in a noisy campus environment. It is very challenging due to illumination and viewpoint changes across cameras, as well as occlusions in the views [175].

The DukeMTMC-ReID dataset

The DukeMTMC-ReID dataset is a subset of the DukeMTMC dataset. It contains 1,812 identities captured by eight cameras. A total of 1,404 identities appear in more than two cameras, and the other 408 IDs are distractor images [73].

The CUHK03 dataset

The CUHK03 dataset contains 13,164 images of 1,360 pedestrians captured by six cameras. Each identity appears in two disjoint camera views. This dataset is challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, and occlusions [33].

The PETA dataset

The PETA dataset is a substantial person attribute recognition dataset annotated with 61 binary attributes and four multi-class attributes for 19,000 images [176].

3.1.2 Video Datasets

The DukeMTMC-VideoReID dataset

The DukeMTMC-VideoReID dataset is a subset of DukeMTMC, which is specially designed for video-based Re-ID. Since this dataset is manually annotated, each identity has only one tracklet per camera. The pedestrian images are cropped from the videos for 12 frames every second to generate a tracklet. The dataset of identities is split according to the protocol in [175], i.e., 702 for training, 702 for testing, and 408 as the distractors. In total, 369,656 frames of 2,196 tracklets are generated for training, and 445,764 frames of 2,636 tracklets are generated for testing and distractors [177].

The MARS dataset

The MARS dataset is the largest video Re-ID dataset, which contains 1,261 individuals and around 20,000 video sequences captured by six cameras. The MARS dataset is split into train and test sets, containing 631 and 630 identities, respectively. Each identity has an average of 13.2 tracklets. The dataset is captured in a noisy campus environment, thus suffering from significant viewpoint changes, pose variations, and illumination changes [118].

The iLIDS-VID dataset

The iLIDS-VID dataset contains 300 individuals, where two cameras capture each annotated identity. The dataset contains approximately 600 hand-labelled tracklets [178].

The LPW dataset

The LPW dataset contains 2,731 pedestrians in three scenes, where four cameras capture each annotated identity. The data features a notable scale of 7,694 tracklets, known for their cleanliness, with over 590,000 images [179].

FIGURE 3.1: Compared with existing datasets, LS-VID presents more substantial variations of lighting, scene, and background.

The LS-VID dataset

The LS-VID dataset, represented in Figure 3.1, contains 3,772 individuals and around 14,943 video sequences captured by 15 cameras. The dataset is divided into train and test sets, containing 842 and 200 identities, respectively [180].

The AG-VPReID dataset

The AG-VPReID dataset shown in Figure 3.2, is a large-scale benchmark for aerial-ground video-based person re-identification. It comprises 6,632

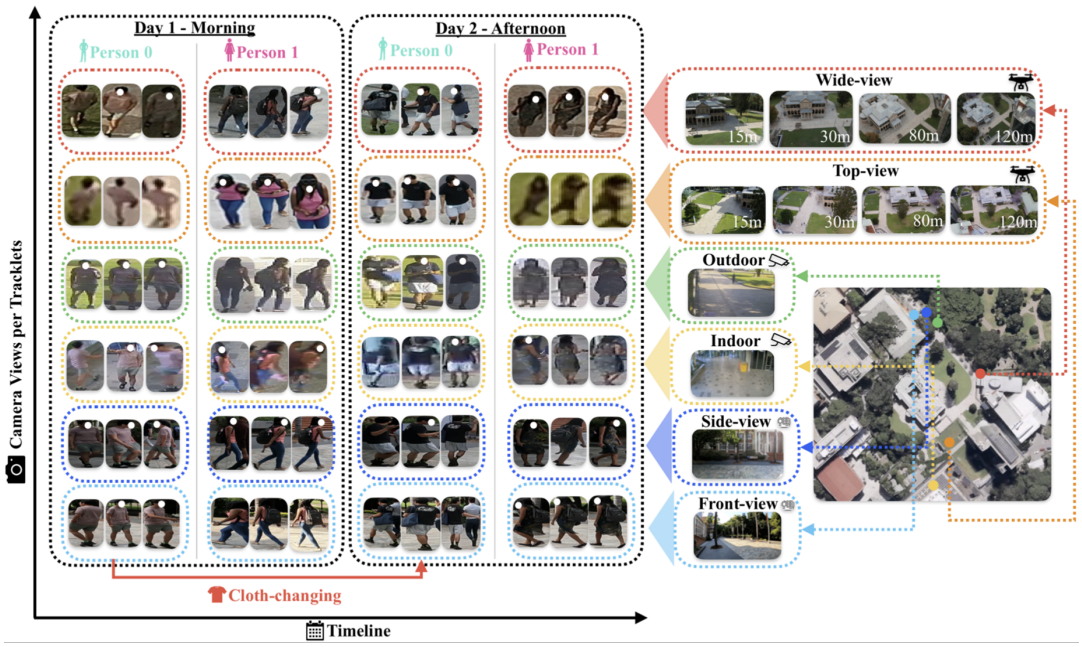


FIGURE 3.2: AG-VPReID was captured using six cameras (drones, CCTVs, and GoPros). Sample images and camera placements (right) illustrate cross-camera appearance variations for two pedestrians across different sessions and times of day (left).

subjects, 32,321 tracklets, and over 9.6 million frames captured using multiple platforms, including aerial drones operating at altitudes between 15–120 meters, stationary CCTV cameras, and wearable cameras. The dataset is collected over multiple sessions across 20 days, which introduces significant variations in viewpoint, scale, and resolution between aerial and ground observations. Following the evaluation protocol provided with the dataset, a balanced subset of 3,013 identities that appear in both aerial and ground views is used for training and testing, where 1,555 identities are used for training, and 1,456 identities are used for testing, with additional identities used as distractors in some evaluation settings [153].

The datasets described above vary in scale, number of cameras, and sequence characteristics, reflecting different challenges in video-based person re-identification. Table 3.1 provides a consolidated summary of these datasets, including their key statistics and properties.

TABLE 3.1: Summary of video-based person Re-ID datasets.

Dataset	Train Set	Test Set	Total Tracklets	Avg. F/T	Cameras #
Duke	702 identities	702 identities	2,196 tracklets	167	8
MARS	631 identities	630 identities	20,000 tracklets	59	6
iLIDS-VID	300 identities	300 identities	600 tracklets	40	2
LPW	2,731 identities	2,731 identities	7,694 tracklets	76	4
LS-VID	842 identities	200 identities	14,943 tracklets	199	15
AG-VPReID	1,555 identities	1,456 identities	32,321 tracklets	297	6

3.1.3 Common Preprocessing and Data Preparation

Although the three methodological settings studied in this thesis differ in their training strategies, several preprocessing conventions are shared across the proposed models. First, video frames are resized to a fixed spatial resolution before being passed to the network, ensuring consistent input dimensions and a fair comparison across datasets. Unless otherwise stated in later implementation sections, the input resolution is 256×128 pixels.

Second, frames are converted to tensors and normalized channel-wise to make the input distribution compatible with ImageNet-pretrained backbones. Third, standard data augmentation is applied during training to improve robustness to appearance variation and occlusion. The main augmentations used throughout this thesis include horizontal flipping and random erasing, while any chapter-specific additions are described in the relevant implementation section.

Fourth, each video tracklet is represented by a controlled subset of frames rather than by every frame in the original sequence. This reduces redundancy and keeps computation tractable while preserving the temporal information needed for video-based person Re-ID. Unless otherwise specified, tracklets are sampled using 16 keyframes distributed uniformly across the temporal duration of the sequence. The exact frame-selection or sampling strategy may vary depending on the learning setting and is therefore specified for each proposed model where deviations from this default occur. Finally, when official dataset protocols are available, they are followed for train/test

TABLE 3.2: Mapping between the video-based datasets introduced in Chapter 3 and their use across Chapters 4, 5, and 6.

Dataset	Chapter 4	Chapter 5	Chapter 6
Duke	Comparison only	–	Comparison only
MARS	Used	Used	Used
iLIDS-VID	–	Used	–
LS-VID	Used	Used	Used
AG-VPReID	Used	Used	Used

partitioning and evaluation. When such protocols do not fully match the intended deployment setting of a chapter, an adapted protocol is defined explicitly and justified in that chapter.

3.1.4 Dataset Usage Across Thesis Chapters

Not all datasets described in this chapter are used uniformly throughout the thesis. Since Chapters 4, 5, and 6 address different learning settings, the role of each dataset varies accordingly. In general, the image-based datasets listed above are included for completeness and background context, whereas the main experimental evaluation in this thesis relies on video-based benchmarks. Table 3.2 summarizes the mapping between the datasets introduced in Chapter 3 and their use in the later chapters.

DukeMTMC-VideoReID is not used for experimental validation in this thesis. In Chapter 5, iLIDS-VID is used instead for smaller-scale supervised evaluation. DukeMTMC-VideoReID is included only in Chapters 4 and 6 for comparison with prior literature, given its historical prominence as a benchmark despite the ethical concerns associated with DukeMTMC and its derived subsets.

3.2 Evaluation Metrics

In this section, we discuss the two commonly used evaluation metrics of person Re-ID algorithms, followed by other notable metrics mentioned in the literature.

3.2.1 Cumulative Matching Characteristics

Cumulative Matching Characteristics (CMC) [181] is generated by the sum of each image's accuracy, divided by the total number of images in a query. Ranked CMC [181] expresses the probability of a correct matching in the k-th rank target in a gallery of targets. Since CMC accuracy can be affected by multiple ground truths in a query, we usually couple it with other evaluation metrics. This limitation stems from the fact that it only retrieves the first match it encounters. However, most large-scale datasets contain multiple ground truths, and CMC cannot fully capture a model's capability to discriminate across multiple cameras.

To further illustrate the construction of a Rank-K list, let χ represent the query image and let G represent the gallery set containing N images, where :

$$G = \{g_1, \ldots, g_N\}$$

The Mahalanobis distance between χ and each gallery image $g_i \in G$ is then computed using Equation 3.1 :

$$D(\chi, g_i) = (f_\chi - f_{g_i})^\top \mathbf{M} (f_\chi - f_{g_i}) \tag{3.1}$$

where f_χ and f_{g_i} denote the feature vectors of the query χ and gallery image g_i, respectively, and $\mathbf{M}$ is a real symmetric positive semidefinite matrix that defines the Mahalanobis metric. The positive semidefinite constraint ensures that, for any feature difference vector $\mathbf{z} = f_\chi - f_{g_i}$, the quadratic form

$\mathbf{z}^T\mathbf{M}\mathbf{z} \geq 0$, so the resulting distance is always non-negative. A simple special case is the identity matrix, which is also positive semidefinite.

We can define the K-Rank list K as a list of κ nearest neighbour gallery images $G = \{g_1, ..., g_\kappa\}$ to the query χ, where $D(\chi, g_i) < D(\chi, g_{i+1})$. As shown in equation 3.2, where $\|\cdot\|$ denotes the number of candidates in the k-Rank list K.

$$K(\chi, \kappa) = \{g_1, ..., g_\kappa\}, where \|K(\chi, \kappa)\| = \kappa \tag{3.2}$$

Given this logic, a $K(\chi, 5)$ list contains the top 5 candidates for the query image χ. It is widespread in the literature to have $K(\chi, 1)$ (known as rank-1) when the model's accuracy is very high and only include $K(\chi, 5)$ and $K(\chi, 20)$ if rank-1 is low.

3.2.2 Mean Average Precision

The mean Average Precision (mAP) [175] is a ranking-based evaluation metric that measures how well a retrieval system returns all relevant matches and how early these matches appear in the ranked list. In person Re-ID, this is particularly important because a query may have multiple correct matches in the gallery. Unlike CMC, which mainly reflects whether a correct match appears within the top ranks, mAP evaluates the overall quality of the ranked retrieval results across all relevant matches. For this reason, mAP is typically reported together with CMC.

More specifically, person Re-ID evaluation can be formulated as a ranking problem in which, for each query, the model produces a ranked list of gallery samples ordered by similarity or distance. Precision at rank k, denoted as $P(k)$, represents the fraction of correct matches among the top-k retrieved samples. However, precision alone does not indicate whether the correct matches appear early in the ranking or only after many incorrect candidates.

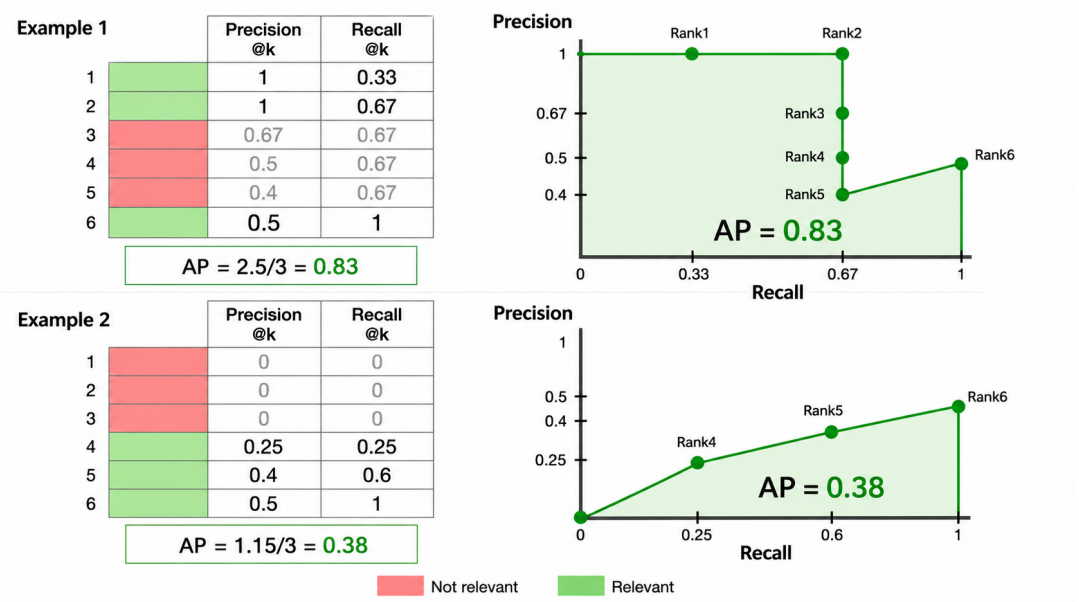


FIGURE 3.3: Illustration of Average Precision (AP) for two ranked retrieval lists. Relevant matches are shown in green and non-relevant matches in red. When relevant matches appear early in the ranking, AP is higher; when they appear later, AP decreases.

Average Precision (AP) addresses this limitation by averaging precision values only at the ranks where relevant matches are retrieved. Let $\text{rel}(k)$ be equal to 1 if the item at rank k is relevant and 0 otherwise. For a ranked list of length K, AP can be written as:

$$AP@K = \frac{1}{R} \sum_{k=1}^{K} P(k)\,\text{rel}(k), \tag{3.3}$$

where R is the number of relevant gallery samples for the query. Therefore, correct matches retrieved near the top of the ranked list contribute more to AP than correct matches retrieved later. Figure 3.3 illustrates this behaviour: the first ranking achieves a higher AP because relevant samples appear early, while the second ranking has a lower AP because relevant samples are delayed until lower ranks.

The mean Average Precision is then obtained by averaging AP over all queries:

$$mAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q), \tag{3.4}$$

where Q is the total number of queries and $AP(q)$ is the average precision for query q. Thus, mAP reflects both retrieval completeness and ranking quality, making it more informative than CMC when multiple ground-truth matches exist for the same query.

3.2.3 Other Metrics

Even though CMC and mAP can demonstrate the model's accuracy, they don't reveal where the model struggles or its complexity. Therefore, additional metrics like Mean Inverse Negative Penalty and Floating Point Operations per second provide a more comprehensive evaluation.

- **Mean Inverse Negative Penalty** (mINP) [140] can be a supplementary metric to the commonly used CMC and mAP metrics. By applying an inverse negative penalty for the hardest correct matches, mINP avoids the domination of easy matches in the mAP/CMC evaluation. In practice, it emphasizes whether the model can retrieve *all* correct matches for a query without pushing the last (hardest) correct match too far down the ranked list.

  An intuitive way to understand mINP is to focus on the position of the hardest correct match. A model may retrieve a few easy matches early, which can still lead to good CMC and even reasonable mAP, while the final correct match appears much later after many incorrect candidates. This situation indicates that the ranking is not consistently reliable across all true matches. This behaviour is explicitly penalized by mINP, tying the score to the rank position of the last correct match (relative to the number of correct matches) and then using an inverse formulation so

that higher values still indicate better performance. As a result, mINP complements mAP by highlighting cases where a model performs well on easy matches but struggles with the most difficult correct matches.

- **Floating-point Operations Per second** (FLOPs) is a crucial evaluation metric when the training/testing device has limited computational resources. This metric will be considered when the efficiency and complexity of the training model are significant. Recent works [182], [183] have used it to evaluate the efficiency of their models.

3.3 Efficiency Considerations

One of the targets of this thesis is to design a Re-ID model that can run on minimal resources, addressing both efficiency and scalability for practical model deployment.

3.3.1 Lightweight Models

One way to address the scalability issue is to implement a lightweight Re-ID model. Various studies [181], [182], [183] modified some of the existing network architectures to achieve lightweight models. This was achieved by optimizing the neural network model to reduce the memory and computational demands of the algorithm [184]. Recent advancements have furthered this approach by moving away from heavy, general-purpose backbones like ResNet toward specialized architectures. For instance, the Combined Depth Space (CDS) network utilizes width and resolution multipliers to create an efficient backbone specifically tailored for pedestrian retrieval, balancing triplet and SoftMax losses through a fine-grained balance neck [185]. Furthermore, hybrid architectures like MixNet integrate shallow CNN modules with deep Transformers to capture both local features and long-distance dependencies

while employing pruning algorithms to minimize parameters and computational costs [186]. To maintain high accuracy in complex environments, lightweight models such as MSFENet now incorporate spatial-frequency fusion and multi-granularity modules to suppress background noise and prevent information loss during feature interaction [187]. These optimizations are essential for deploying Re-ID systems on edge devices where computational resources are limited but real-time processing is required [188].

3.3.2 Video Representation

One way to address the efficiency issue is to implement a keyframe extraction model to preprocess tracklets into representative keyframes. This approach aims to reduce the Re-ID model's processing time and complexity. Since most frames in a tracklet are similar and non-essential to the learning process, we can omit them and keep only the most representative frames, thus reducing the size of the dataset and the processing power required to run the model. Some studies [189], [190] proposed video summarization models for resource-constrained devices in real-time. While this approach might seem to shift the problem from video Re-ID to image Re-ID, it does not. The problem remains in the video Re-ID domain if each tracklet contains enough frames to be considered a video.

3.3.3 Distributed Learning

When considering edge computing devices, their limited resources challenge many Re-ID solutions that require resource-intensive tasks for training machine learning models. Gutierrez-Torre et al. [191] proposed a method for efficiently training deep learning models by utilizing low-power and resource-constrained edge devices while ensuring good estimation accuracy, thus contributing to the progress towards distributed learning.

Chapter 4

One-Shot Video Re-ID

This chapter studies the cost-sensitive one-shot regime introduced earlier in the thesis, where video-based person Re-ID must operate with minimal manual intervention and limited computational resources. Building on the deployment inspiration established in Chapter 1, the weakly supervised and pseudo-labelling literature reviewed in Chapter 2, the datasets and evaluation protocol defined in Chapter 3, this chapter develops a resource-aware framework that starts its training phase with a labelled tracklet per identity and progressively expands supervision through conservative pseudo-labelling.

4.1 Introduction

As discussed in Chapter 1, the thesis examines video-based person Re-ID under three deployment regimes that differ in resource availability, supervision level, and deployment difficulty. This chapter focuses on the first of these regimes: a cost-sensitive setting in which only minimal identity supervision is available and the Re-ID model must remain practical for small-to-medium camera networks.

In this setting, the role of Re-ID is deliberately limited. Rather than defining it as a fully general end-to-end surveillance solution, it's considered a re-association module used after tracking failure, consistent with the earlier-defined thesis scope. A lightweight tracker is assumed to operate

continuously within each camera stream, while the Re-ID component is invoked only when identity continuity is lost because of occlusion, target disappearance, or re-entry into the scene. This deployment assumption keeps the problem operationally focused and matches the resource-aware objective of this chapter.

The main difficulty in this regime is that fully labelled video Re-ID training data are rarely available in practice. To make learning feasible under this constraint, we adopt the one-shot assumption introduced earlier in the thesis: one labelled tracklet per identity can be acquired at a designated entry point in the camera network, such as a gate, doorway, or corridor. This yields a small anchor set that does not capture full intra-identity variation but still provides a reliable starting point for learning.

The methodological problem addressed in this chapter is therefore not how to train a fully supervised Re-ID model, but how to expand supervision from a minimal anchor set while controlling pseudo-label noise and maintaining low computational cost. To do so, the proposed framework progressively adds only the most confident pseudo-labelled tracklets to the training set, allowing the representation to improve gradually without relying on heavy clustering machinery or large-capacity training pipelines.

Accordingly, this chapter develops a one-shot video Re-ID framework with three main design priorities: reliable supervision expansion, video-level temporal representation, and practical efficiency. The framework begins from a single labelled tracklet per identity, uses conservative similarity-based pseudo-labelling to enlarge the labelled pool over time, and represents each tracklet as a compact video descriptor to ensure that temporal information can be used without abandoning the deployment-oriented constraints of this regime.

It is important to clarify the scope of this regime. The proposed framework does not attempt to solve full online identity management for an

unconstrained surveillance system. Instead, consistent with the thesis scope defined in Chapter 1, it assumes that multi-person scenes have already been processed by a tracker, meaning the Re-ID module operates on person tracklets rather than on raw frames containing multiple detections. The entry-point mechanism is therefore used only as a practical way to obtain an initial anchor set in controlled deployments, not as a claim that every active identity can always be captured perfectly and assigned automatically at the moment of entry. In practice, cases such as missed detections at entry, ambiguous re-entries, or tracklets that are not reliably represented in the initial anchor set can be handled during a limited preprocessing stage through manual verification, similarity-based grouping, and targeted inspection of atypical or highly divergent tracklets. Although the entry-point mechanism introduces some additional annotation cost, it remains substantially cheaper than exhaustively annotating the full dataset. Accordingly, this chapter studies whether such a partially constructed anchor set is sufficient to support useful Re-ID behaviour within the temporally bounded gallery setting introduced earlier in the thesis.

The remainder of this chapter is organized as follows. Section 4.2 presents the problem formulation, pseudo-labelling strategy, progressive learning scheme, and a temporal modelling approach. Section 4.3 describes the experimental protocol used in this regime. Section 4.4 reports the main comparative results, and the ablation study analyzes the contribution of the backbone choice, progressive expansion policy, and pseudo-label assignment strategy. Finally, the chapter concludes with the main findings for this one-shot setting.

4.2 Methodology

4.2.1 Problem Statement

Given a video tracklet dataset D_α collected from a camera network, we assume the availability of a small initial labelled set $D_\iota \subset D_\alpha$ containing a single labelled tracklet per identity in the network. This labelled set is operationally approximated by defining an entry point in the camera network and using it to collect an initial labelled tracklet for each active identity within the current surveillance window. The remaining tracklets form an unlabelled set D_μ such that $D_\alpha = D_\iota \cup D_\mu$ and $D_\iota \cap D_\mu = \emptyset$.

In this chapter, D_ι should be understood as an operationally constructed anchor set rather than as proof of a fully automated real-world labelling system. In practice, the entry-point assumption approximates a controlled deployment in which a designated camera region is used to obtain one initial tracklet per active identity within the current surveillance window. Under this formulation, the method is not required to maintain a globally fixed, permanently complete class list. Cases in which a person is not identified at entry, re-enters ambiguously, or is not reliably represented in D_ι can be handled during a limited preprocessing stage through manual verification, similarity-based grouping, and targeted inspection of atypical or highly divergent tracklets. Although this adds some annotation cost, it remains substantially cheaper than exhaustively annotating the full dataset. The purpose of the experimental protocol is therefore to evaluate the learning behaviour of the proposed framework under this constrained operating regime, rather than to claim a complete solution to open-ended online identity bookkeeping.

Our objective is to learn a discriminative visual representation model f_θ for one-shot video-based person Re-ID by leveraging both the labelled anchors in D_ι and the large unlabelled pool D_μ. The learning process is progressive:

at each iteration, the current model is used to estimate pseudo-labels for a subset of unlabelled tracklets. The most confident pseudo-labelled candidates are moved from D_μ to D_ι, and the model is retrained using the expanded labelled set. After each iteration, the model can be evaluated by ranking a query tracklet χ against a gallery set $G = \{g_1, \ldots, g_n\}$ according to distances computed in the learned feature space.

This setting is more challenging than typical semi-supervised Re-ID because supervision is limited to a single tracklet per identity. In addition, the initial labelled set is not intended to cover intra-identity variation; instead, it serves as an anchor that enables reliable association and gradual generalization as pseudo-labels are added.

4.2.2 Pseudo Labels Assignment Strategy

In the one-shot setting studied in this chapter, pseudo-label assignment must prioritize reliability over coverage. Since the initial supervision contains only one labelled tracklet per identity, early pseudo-label errors can easily propagate through later training iterations and degrade the representation. For this reason, the pseudo-labelling mechanism used here is designed to expand supervision conservatively, starting from the most reliable matches and gradually incorporating more difficult samples as the embedding becomes more discriminative.

Two broad design choices are possible in this setting: similarity-based assignment and clustering-based assignment. In similarity-based assignment, each unlabelled tracklet is matched directly to the current labelled anchor set in the learned feature space, and only the most confident matches are accepted as pseudo-labels. In clustering-based assignment, unlabelled tracklets are first partitioned into clusters, after which cluster structure is used to infer

pseudo-identities. Although both directions are relevant to weakly supervised Re-ID, they behave differently under the constraints of this chapter.

Our framework adopts similarity-based pseudo-labelling because it is more consistent with the one-shot and resource-aware assumptions of this chapter. First, the labelled anchors provide an explicit identity reference from the beginning of training, which makes nearest-anchor matching a natural way to expand supervision progressively. Second, this strategy avoids the need to estimate or refine cluster structure under extremely limited supervision, where cluster impurity can introduce merged identities or fragmented classes. Third, similarity-based assignment is simpler computationally and therefore better aligned with the deployment-oriented objective of this chapter, where the model is intended for small-to-medium camera networks with limited resources.

Accordingly, after each training stage, unlabelled tracklets are matched to the current labelled set using the learned embedding, and only the high-confidence matches are promoted into the labelled pool. This choice allows the model to benefit from progressive supervision expansion while reducing the risk of large-scale error injection. For completeness, clustering-based alternatives are still evaluated later in the ablation study, but they are treated as comparison strategies rather than as the primary design adopted in this chapter.

4.2.3 Progressive Labelling and Learning Strategies

Our framework alternates between two steps: (i) training a discriminative embedding model on the current labelled set D_ι, and (ii) expanding D_ι by adding a small set of high-confidence pseudo-labelled tracklets from D_μ. This gradual expansion is critical, since adding too many noisy pseudo-labels early can degrade the embedding and amplify errors in later iterations.

Let the current labelled set be $D_\iota = \{(\tau_1, \iota_1), \ldots, (\tau_n, \iota_n)\}$, where each tracklet τ_i belongs to identity ι_i, and let the unlabelled set be $D_\mu = \{\tau_1, \ldots, \tau_m\}$. After each expansion of D_ι, we retrain the model using an identification loss. Given a training sample x_i with an identity class C_i, we use the standard cross-entropy identification loss:

$$\ell_{id}(x, C) = -\sum_i \log p(C_i \mid x_i) \tag{4.1}$$

where $p(C_i \mid x_i)$ is the softmax probability predicted for class C_i.

After training, we estimate pseudo-labels for unlabelled tracklets by matching them to the labelled anchors in the embedding space. Let f_x be the embedding of an unlabelled tracklet τ_x and f_y be the embedding of a labelled tracklet $\tau_y \in D_\iota$. We define the verification (matching) cost as the squared Euclidean distance:

$$\ell_V(\tau_x, \tau_y) = \left\| f_x - f_y \right\|_2^2 \tag{4.2}$$

For each unlabelled tracklet τ_x, we find its closest labelled tracklet τ_y under ℓ_V and assign the corresponding identity ι_y as a pseudo-label for τ_x. We then select only the most confident pseudo-labelled candidates (smallest matching costs) to form a candidate set D_σ, which is added to D_ι.

Following the observation in [72] that pseudo-label set size impacts both performance and compute, we control the expansion rate using a generalization percentile (GP). Let t be the current iteration index and let $|D_\mu|$ be the number of remaining unlabelled tracklets. We define the GP at iteration t as

$$\mathrm{GP}(t) = \min(\mathrm{GP}_{\max}, \delta \cdot t) \tag{4.3}$$

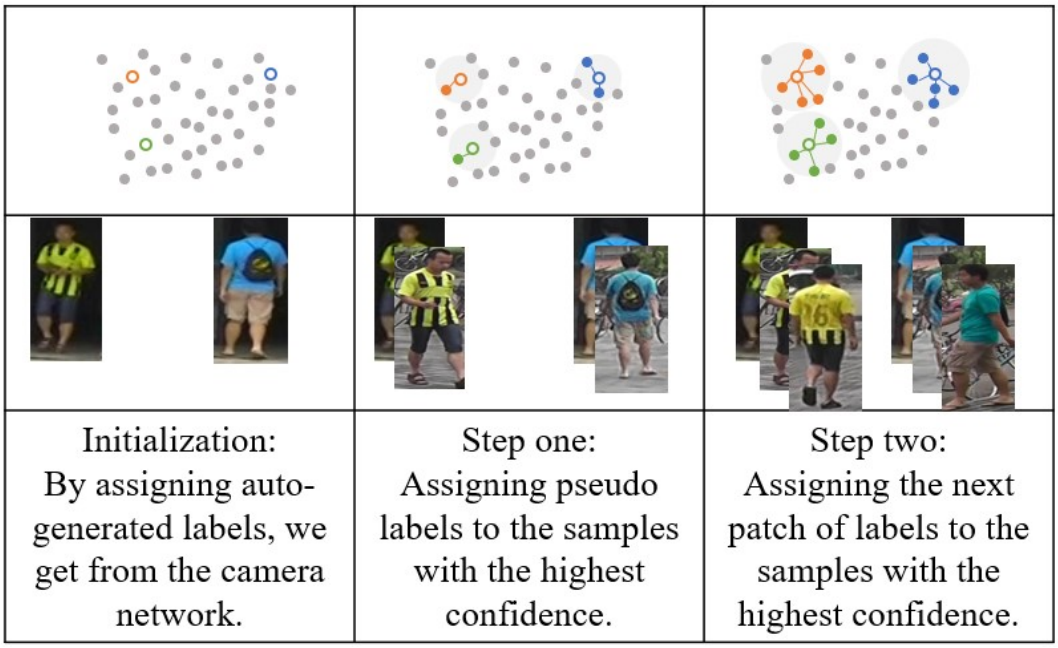


FIGURE 4.1: Our progressive labelling strategy starts from the auto-generated one-shot anchors. At each iteration, the model assigns pseudo-labels to the most confident unlabelled tracklets first, then gradually expands the labelled set.

where δ is a confidence saturation factor and $\mathrm{GP}_{\max}$ caps the maximum expansion rate. The pseudo-label set size is then

$$|D_\sigma| = \lceil \mathrm{GP}(t)\,|D_\mu| \rceil \tag{4.4}$$

In each iteration, we add the $|D_\sigma|$ pseudo-labelled candidates with the lowest verification costs to reduce the probability of injecting noisy labels early in training.

Figure 4.1 illustrates this progressive expansion process, where the labelled anchor set is enlarged iteratively by promoting only the most confident pseudo-labelled tracklets from the unlabelled pool.

4.2.4 Temporal Modelling

For video Re-ID, each tracklet contains N frames, and the backbone produces a frame-level feature vector f_i^n for each frame index $n \in [1, N]$. A tracklet-level

descriptor f_v is obtained by aggregating the frame-level features.

Max-pooling aggregates by selecting the strongest activations:

$$f_v = \max_n(f_i^n) \tag{4.5}$$

However, prior work has shown that average pooling is often more effective for set-based video descriptors [95]:

$$f_v = \frac{1}{N}\sum_{n=1}^{N} f_i^n \tag{4.6}$$

While average pooling provides a stable representation, it assigns equal importance to all frames, including blurred or partially occluded frames.

To better emphasize informative frames, we introduce a weighted temporal pooling strategy inspired by attention-based weighting [19], [192]. We first apply a convolutional layer to the stacked frame features, followed by a non-local attention layer [14] that predicts a weight $\omega^n \in [0, 1]$ for each frame. The tracklet descriptor is then computed as:

$$f_v = \sum_{n=1}^{N} \omega^n f_i^n \tag{4.7}$$

This formulation reduces the influence of noisy frames while retaining robust aggregation across the full sequence. In the results section, we compare set-to-set average pooling against this weighted temporal pooling.

4.2.5 Framework

Given the components above, our framework can be summarized as an iterative procedure that alternates between training on the current labelled set and expanding it using high-confidence pseudo-labels. The input to the algorithm is the initial labelled anchors D_ι, the unlabelled pool D_μ, the

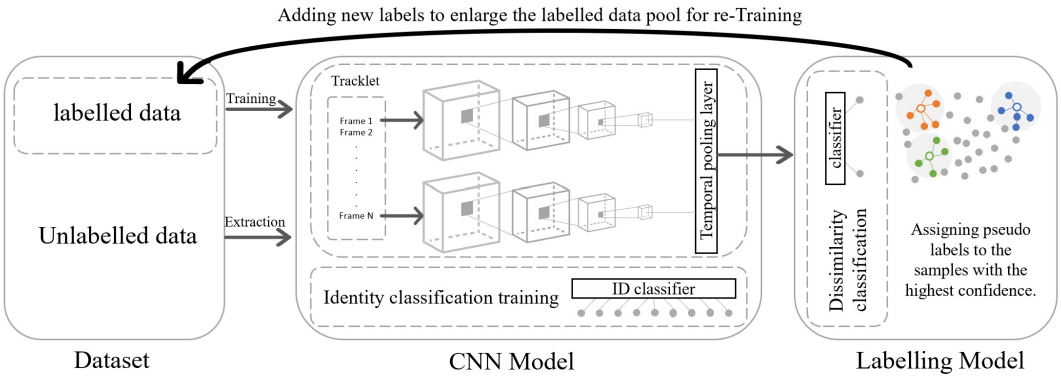


FIGURE 4.2: A schematic illustration of our framework. The model is trained using the labelled anchor set and progressively pseudo-labelled tracklets. After each iteration, new pseudo-labels are assigned to a controlled fraction of the unlabelled pool and added to the labelled set for the next iteration.

confidence saturation factor δ, and an ImageNet-pretrained CNN backbone f_θ. The output is the final trained model for one-shot video Re-ID.

Algorithm 1 Training procedure for the proposed one-shot video Re-ID framework

Input: labelled data D_ι, unlabelled data D_μ, confidence factor δ, initialized CNN model f_θ

Output: trained CNN model f_θ

step $t \leftarrow 0$
while $D_\mu \neq \emptyset$ **do**
    $t \leftarrow t + 1$
    Train f_θ on D_ι using the identification loss in (4.1)
    Compute embeddings for D_ι and D_μ, then assign pseudo-labels by nearest labelled match using (4.2)
    Compute the expansion fraction using (4.3) and select $|D_\sigma|$ candidates using (4.4)
    $D_\sigma \leftarrow$ the selected pseudo-labelled candidates with the lowest verification costs
    $D_\iota \leftarrow D_\iota \cup D_\sigma$
    $D_\mu \leftarrow D_\mu - D_\sigma$

The overall interaction between the labelled anchor set, the pseudo-labelling step, and the iterative retraining process is illustrated in Figure 4.2. Algorithm 1 summarizes the corresponding training procedure.

4.3 Experiments

This section evaluates our one-shot video Re-ID framework under the deployment regime emphasized in this chapter: limited supervision and practical compute constraints. We follow standard dataset protocols whenever available, and we enforce the same supervision assumption across datasets by labelling only one tracklet per identity to initialize the labelled set D_l, while treating the remaining tracklets as unlabelled data D_μ for progressive pseudo-labelling. In addition to reporting retrieval accuracy, we explicitly study the efficiency-accuracy trade-off by testing multiple backbone networks, including lightweight options suitable for cost-sensitive and small-scale camera deployments.

Unless otherwise stated, we adopt an ImageNet-pretrained backbone and train with a warm-up-based schedule for stable optimization. Performance is reported using standard video Re-ID metrics, and we provide ablations to quantify the contributions of backbone choice, regularization, and pseudo-labelling strategy under the one-shot constraint.

4.3.1 Dataset Splits

We evaluate our one-shot video Re-ID framework on four benchmarks: DukeMTMC-VideoReID, MARS, LS-Vid, and AG-VPReID. Full dataset descriptions are provided in Section 3.1.2. Here, we only describe the split and initialization protocol used in this chapter.

For all datasets, we initialize the labelled anchor set D_l using one labelled tracklet per identity and treat the remaining training tracklets as unlabelled data D_μ for progressive pseudo-labelling. When an official protocol is established, we follow it; otherwise, we define a one-shot initialization strategy consistent with the entry-point assumption described earlier in this chapter.

- **Duke:** We follow the standard train/test split in [73]. Within the training set, one tracklet per identity is retained as the labelled anchor, while the remaining training tracklets are treated as unlabelled.
- **MARS:** We follow the standard train/test split of [118] and apply the same one-shot initialization, keeping one labelled tracklet per identity and treating the remaining training tracklets as unlabelled.
- **LS-Vid:** Since there is no established one-shot split protocol matching our entry-point assumption, we construct D_l by selecting one tracklet per identity from camera 0, which serves as the designated entry-point camera. All remaining tracklets of those identities are used as unlabelled data D_μ.
- **AG-VPReID:** We follow the original split defined by the dataset protocol for the aerial-to-ground and ground-to-aerial evaluation directions. Within the training portion of each direction, one tracklet per identity is used to form D_l, and the remaining training tracklets are treated as unlabelled.

4.3.2 Experimental Settings

We evaluate the proposed one-shot video Re-ID framework using the standard video Re-ID retrieval metrics defined in Chapter 3. In particular, we report CMC at Rank-1, Rank-5, and Rank-20, together with mean average precision (mAP). Since evaluation is performed at the tracklet level, each query corresponds to a query tracklet, and the gallery contains tracklets from other cameras or from the corresponding evaluation split, depending on the dataset protocol. In this chapter, CMC is used to reflect top-k retrieval accuracy, while mAP provides a more complete assessment when multiple ground-truth matches exist in the gallery.

4.3.3 Implementation Details

All experiments are conducted using a video tracklet representation in which each tracklet is processed into a fixed number of frames and encoded by a CNN backbone. Unless otherwise stated, we resize each frame to 256×128, apply standard data augmentation (random horizontal flipping, cropping, and erasing), and normalize using ImageNet statistics. To study the efficiency-accuracy trade-off emphasized in this chapter, we evaluate multiple backbone networks, ranging from lightweight architectures to stronger, more resource-demanding models. In particular, we report results using MobileNet and ShuffleNet as resource-efficient backbones, and ResNet-50 variants as higher-capacity baselines. For the ResNet-50 baseline, we also evaluate the effect of dropout regularization, since it consistently improves generalization under limited supervision.

One-shot initialization. For each dataset, we construct the labelled anchor set D_l by selecting one tracklet per identity, following the dataset protocol when available and the dataset-specific initialization strategy described in the Datasets subsection. The remaining training tracklets form the unlabelled set D_μ. During training, we maintain the one-shot constraint by never using additional manually verified labels beyond D_l.

Progressive pseudo-labelling. Training proceeds in stages. At each stage, the model is trained on D_l and a subset of D_μ that is assigned pseudo-labels based on the current model's predictions. We only accept pseudo-labels that satisfy a confidence criterion to reduce error propagation, and we gradually expand the pool of pseudo-labelled examples as training progresses. This strategy allows the model to improve its representation while limiting the impact of noisy assignments, which is critical under the one-shot supervision assumption.

Optimization and schedule. We initialize the backbone with ImageNet pretraining and optimize it using Adam with weight decay. The learning rate follows a warm-up schedule during the early phase to stabilize optimization, then decays in steps. The training procedure is carried out progressively in stages, with each stage incorporating an additional batch of pseudo-labelled samples. For each stage, the backbone is trained for a fixed maximum number of intervals, and early stopping is used only as a local criterion within that stage when validation performance stops improving after several iterations. Thus, early stopping is used to conclude optimization for the current pseudo-label batch rather than to terminate the full algorithm. Once this condition is met, the method proceeds to the next pseudo-label expansion step, and the overall process continues until the unlabelled pool has been fully consumed, unless otherwise stated. We use a batch sampling strategy that selects 8 identities per batch, with 4 tracklets per identity when available, stabilizing the metric learning objective, unless otherwise noted.

Inference. At inference time, each query and gallery tracklet is encoded into a single descriptor by aggregating frame-level features using temporal pooling. Retrieval is then performed by ranking gallery descriptors according to their distance from the query descriptor, and performance is reported using the metrics described above.

4.4 Results and Comparison

We evaluate the proposed framework against representative prior video Re-ID methods on the MARS and Duke datasets. The results are summarized in Table 4.1. Using only a ResNet50 backbone with dropout, our method (GVOUReID) achieves 55.5% mAP and 72.2% Rank-1 accuracy on MARS and 72.9% mAP and 78.6% Rank-1 accuracy on Duke. Relative to the baseline, this

TABLE 4.1: Evaluation results on video-based datasets DukeMTMC-VideoReID and MARS using mAP and CMC

Method	Year	MARS				Duke			
		mAP	Rank-1	Rank-5	Rank20	mAP	Rank-1	Rank-5	Rank20
baseline	-	15.45	36.16	50.20	61.86	33.27	39.60	56.84	66.95
OIM [135]	2017	13.50	33.70	48.10	-	43.80	51.10	70.50	-
Stepwise [95]	2017	19.65	41.21	55.55	66.76	46.76	56.26	70.37	79.20
RACE [97]	2018	24.50	43.20	57.10	67.60	-	-	-	-
DAL [136]	2018	23.00	49.30	65.90	77.90	-	-	-	-
EUG [72]	2018	42.45	62.67	74.94	82.57	63.23	72.79	84.18	91.45
PGOR [52]	2018	46.51	66.50	76.38	84.69	68.42	72.47	86.38	94.45
PL [96]	2019	42.60	62.80	75.20	83.80	63.30	72.90	84.30	91.40
DGM [98]	2019	16.87	36.81	54.01	68.51	33.62	42.36	57.92	69.31
DBC [99]	2019	43.8	64.3	79.2	87.8	66.1	75.2	87	-
TCPL [100]	2020	43.6	65.2	77.5	-	67.9	76.8	87.8	-
SSLR [92]	2020	43.6	62.8	77.2	-	69.3	76.4	88.7	-
SCL [101]	2022	46.6	66.6	77	-	78.4	82.2	93.2	-
RPE [133]	2024	40.4	63.3	75.4	80.6	71.5	77.8	89.3	91.7
Weakly [112]	2024	**75.2**	**83.9**	**93.6**	**95.5**	87.4	89.7	97.2	98.0
DRMPCL [134]	2025	71.6	82.0	91.1	93.4	**92.7**	**94.0**	**98.6**	**99.4**
StS	**-**	49.7	68.2	82	88.40	66.50	74.10	86.20	92.30
GVOUReID	**-**	55.5	72.2	84.4	89.6	72.9	78.6	88.5	93.7

corresponds to improvements of +13.4% mAP and +15.8% Rank-1 on MARS and +9.0% mAP and +2.7% Rank-1 on Duke.

The strongest recent entries in Table 4.1 should be interpreted in light of their training assumptions. In particular, the last two methods in the table (Weakly and DRMPCL) rely on stronger supervision and/or more compute-intensive training pipelines than the lightweight one-shot regime considered in this chapter. The Weakly method uses video-level labels beyond the one-shot assumption, while the DRMPCL method builds on a higher-capacity contrastive training framework. These methods are therefore included for a broader context, but they are not the most directly matched comparators for the deployment setting studied here.

An extended assessment of our methodology on more recent video-based datasets, specifically LS-Vid and AG-VPReID, is presented in Table 4.2. There

TABLE 4.2: Evaluation Results on Newer Video-Based Datasets

Method	LS-Vid		Aerial→Ground		Ground→Aerial	
	mAP	Rank1	mAP	Rank1	mAP	Rank1
ShuffleNet	53.2	61.7	58.5	67.8	51.8	62.5
MobileNet	63.4	72.8	61.5	70	54.9	65.7
ResNet50	**66**	**74.4**	**61.3**	**71.3**	**55.8**	**66.3**

are no directly comparable unsupervised baselines under the same experimental assumptions, as these datasets are typically analyzed in supervised contexts in the literature. Accordingly, this table is included primarily to demonstrate how the proposed framework behaves beyond MARS and Duke and to provide additional evidence on larger, more challenging datasets. Among the tested backbones, ResNet50 provides the strongest overall results, while MobileNet remains a competitive and resource-efficient option when memory and processing power are limited. Crucially, compared to MARS and Duke, LS-Vid and AG-VPReID are bigger and more diverse, making them more difficult benchmarks and superior stress tests for robustness in real-world scenarios.

4.4.1 Analysis

The results in Tables 4.1 and 4.2 indicate that the proposed framework performs competitively given the one-shot setting studied in this chapter. On MARS and Duke, the method improves substantially over the baseline, which suggests that progressive pseudo-labelling combined with tracklet-level temporal aggregation is effective even when supervision is limited to one labelled tracklet per identity. These gains are particularly relevant in the context of this chapter, where the objective is not to match the strongest high-capacity training pipelines but to obtain a practical and resource-aware Re-ID solution under restricted supervision.

The results also suggest that the video-oriented design of the framework is important. In contrast to methods that represent a tracklet using only a

small subset of frames, the proposed formulation aggregates information at the tracklet level, making greater use of the temporal evidence available in the sequence. This likely contributes to the improvements observed compared to the baseline, particularly in settings where identity evidence is distributed across multiple frames rather than concentrated in a single observation.

Figure 4.3 shows qualitative retrieval examples on the MARS dataset. Despite its lightweight design, the proposed model often retrieves the correct identity at Rank-1. The main errors occur under large viewpoint changes and inter-camera domain shifts. These same queries are revisited in later chapters to show that a more robust Re-ID system becomes beneficial beyond this initial cost-sensitive setting.

At the same time, the comparisons in Table 4.1 should be interpreted with appropriate caution. Several recent methods, especially the strongest entries at the bottom of the table, rely on training settings that are not directly aligned with the lightweight one-shot regime considered in this chapter. Some use weak supervision beyond the one-shot assumption, whereas others benefit from more compute-intensive contrastive or transfer-based pipelines. For this reason, the most meaningful conclusion is not that the proposed method surpasses every reported approach under every condition, but that it remains competitive while operating under a more constrained, deployment-oriented setting.

The extended results on LS-Vid and AG-VPReID further clarify this point. These datasets are larger and more diverse than MARS and Duke, and they therefore provide a stronger test of robustness under realistic variations. Although no directly matched one-shot baselines are available for these benchmarks, the results show that the proposed method remains effective when moved beyond the standard datasets most commonly reported in earlier work. In particular, ResNet50 gives the strongest overall performance, while MobileNet remains attractive when efficiency is prioritized, which is

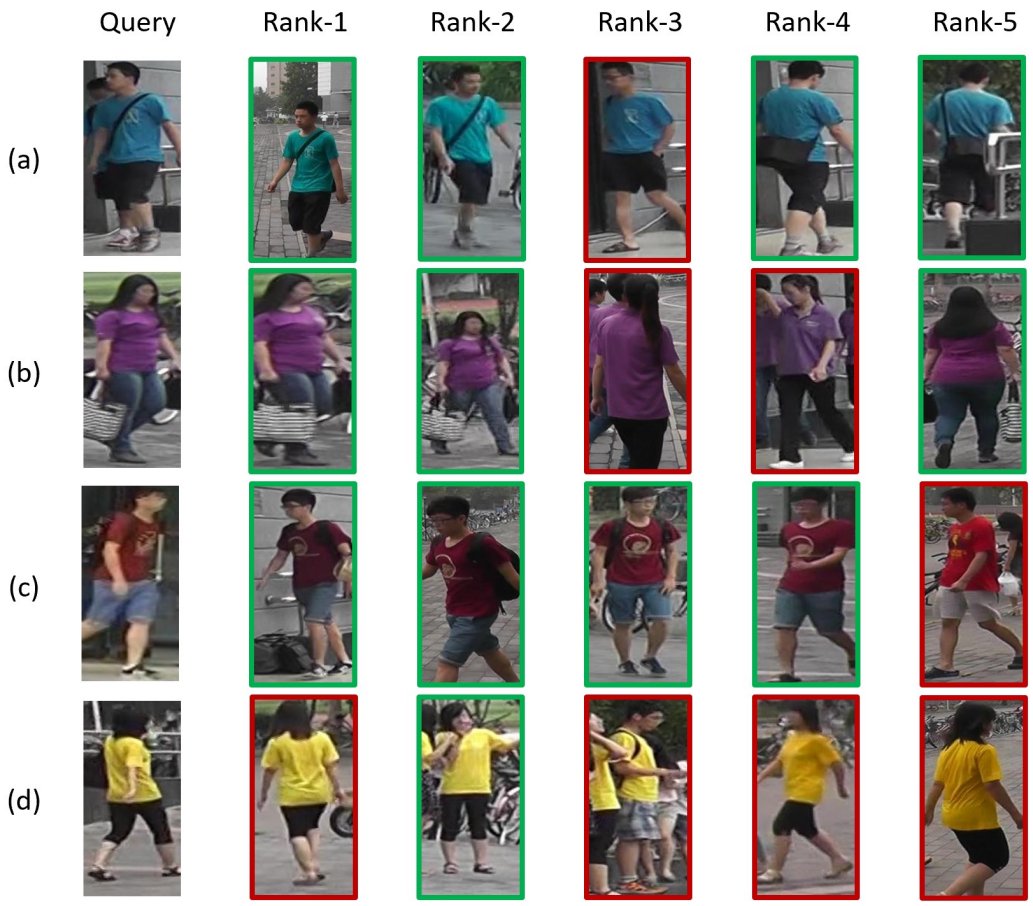


FIGURE 4.3: Qualitative retrieval examples on the MARS dataset. For each query, the top-5 retrieved tracklets are shown. Green boxes indicate true matches, and red boxes indicate false matches. Examples (a)–(c) show successful retrieval behaviour where most top-ranked results correspond to the correct identity, while example (d) illustrates a more challenging case with several false matches among the returned candidates.

consistent with the deployment objectives of this chapter.

The comparison with clustering-based pseudo-labelling is also valuable. Relative to methods such as SSLR [92], the anchor- based formulation avoids relying entirely on cluster purity during the early stages of learning. Under one-shot supervision, this is important because clustering can merge visually similar identities or split a single identity across multiple clusters. By starting from one labelled anchor per identity and expanding the labelled set conservatively, the proposed strategy provides a more stable mechanism for supervision growth in the specific setting considered in this chapter.

A similar distinction applies when comparing the proposed framework with relation-refinement methods such as RPE [133]. Their approach strengthens the embedding space through an additional relation-preserving refinement stage, whereas the present chapter focuses on a simpler pipeline that improves video-level representation learning without adding graph-based refinement modules or heavy post-processing. These two directions are therefore better understood as emphasizing different design priorities: relational refinement on the one hand and compact, deployment-oriented video Re-ID on the other.

Overall, the results support the main claim of this chapter: in the cost-sensitive one-shot regime, a conservative similarity-based expansion strategy combined with tracklet-level temporal modelling can provide a strong practical alternative to more heavily supervised or computationally demanding solutions.

4.4.2 Ablation Study

In this section, we study the contribution of the main design choices in our one-shot learning framework. Since this chapter targets resource-constrained deployment, we focus on both retrieval accuracy and practical efficiency. We first analyze the impact of backbone choice and regularization, then examine how the progressive labelling pace (controlled by the generalization percentile through δ), affects convergence and final accuracy. Finally, we compare pseudo-label assignment strategies to quantify the importance of conservative, confidence-driven label expansion.

Backbone Model and the Effect of Regularization

In this section, we investigate the impact of different backbone models and the effect of regularization on the performance of our Re-ID system. We

TABLE 4.3: Performance comparison of different backbone models on MARS and Duke datasets.

Model	MARS				Duke				Par. (M)
	mAP	Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20	
Res50	45.10	63.50	77.10	83.60	66.5	75.1	85.6	92.2	25.5
Res50v2	52.10	69.60	81.60	87.20	68.6	76.2	87.5	93.3	25.5
drop50	**55.5**	**72.2**	**84.4**	**89.6**	**72.9**	**78.6**	**88.5**	**93.7**	25.5
MobileN	53.5	71.7	84.0	89.1	68.9	76.6	87.9	93.6	4.7
Shuff. Net	51.7	67.5	79.3	85.7	68.7	76.4	87.6	93.3	**1.8**

conducted experiments using five backbone networks: three variants of the ResNet architecture (ResNet50, ResNet50v2, and ResNet50 with dropout layers), as well as two lightweight CNNs, MobileNet and ShuffleNet. The results of these experiments are summarized in Table 4.3.

The results indicate that ResNet50 with dropout layers provides the strongest performance among the tested backbones across the reported metrics on the MARS and Duke datasets. Specifically, ResNet50 with dropout layers achieves a mean Average Precision (mAP) of 55.5% on MARS and 72.9% on Duke, which are the highest among the tested models. Similarly, for Rank-1 accuracy, ResNet50 with dropout layers attains 72.2% on MARS and 78.6% on Duke, again surpassing the other models.

While ResNet50 with dropout provides the best overall performance, MobileNet emerges as a particularly important backbone for practical deployment. MobileNet achieves very similar results while being substantially more resource-efficient, reaching 53.5% mAP and 71.7% Rank-1 on MARS and 68.9% mAP with 76.6% Rank-1 on Duke, using only 4.7M parameters compared to 25.5M for the ResNet-based backbones. This efficiency gap has direct operational implications: MobileNet is well-suited for rapid initial testing and iterative prototyping, where repeated training runs and frequent configuration changes are needed but access to high-end GPUs is limited.

More importantly, it is well-suited to small-scale deployments and constrained network setups (e.g., small facilities, limited camera networks, or edge devices), where investment in expensive compute infrastructure may not be justified. In such cases, MobileNet offers a practical and cost-efficient option for deploying a Re-ID-based tracking system with competitive performance under limited hardware resources.

ResNet50v2 shows improved performance over the original ResNet50, achieving mAPs of 52.10% on MARS and 68.6% on Duke. This suggests that the architectural improvements in ResNet50v2 contribute to better feature extraction and overall performance. However, adding dropout regularization to ResNet50 with dropout layers further enhances performance, likely due to its ability to prevent overfitting and improve generalization.

In contrast, the standard ResNet50 model achieves the lowest performance among the ResNet variants, with an mAP of 45.10% on MARS and 66.5% on Duke. The results from this model reinforce the benefit of incorporating architectural refinements and regularization to enhance the ReID system's robustness and accuracy. Finally, ShuffleNet provides a compact configuration with only 1.8M parameters, achieving 51.7% mAP with 67.5% Rank-1 on MARS and 68.7% mAP with 76.4% Rank-1 on Duke, further supporting the effectiveness of lightweight backbones when model size is a key constraint.

In summary, the results indicate that ResNet50 with dropout layers provides the strongest performance among the evaluated backbones while preserving a reasonable balance between model capacity and generalization. At the same time, MobileNet provides a highly attractive, resource-efficient alternative for initial testing, rapid experimentation, and cost-sensitive deployments, enabling practical Re-ID adoption in settings where large infrastructure investments are not warranted. The results underscore the significance of selecting appropriate backbone models and incorporating regularization strategies to optimize Re-ID performance.

Generalization Percentile for Progressive Labelling

In this section, we investigate how our progressive labelling method's generalization performance is affected by the confidence saturation factor δ. The experiments indicate that setting δ to $\frac{D_\alpha}{2}$ provides the best balance between the number of iterations and the final retrieval accuracy. This configuration provides a practical compromise between computational cost, convergence speed, and retrieval performance.

Setting δ to 1 yields the highest observed accuracy. However, this gain is accompanied by substantially higher computational cost and diminishing returns relative to the extra training effort.

By comparison, the dynamic pacing strategy performs better than fixing the number of iterations at 20, while still retaining a practical training schedule. Although the fixed approach is less computationally complex, it compromises some of the accuracy that could be achieved by making adaptive adjustments to δ.

When δ is set to D_α, convergence is at its fastest, but retrieval performance is at its weakest. This configuration rapidly reaches saturation, leading to premature convergence and inadequate model refinement. The expedited completion of iterations may be attractive for conserving resources, but the trade-off in accuracy diminishes its desirability.

The results of this study are presented in Figure 4.4, where mAP, Rank-1, Rank-5, and Rank-20 performance metrics are plotted against the proportion of pseudo-labelled samples in dataset D_α. These results show that the confidence saturation factor plays an important role in balancing convergence speed, computational cost, and retrieval accuracy.

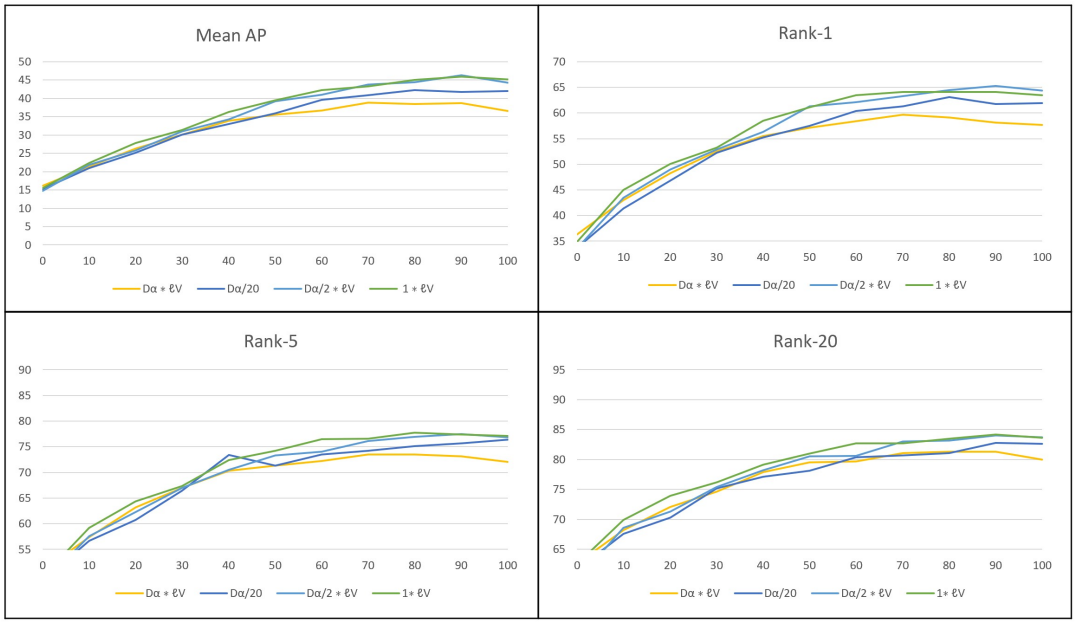


FIGURE 4.4: mAP, Rank-1, Rank-5, and Rank-20 results against the percentage of pseudo-labelled samples in D_α for the Mars Dataset. The figure illustrates the trade-offs involving convergence speed, resource consumption, and accuracy for different values of the confidence saturation factor δ. Setting δ to $\frac{D_\alpha}{2}$ provides the best balance between the number of iterations and accuracy.

Pseudo-Labels Assignments

This section evaluates the efficacy of our progressive labelling approach in comparison to other pseudo-labelling strategies. These strategies include generating all pseudo-labels using k-means clustering (Clustering@0%), performing k-means clustering after pseudo-labelling half of the dataset (Clustering@50%), and relying exclusively on progressive labelling (Prog_learning). Table 4.4 presents the findings of these experiments.

The initial strategy, clustering at 0%, entails generating all pseudo-labels by employing k-means clustering solely on the initially labelled data. This method requires only a single training iteration. Although this approach is simple and requires minimal computational resources, it achieved the lowest performance on both datasets, with mAPs of 42.00% on MARS and 52.1% on Duke. The Rank-1 accuracy was the lowest compared to the other

TABLE 4.4: Performance comparison of different pseudo-labelling strategies on MARS and Duke datasets.

Model	MARS				Duke			
	mAP	Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20
Clustering@0%	42.00	61.90	76.40	83.60	52.1	69.6	81.6	87.2
Clustering@50%	48.70	66.20	79.70	85.30	63.40	72.40	83.90	91.20
Prog_learning	55.5	72.2	84.4	89.6	68.9	76.6	88.5	93.7

tested methods, suggesting that clustering alone is insufficient to achieve high-quality pseudo-labelling in Re-ID.

The second strategy, clustering at 50%, employs k-means clustering after pseudo-labelling half of the dataset. This approach requires only half the number of training iterations when compared to the progressive labelling strategy. It demonstrates improved performance over clustering alone, achieving mAPs of 48.70% on MARS and 63.40% on Duke. The Rank-1 accuracy experienced a significant improvement, indicating that performing partial progressive labelling prior to clustering can enhance the quality of pseudo-labels.

The last approach, progressive learning, depends entirely on gradually assigning labels to the dataset without utilizing any clustering techniques. Although this strategy requires multiple iterations, it performs better than both of the clustering-based alternatives, achieving the highest mAP scores, 55.5% on MARS and 68.9% on Duke. The Rank-1 accuracy was the highest, suggesting that progressive labelling improves pseudo-label reliability in this setting.

In sum, the experiments demonstrate that progressive labelling outperforms clustering-based methods for pseudo-labelling in Re-ID. Although clustering and clustering at 50% provide quicker training times, they sacrifice accuracy. These results support the use of adaptive and iterative labelling strategies when pseudo-label reliability is a primary concern.

4.5 Conclusion

This chapter examined the one-shot, resource-aware regime of video-based person Re-ID introduced earlier in the thesis. The focus in this regime is not on maximizing performance under full supervision but on determining whether useful Re-ID behaviour can still be obtained when only one labelled tracklet per identity is available and computational resources are limited.

The results show that this regime remains viable when supervision is expanded conservatively. Beginning with a minimal set of labelled anchors and gradually using only pseudo-labelled tracklets with high confidence provides a practical way to improve the learned representation while preventing the spread of early label noise. In addition, the use of tracklet-level temporal aggregation allows the model to exploit the information available in video sequences more effectively than a purely frame-level formulation.

The experimental analysis also clarifies the main trade-off of this chapter. Higher-capacity backbones, particularly ResNet50 with dropout, provide the strongest retrieval performance, while lightweight alternatives, such as MobileNet, remain attractive when computational efficiency is a primary constraint. This distinction is important because the goal of the proposed framework is not to replace fully supervised high-capacity systems but to provide a practical option for deployments, where both annotation effort and compute cost must be controlled.

Taken together, the findings of this chapter support the broader thesis argument that video-based person Re-ID should be studied through multiple deployment regimes rather than through a single fixed solution. In the specific cost-sensitive regime considered here, a one-shot anchor-based learning strategy combined with progressive pseudo-labelling and temporal modelling provides a practically meaningful approach within the constrained deployment setting studied in this chapter.

The next chapter moves to the fully supervised regime, where the constraint on manual labelling is relaxed and the design objective shifts from efficiency under weak supervision to maximizing retrieval accuracy through stronger spatiotemporal representation learning.

Chapter 5

Supervised Video-Based Re-ID

This chapter presents our supervised video-based person Re-ID framework for the ideal deployment regime where labelled training data are available and computational resources are sufficient. Unlike the efficiency-oriented regime in Chapter 4 and the domain adaptation setting discussed in Chapter 6, this chapter isolates the highest-level performance achievable under full supervision. Together, these regimes provide a comprehensive view of the trade-offs between accuracy, efficiency, and adaptability in practical video Re-ID deployment. We build on a ViT backbone previously trained on ImageNet and introduce complementary global and local branches that collaborate to combine frame-level and tracklet-level cues. In addition, we incorporate loss fusion and ranking optimization to improve discriminability and retrieval performance across multiple video Re-ID benchmarks.

5.1 Introduction

This chapter focuses on the supervised deployment regime for video-based person Re-ID, where identity labels are available for training, and the main objective is to maximize retrieval accuracy. Unlike the one-shot setting studied in Chapter 4, the emphasis here is not on minimizing annotation requirements or reducing computational costs, but rather on learning high-quality tracklet

representations that remain discriminative under challenging appearance variations.

Video-based Re-ID offers a richer setting than image-based matching because each tracklet contains multiple observations of the same identity across time. These observations may capture complementary poses, viewpoints, and partial visibilities, which can help reduce the effects of blur, occlusion, and background clutter when the information is compiled efficiently. Therefore, in the fully supervised regime, the key challenge is not the lack of labels, but how to exploit the spatiotemporal structure of tracklets in a way that improves discriminability and retrieval robustness.

To address this problem, we adopt a Vision Transformer (ViT) backbone [16] and extend it with two complementary representation paths: (i) a tracklet-level branch that aggregates information across the full sequence using spatiotemporal attention, and (ii) a frame-level local branch that emphasizes fine-grained cues through temporal feature shifting and localized salient-region modelling. In addition, we optimize training using a fusion of identification- and metric-learning objectives, and we improve retrieval at inference time through a re-ranking procedure.

A broader review of supervised video-based person Re-ID and related metric-learning strategies is provided in Chapter 2. In this chapter, we focus only on the design decisions that are most relevant to the supervised setting: stronger spatiotemporal aggregation, more effective local discriminative modelling, and loss optimization to maximize retrieval performance.

The main contributions of this chapter may be summarized as follows:

- We propose an accuracy-oriented supervised video Re-ID framework built on a ViT backbone that explicitly leverages tracklet information rather than treating the task as a set of independent image matches.

- We introduce a dual-branch design that combines (i) tracklet-level spatiotemporal aggregation through a dedicated attention-based module (SAN) and (ii) local discriminative modelling through temporal feature shifts (TFS) and localized salience feature patches (LSFP).
- We replace dense uniform region tokenization with a low-token-count, semantically grounded part-based representation derived from the human body parsing masks, improving robustness to pose variation and background noise while remaining tractable for multi-frame processing.
- We adopt loss fusion (label-smoothed identity loss, hard-mined triplet loss, and centre loss) to improve embedding discriminability, and we apply a ranking optimization (re-ranking) stage that consistently boosts retrieval metrics on larger-scale benchmarks.
- We provide extensive experimental evaluation across standard video Re-ID datasets and a cross-platform setting, and we quantify the effect of each module through ablation studies.

The remainder of this chapter is organized as follows. Section 5.2.1–Section 5.2.3 presents the proposed architecture and feature extraction modules, followed by the loss design and ranking optimization. Section 5.3.2 reports quantitative comparisons on multiple benchmarks and discusses the impact of re-ranking, and Section 5.3.4 provides an ablation study to isolate the contribution of SAN, TFS, and LSFP. Finally, we conclude with a summary of findings and practical considerations for supervised deployment.

5.2 Methodology

This section presents the design of our supervised video Re-ID framework. While recent transformer-based models have demonstrated strong performance in image-based Re-ID, directly applying them to video data does not

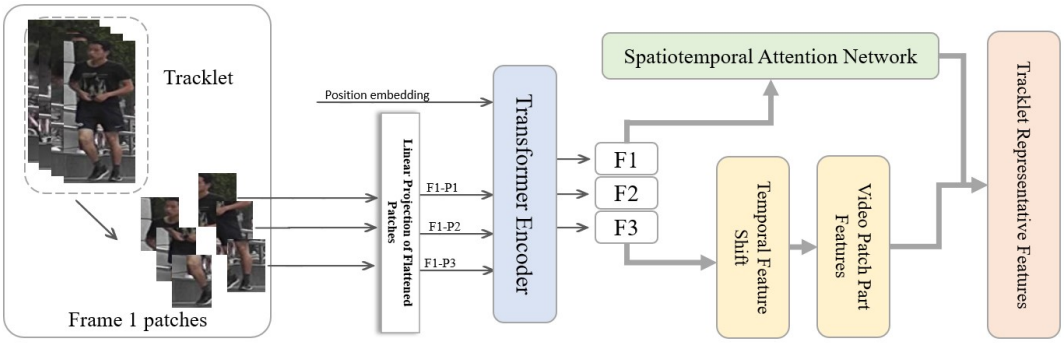


FIGURE 5.1: Our model overview. Each frame is split into patches, and position embeddings are added before feeding them to our transformer model. We then send the generated frame feature vectors to both the SAN and TFS modules to obtain a tracklet-representative feature.

fully exploit the temporal structure of tracklets. In particular, standard ViT architectures primarily model spatial relationships within individual frames and lack mechanisms for effective temporal aggregation and fine-grained local discrimination across sequences.

To address these limitations, we extend a ViT backbone with two complementary feature extraction branches: a global branch that aggregates tracklet-level spatiotemporal context and a local branch that enhances fine-grained discriminative features across frames. In addition, we design a fused optimization strategy and a re-ranking mechanism to further improve retrieval performance.

5.2.1 Feature Extraction

As illustrated in Figure 5.1, the proposed framework comprises a shared transformer backbone, followed by complementary global and local branches.

A key challenge in video-based Re-ID is how to represent a tracklet, ensuring that both global temporal context and fine-grained appearance cues are preserved. Relying solely on frame-level representations can lead to

unstable matching due to occlusion or blur, while naive temporal aggregation may dilute discriminative details.

To address this, we adopt a dual-level representation strategy based on a shared ViT backbone. Each frame is first converted into a sequence of patch tokens, which are then processed by a transformer encoder to produce frame-level features. These features are then propagated through two complementary branches: a global branch that aggregates spatiotemporal information across the tracklet and a local branch that enhances fine-grained discriminative cues. Finally, the outputs of both branches are combined to form a unified tracklet representation used for training and retrieval.

Formally, for each frame $I \in \mathbb{R}^{H \times W \times C}$, where H, W, and C denote the frame height, width, and channels, respectively, the input is divided into n equal-sized patches. We calculate n as follows:

$$n = \left[\frac{H-P}{S}+1\right] \times \left[\frac{W-P}{S}+1\right] \tag{5.1}$$

where P and S are the convolutional operation's patch and stride sizes. We use a stride equal to the patch size ($S = P$) to ensure non-overlapping patches, and no padding is applied. Each patch $I_{p(n)}$ is then flattened and mapped to D dimensions with a linear projection mapping function $\mathcal{F}$. Lastly, both the learnable class embedding I_{class} and the learnable position embedding pos are prepended to the frame vector as follows:

$$I = \left[I_{class}; \mathcal{F}\left(I_{p(1)}\right); \mathcal{F}\left(I_{p(2)}\right); \cdots; \mathcal{F}\left(I_{p(n)}\right)\right] + pos, \tag{5.2}$$

where $pos \in \mathbb{R}^{(n+1)\times D}$ and the frame vector I is used as input for the encoder. The transformer encoder consists of multiple blocks. For the i-th block, the input token sequence is first normalized and passed through the multi-head self-attention (MSA) layer with a residual connection:

$$\mathbf{X}'_i = \mathbf{X}_{i-1} + MSA\left(LN\left(\mathbf{X}_{i-1}\right)\right) \tag{5.3}$$

The resulting sequence is then normalized and passed through the multi-layer perceptron (MLP), again with a residual connection:

$$\mathbf{O}_i = \mathbf{X}'_i + MLP\left(LN\left(\mathbf{X}'_i\right)\right) \tag{5.4}$$

where $\mathbf{X}_{i-1}$ is the input to the i-th transformer block, $\mathbf{X}'_i$ is the intermediate output after self-attention, and $\mathbf{O}_i$ is the output of the block. Since the output dimension of each block is the same as the input dimension, $\mathbf{O}_i$ can be passed directly to the next transformer block, as discussed further in Section 5.2.3.

The encoder outputs are then processed by the two complementary branches of our framework. In the global branch, frame-level features are aggregated via a spatiotemporal attention network to produce a tracklet-level descriptor. In parallel, the local branch applies temporal feature shift and localized salience feature patches to improve fine-grained discriminative modelling. Finally, the outputs of both branches are combined to form the tracklet's representation feature and used for subsequent loss optimization and retrieval stages.

5.2.2 Mini-Global Model

A major challenge in tracklet-level modelling is the trade-off between capturing long-range temporal dependencies and maintaining computational efficiency. Processing all frames within a tracklet can be computationally prohibitive and may introduce redundancy, as consecutive frames often contain highly correlated information.

To balance these factors, we design a global branch that selectively aggregates temporal information while controlling redundancy through frame normalization.

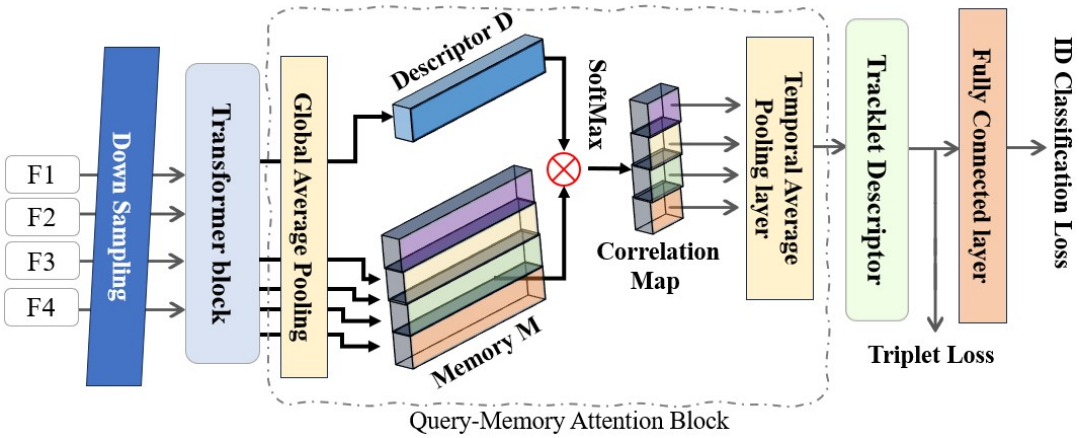


FIGURE 5.2: An illustration of our Spatiotemporal Attention Network. Here, the query-memory attention block is connected to the transformer block to generate the attention maps for each frame of the tracklet.

The mini-global branch is designed to learn tracklet-level representations that capture spatiotemporal dependencies across the sequence. This branch, unlike the local branch, processes all frames within a tracklet rather than a subset of sampled keyframes, generating more comprehensive video-level features that capture spatial-temporal information across the tracklet. This approach is computationally expensive and subject to diminishing returns [14]. Downsampling reduces redundancy across highly correlated frames and mitigates the impact of low-quality or noisy observations within the tracklet. To control this variability, tracklets exceeding the average length are subsampled by removing non-informative frames, while shorter tracklets are augmented via frame duplication. This normalization reduces temporal imbalance and stabilizes training.

However, simple temporal pooling is insufficient to capture complex inter-frame dependencies, particularly when frames vary significantly in quality or relevance. Therefore, we introduce a spatiotemporal attention mechanism to selectively emphasize informative frames. As illustrated in Figure 5.2, the SAN module computes attention weights across frames based on their semantic relevance.

Our network starts by feeding the tracklets to the transformer encoder as frames. In the global branch, all frames that belong to the same tracklet will be aggregated and downsampled to become an input to our spatiotemporal attention network (SAN), inspired by [138], [164], [193]. This network consists of a query-memory attention block, where each frame in the tracklet is squeezed into a descriptor $\mathcal{D}$ using global average pooling (GAP), and the memory $\mathcal{M}$ consists of the descriptors of the remaining frames in the tracklet. By calculating the correlation map of each descriptor with the normalized cross correlation (NCC), we compute the semantic relevance between $\mathcal{D}$ and all the other descriptors in $\mathcal{M}$. The final attention mask is calculated as a softmax [165] layer applied to the correlation map. This block is followed by a temporal average pooling (TAP) layer that generates a temporal attention score, thereby creating a tracklet-level descriptor. A fully connected layer is used at the end to predict the probability of the person's identity.

5.2.3 Max-Local Model

While global aggregation captures overall temporal consistency, it may overlook fine-grained local cues that are critical for distinguishing visually similar identities. In particular, subtle differences in clothing texture, accessories, or body parts may be diluted during global pooling.

To address this limitation, we introduce a local branch that enhances discriminative features at the frame level while still incorporating temporal information. In this branch, there is no downsampling. However, we do not process every frame; instead, we randomly choose some keyframes.

To achieve this, we implemented two blocks: the temporal feature shift (TFS) and localized salience feature patches (LSFP).

Since we still want to preserve temporal information in this branch, we

developed our TFS module, which is inspired by shuffle networks for convolutional neural networks (CNNs) [194], [195], [196]. The original premise of shuffle networks is to introduce temporal vision into a conventional 2D convolutional network without requiring kernel expansion. Thus, our proposed block is a parameter-free module that reallocates the transformer decoder's output to capture the temporal dimension.

Temporal Feature Shift

Standard Vision Transformers are inherently designed for spatial modelling and do not explicitly encode temporal relationships across frames. As a result, they may fail to capture motion patterns and temporal dependencies that are crucial in video-based Re-ID.

To introduce temporal awareness without modifying the transformer architecture, we propose a TFS module that reallocates features to incorporate a temporal receptive field. As illustrated in Figure 5.3, the TFS module rearranges channel groups across frames to incorporate temporal context without modifying the transformer architecture.

Since the transformer architecture remains unchanged, the input representation must preserve both the original spatial structure and temporal information across frames. Therefore, the transformed features maintain the same dimensionality as the original input while incorporating temporal context through feature reallocation. Given that each tracklet will have T frames, a tensor [C, H, W] of H and W spatial dimensions and a channel size of C will represent each frame. We divided each channel into T groups, each with a C/T channel size. Each set of newly grouped features $\mathcal{G}$ will have the shape of [C/T, H, W] and contain part of the frame spatial features:

$$f_i = \left[\mathcal{G}_{i,1}^{C/T}, \mathcal{G}_{i,2}^{C/T}, ..., \mathcal{G}_{i,T}^{C/T}\right] \tag{5.5}$$

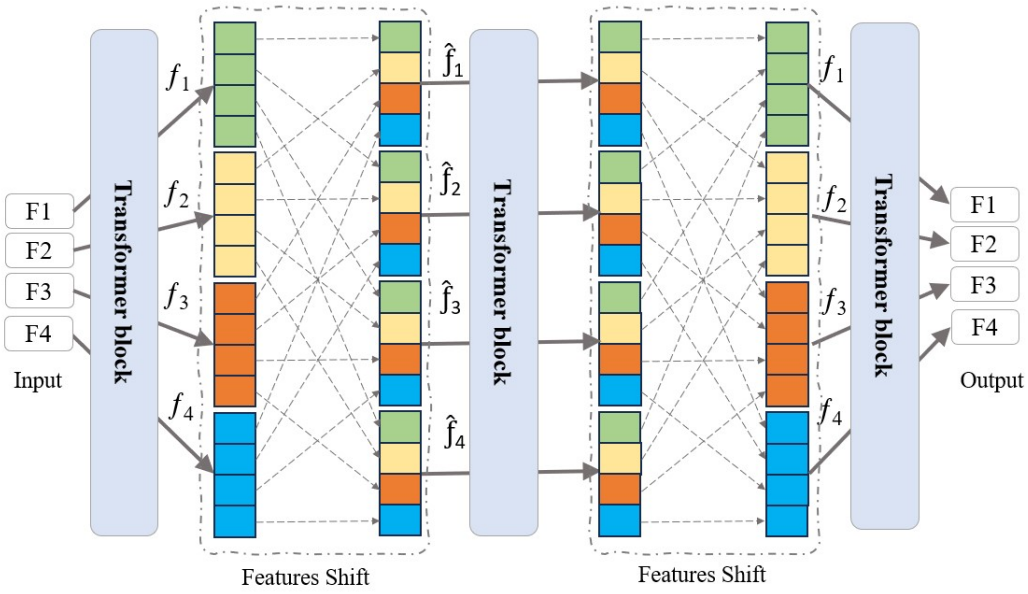


FIGURE 5.3: Graphical representation of our temporal feature shift network. Here, we aggregate the frames' feature tensors into temporal feature tensors $\hat{f}_i$, which contain the spatial information of all sequential frames, thus providing non-local perception.

denoting each group of index j at the i-th frame as $\mathcal{G}_{i,j}$. Here, i indexes the frame and j indexes the channel group, such that each transformed feature aggregates information across frames at a fixed channel position. TFS transforms the original feature tensor of the i-th frame f_i to a new feature tensor $\hat{f}_i$ using the following equation:

$$\hat{f}_i = \left[\mathcal{G}_{1,j}^{C/T}, \mathcal{G}_{2,j}^{C/T}, ..., \mathcal{G}_{T,j}^{C/T}\right] \tag{5.6}$$

where $1 < i < T$, and j will be the group at the i-th index of the slicing operation along the channel dimension. For instance, $\mathcal{G}_{1,1}$ indicates the first feature group from the first frame tensor. As a result, the new feature tensor $\hat{f}_i$ contains the spatial information of all sequential frames and serves as an input to the following transformer block. Furthermore, we argue that—unlike He et al.'s [196] and Alsehaim and Breckon's [125] implementations, which only shift features by S steps between neighbour frames, weakly representing

temporal information—our implementation takes advantage of the non-local details in the tracklet. Our experiments demonstrate the effectiveness of our approach to feature shifting in section 5.3.4.

Localized Salience Feature Patches

Existing Re-ID models often utilize horizontal stripes to extract localized features [41], [125], [197], [198]. The horizontal-striping approach divides the human image into several fixed horizontal bands (stripes) and pools the features within each band to obtain local descriptors. This design is simple and effective for many image-based Re-ID tasks because it provides alignment of body regions (e.g., head/upper-body/lower-body) with relatively low computational costs. While horizontal partitioning is simple and effective in many image-based settings, it can be compromised when body part locations shift due to pose, viewpoint, or occlusion. Thus, Lee et al. [174] propose REET (region-enhanced tokenization), which generates many region tokens via dense horizontal and vertical slicing to enrich local representations. While REET demonstrates improvements in image-based settings by increasing region coverage and diversity, it has two practical limitations for video/tracklet Re-ID. First, REET produces a large number of regions (e.g., 64 regions in their example), which increases the computational cost of the multi-head self-attention because its complexity grows quadratically with the number of tokens [141]; this becomes prohibitive when processing multiple frames per tracklet. Second, REET is not specifically designed to address intra-tracklet temporal variations (pose changes, occlusions, and vertical misalignment across frames), which are central challenges in video Re-ID. Motivated by REET's insight (region-enhanced tokenization) but mindful of these limitations, we propose a semantically grounded, low-token-count alternative better suited for video tracklets. To overcome the limitations of fixed horizontal stripes and dense uniform slicing for video Re-ID, we replace horizontal

striping with human-body-parsing masks. Specifically, our parser produces four semantic masks per frame: head, torso, legs, and hands. These masks yield semantically coherent regions that are robust to pose and viewpoint variations, reduce background noise, and keep the number of tokens low for efficient temporal aggregation. The parser incorporates boundary-aware, hybrid-resolution, and edge-guided strategies to improve part localization and small-part delineation [199].

Let $\mathcal{M} = \{\text{head}, \text{torso}, \text{legs}, \text{hands}\}$ denote the set of semantic part masks, and let $P = |\mathcal{M}| = 4$. For a tracklet t with T frames, and for part $p \in \mathcal{M}$, let $\mathbf{S}_{p,j} \in \mathbb{R}^d$ denote the pooled feature vector extracted from frame j restricted to part p. The vector $\mathbf{S}_{p,j}$ is obtained by applying mask p to the feature map of frame j (pixel-wise multiplication of the mask and feature map), followed by spatial pooling. We arrange the per-part, per-frame features into a tensor:

$$\mathbf{F}_t^{\text{parts}} = \begin{bmatrix} \mathbf{S}_{\text{head},1} & \cdots & \mathbf{S}_{\text{head},T} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{\text{hands},1} & \cdots & \mathbf{S}_{\text{hands},T} \end{bmatrix} \in \mathbb{R}^{P \times T \times d} \tag{5.7}$$

To obtain a robust per-part descriptor across the tracklet, we perform part-based temporal aggregation with area normalization to compensate for variable mask sizes and partial occlusions:

$$\mathbf{H}_p = \frac{1}{T} \sum_{j=1}^{T} \frac{\mathbf{S}_{p,j}}{Z_{p,j}}, \qquad p \in \mathcal{M}, \tag{5.8}$$

where $Z_{p,j}$ is a normalization scalar proportional to the mask area in frame j. For binary masks, $Z_{p,j} = |\text{mask}_{p,j}|$ (number of mask pixels); and for soft/probability masks, $Z_{p,j} = \sum \text{mask}_{p,j}$. This normalization encourages consistent magnitudes across parts and frames, reducing the influence of small/partial masks (e.g., when a part is heavily occluded).

This formulation provides inherent robustness to partial occlusion. When

a body part is occluded in a given frame, the corresponding mask area $Z_{p,j}$ becomes small, reducing its contribution to the aggregated feature. At the same time, temporal aggregation across frames ensures that information from visible instances of the same part is preserved. Moreover, since the representation is constructed from multiple semantic parts, occlusion affecting one region (e.g., hands or legs) does not eliminate the contribution of other visible parts (e.g., torso or head), allowing the model to maintain discriminative identity cues despite partial visibility.

After computing the per-part descriptors $\mathbf{H}_p$, the tracklet descriptor is formed by concatenation, followed by an optional projection and normalization:

$$\mathbf{D}_t = \mathrm{Norm}\big(\mathbf{W}_{\text{proj}} \left[\mathbf{H}_{\text{head}} \,\|\, \mathbf{H}_{\text{torso}} \,\|\, \mathbf{H}_{\text{legs}} \,\|\, \mathbf{H}_{\text{hands}}\right]\big)\,, \tag{5.9}$$

where $\|$ denotes concatenation, $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d' \times 4d}$ is an optional learnable linear projection, and $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization applied to the final descriptor. As illustrated in Figure 5.4, the proposed LSFP approach replaces conventional uniform patching strategies (e.g., horizontal or grid-based partitions) with semantically aligned body-part masks. This design allows the model to focus on discriminative regions while suppressing background noise, which is particularly beneficial under occlusion and viewpoint variation.

5.2.4 Loss Optimization

In supervised Re-ID, relying solely on classification loss may lead to suboptimal embedding structures, while metric learning alone may struggle from unstable convergence. To address this, we combine both objectives in our fused loss as follows:

$$\ell_{\mathcal{F}} = \alpha \cdot \ell_{_{ls}Id} + \beta \cdot \ell_{HTri} + \gamma \cdot \ell_{center} \tag{5.10}$$

FIGURE 5.4: Examples of features patching. A) VPPF [125]: Horizontal stripes-based features patching with strip size of 4. (B) LSFP (ours): human body parsing masks with self-attention and background suppression. (C) REET [174]: Regional stripes-based with 8×8 strip size.

where α, β, and γ are parameters to control the weight of each one of the three loss functions. Where $\ell_{_{ls}Id}$ is the label smoothing classification loss function, adapted from [200], and ℓ_{HTri} is the hard-mined triplet loss function introduced in [81]. We also adopted both the centre loss function used in [201] and their inference stage between the feature embedding and the fully connected layer to improve loss convergence.

5.2.5 Ranking Optimization

Although the learned embeddings provide strong initial retrieval performance, the ranking process may still fail to recover hard positive samples due to subtle appearance variations, occlusions, or viewpoint changes. In such cases, relevant identities may not appear in the top-ranked results, despite being close in the embedding space.

To address this limitation, we apply a re-ranking strategy that exploits the local neighbourhood structure of the feature space. By incorporating

reciprocal nearest neighbours, the re-ranking process refines the similarity relationships between query and gallery samples, allowing hard positive matches to be recovered and improving the overall ranking quality.

The rank-κ list of candidates for a query image χ is defined as :

$$\mathcal{K}(\chi, \kappa) = \{g_1, ..., g_\kappa\}\,, \text{ where } \|\mathcal{K}(\chi, \kappa)\| = \kappa \tag{5.11}$$

Inspired by [25], [202], the re-ranked list $\mathcal{K}_\mathcal{R}$ is then defined as:

$$\begin{aligned} &\mathcal{R}(\chi, \kappa) \leftarrow \mathcal{K}(\chi, \kappa) \cup \mathcal{K}(\varrho, \frac{1}{2}\kappa) \\ &\text{s.t.} \left\| \mathcal{K}(\chi, \kappa) \cap \mathcal{K}(\varrho, \frac{1}{2}\kappa) \right\| \geq \frac{2}{3} \left\| \mathcal{K}(\varrho, \frac{1}{2}\kappa) \right\| \\ &\forall \varrho \in \mathcal{K}(\chi, \kappa) \end{aligned} \tag{5.12}$$

The resulting list will have more positive samples related to the candidates than the query image itself. We can express those samples as hard positive candidates not included in the original list due to variations in occlusion, pose, or illumination. Next, we re-calculate the distance between the query image χ and any gallery image g_i as the overlap percentage between their $\mathcal{R}$ sets:

$$\mathcal{D}\left(\chi, g_i\right) = 1 - \frac{\|\mathcal{R}(\chi, \kappa) \cap \mathcal{R}(g_i, \kappa)\|}{\|\mathcal{R}(\chi, \kappa) \cup \mathcal{R}(g_i, \kappa)\|} \tag{5.13}$$

This formulation effectively captures contextual similarity by considering shared neighbourhood structure rather than relying solely on pairwise distances. As a result, it improves robustness to local ambiguities in the embedding space.

We can then define the K re-ranked list $\mathcal{K}_\mathcal{R}$ as a list of κ nearest neighbours' gallery images $G = \{g_1, ..., g_\kappa\}$ from the query χ, where $\mathcal{D}\left(\chi, g_i\right) < \mathcal{D}\left(\chi, g_{i+1}\right)$. Our experimental results demonstrate that re-ranking raises overall accuracy when the dataset is large. However, we can observe that it

negatively affects the results when dealing with small datasets. This degradation is likely due to unreliable neighbourhood structure in smaller datasets, where limited samples reduce the effectiveness of reciprocal nearest-neighbour relationships.

5.2.6 Framework

We adopt the Bag-of-Tricks [201] as a strong and well-established reference framework and extend it with a transformer-based backbone [16] and the proposed strategies described in [125], [135], [139] to better capture spatiotemporal dependencies.

Algorithm 2 Supervised video Re-ID

Input: Training set D_T , evaluation set D_E , pre-trained ViT model f_θ
Output: best ViT model f_θ

Do random erasing augmentation [203] on D_T
Remove the last stride [41] from the pre-trained ViT model f_θ
Add attention layers to the pre-trained ViT model f_θ
Add TFS and LSFP to the pre-trained ViT model f_θ
Split D_T into training set D_{train} and validation set D_{val}
while not converged **do**
    Train the model using a batch from D_{train}
    Update ℓ_{HTri} via the fusion loss $\ell_{\mathcal{F}}$ in Eq. 5.10.
    Validate the model using a batch from D_{val}
    Update $\ell_{_{ls}Id}$ via the fusion loss $\ell_{\mathcal{F}}$ in Eq. 5.10.
    Update ℓ_{center} via the fusion loss $\ell_{\mathcal{F}}$ in Eq. 5.10.
Evaluate the model using D_E
Apply re-ranking using Eq. 5.13.
Evaluate the model again using D_E after re-ranking

5.3 Experiments

We follow a similar setting to the baseline and most state-of-the-art works to facilitate comparison. A ViT pre-trained on ImageNet is used as a backbone. Before training, each frame is flipped horizontally and then applied to random

erasing [203]. We kick-started network performance using the warm-up strategy [77] and optimized our model by setting the gradient descent momentum to 0.9. We resized all frames to 256×128 and set the batch size to 8, with each tracklet consisting of four randomly selected frames.

5.3.1 Experimental Settings

We evaluate the proposed supervised video Re-ID framework using the standard retrieval metrics defined in Chapter 3. In particular, we report CMC at Rank-1 together with mean average precision (mAP), and include mean inverse negative penalty (mINP) to better reflect retrieval quality for harder matches. Since evaluation is performed at the tracklet level, each query corresponds to a query tracklet, and the gallery contains tracklets defined by the corresponding dataset protocol. In this chapter, these metrics are used to assess both top-rank retrieval accuracy and the overall quality of the ranked retrieval list.

5.3.2 Results and Comparison

As shown in Table 5.1, the proposed model achieves competitive performance across all three datasets prior to re-ranking (MARS: mINP 58.5%, mAP 84.5%, Rank-1 94.3%; LS-Vid: mINP 61.8%, mAP 83.8%, Rank-1 90.2%; iLIDS-VID: mINP 73.5%, mAP 88.6%, Rank-1 98.6%). After applying re-ranking, consistent improvements are observed across all benchmarks (MARS: mINP 79.5%, mAP 91.6%, Rank-1 96.3%; LS-Vid: mINP 71.4%, mAP 85.3%, Rank-1 93.6%; iLIDS-VID: mINP 77.9%, mAP 89.2%, Rank-1 98.6%).

On MARS, the re-ranked model achieves performance comparable to or exceeding several transformer-based approaches on Rank-1 accuracy, while remaining competitive on mAP. On iLIDS-VID, the model achieves Rank-1 accuracy on par with the strongest reported results. On LS-Vid, while some

TABLE 5.1: Evaluation results on MARS, LS-Vid, and iLIDS-VID using mINP, mAP, and Rank-1. The best result in each metric is shown in bold.

Method	MARS			LS-Vid			iLIDS-VID		
	mINP	mAP	R1	mINP	mAP	R1	mINP	mAP	R1
BagOfTricks [201]	62.0	81.6	85.8	–	–	–	82.2	–	74.0
CoSeg [138]	57.8	79.9	84.9	–	–	–	–	–	–
AGW [140]	63.9	83.0	87.6	–	–	–	**89.0**	–	83.2
BiCnet [139]	–	86.0	90.2	–	75.1	84.6	–	90.4	98.3
PiT [147]	–	86.8	90.2	–	–	–	–	–	96.5
ViT [174]	–	90.7	78.6	–	–	–	–	–	74.2
VID-Trans [125]	–	86.4	94.7	–	–	–	–	–	96.6
TMT [148]	–	85.8	91.2	–	–	–	–	91.3	**98.6**
DCCT [150]	–	87.5	92.3	–	–	–	–	91.7	98.4
LSTRL [149]	–	86.8	91.6	–	82.4	89.8	–	92.2	**98.6**
CLIP-ReID [155]	–	88.1	91.7	–	80.6	88.8	–	–	–
TM [126]	–	89.9	95.9	–	–	–	–	–	–
AG [152]	–	91.5	93.2	–	**87.3**	93.2	–	**96.3**	98.5
Ours	58.5	84.5	94.3	61.8	83.8	90.2	73.5	88.6	**98.6**
Ours + Re-ranking	**79.5**	**91.6**	**96.3**	**71.4**	85.3	**93.6**	77.9	89.2	**98.6**

methods report higher baseline mAP, the proposed approach demonstrates clear improvements after re-ranking, particularly in Rank-1 performance.

On the cross-platform AG-VPReID benchmark (Table 5.2), the proposed method achieves competitive cross-domain retrieval performance, with improvements observed after re-ranking. In particular, the model yields strong results in both Aerial→Ground and Ground→Aerial settings, indicating robust generalization across viewpoints.

Overall, these results show that the proposed framework performs consistently at a level comparable to state-of-the-art methods, with re-ranking providing a significant contribution to improving retrieval accuracy across different datasets.

To further illustrate the retrieval behaviour of the proposed model, Figure 5.5 presents qualitative examples of top-ranked results. In most cases (Figure 5.5(a–c)), the correct identity is consistently retrieved within the top ranks despite variations in pose, viewpoint, and background clutter. This indicates that the learned representations are robust to intra-tracklet variations

TABLE 5.2: Evaluation results on video-based dataset AG-VPReID using mAP, and CMC.

Method	Aerial→Ground		Ground→Aerial	
	mAP	Rank-1	mAP	Rank-1
BiCnet [139]	59.8	69.2	54.3	64.7
TMT [148]	60.8	70.5	55.9	65.8
DCCT [150]	61.5	71.2	56.6	66.4
LSTRL[149]	61.7	71.3	56.7	66.5
CLIP-ReID [155]	62.3	71.6	57.2	66.8
TM [126]	54.7	63.9	55.3	68.3
AG [152]	64.0	71.9	58.0	75.6
Ours	65.2	**73.6**	56.4	74.2
Ours Reranked	**65.7**	72.2	**58.6**	**76.1**

and effectively capture discriminative identity cues.

However, faulty cases are also observed, as shown in Figure 5.5(d), where the query originates from Camera 6, which exhibits a larger domain shift compared to other cameras. This leads to an incorrect match at lower ranks, highlighting the impact of cross-camera variation on retrieval performance. This behaviour is consistent with the limitations discussed in Section 4.4.1, where domain shift can degrade feature consistency across views. Compared to the one-shot setting, the supervised framework reduces but does not completely eliminate such effects, particularly in scenarios with significant cross-camera discrepancies.

5.3.3 Analysis

Recent transformer-based and hybrid methods (e.g., TMT, DCCT, VID-Trans, and TM) aim to improve video Re-ID performance through enhanced spatiotemporal modelling. These approaches typically focus on architectural modifications to better capture temporal dependencies and spatial context across frames.

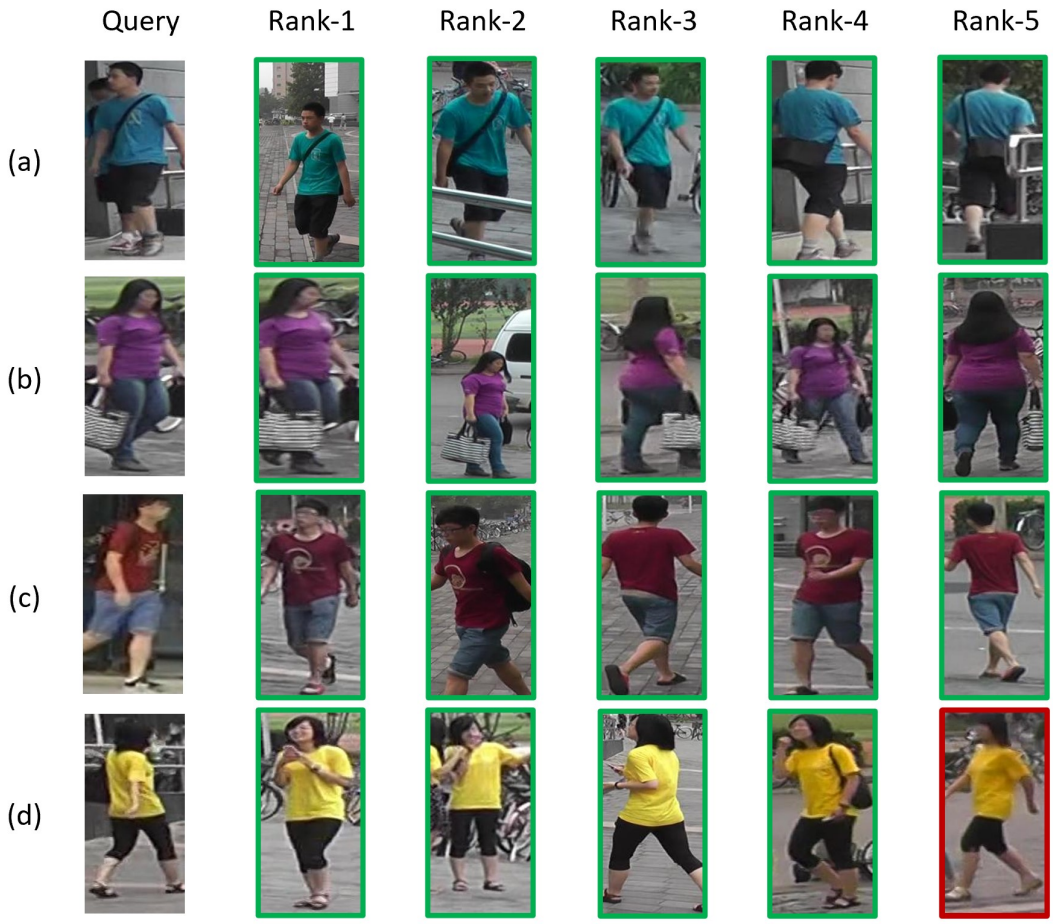


FIGURE 5.5: Qualitative retrieval examples on the MARS dataset. For each query (left), the top-5 retrieved tracklets are shown. Green boxes indicate correct matches, while red boxes indicate incorrect matches. Case (d) illustrates a failure due to cross-camera domain shift (Camera 6).

In comparison, our results show that the proposed framework achieves competitive performance across multiple benchmarks, with particularly strong improvements after re-ranking. On the MARS dataset, the re-ranked pipeline achieves Rank-1 accuracy comparable to or exceeding several transformer-based methods, indicating that the learned embeddings are highly discriminative at the identity level. This suggests that the combination of global spatiotemporal aggregation (SAN) and local discriminative modelling (TFS and LSFP) effectively captures both coarse and fine-grained identity cues.

Furthermore, methods such as DCCT emphasize complementary feature learning via hybrid architectures, whereas our approach achieves similar performance with a unified transformer-based design. This indicates that explicitly separating global and local representation pathways can provide comparable benefits without introducing additional architectural complexity.

On the other hand, approaches that rely heavily on large-scale pretraining (e.g., CLIP-ReID and AG) demonstrate strong baseline mAP performance by leveraging external data and multimodal supervision. While such methods achieve higher performance in certain metrics, our results show that, after re-ranking, the proposed model attains comparable Rank-1 accuracy on several benchmarks. This highlights the effectiveness of our architecture in learning discriminative identity features directly within the supervised video Re-ID setting, without relying on additional pretraining modalities.

Overall, these results suggest that performance gains in supervised video Re-ID are not solely dependent on larger models or external data but can also be achieved through a balanced design that integrates global temporal context with fine-grained local discrimination. The consistent improvements observed after re-ranking further indicate that the learned feature space preserves meaningful neighbourhood structure, enabling more accurate retrieval when contextual relationships are exploited.

TABLE 5.3: Ablation study on Mars and LS-Vid

Method	Mars		LS-Vid	
	Rank-1	mAP	Rank-1	mAP
ours without TFS	93.2	86.0	86.5	78.1
ours without LSFP	92.5	88.4	87.1	77.2
ours without SAN	90.4	85.2	81.4	74.6
ours with TCSS	95.3	89.1	88.1	82.4
ours with NLA-CNN	91.4	85.7	82.1	76.4
ours Full implementation	96.3	91.6	93.6	85.3

5.3.4 Ablation Study

We conduct an ablation study to quantify the contribution of each component in our framework on Mars and LS-Vid (Table 5.3). Our full implementation achieves 96.3% Rank-1 and 91.6% mAP on Mars, and 93.6% Rank-1 and 85.3% mAP on LS-Vid.

TFS and LSFP: We first evaluate the impact of temporal feature shift (TFS) and localized salience feature patches (LSFP). Removing **TFS** decreases performance from 96.3/91.6 to 93.2/86.0 on Mars (3.1 Rank-1, 5.6 mAP) and from 93.6/85.3 to 86.5/78.1 on LS-Vid (7.1 Rank-1, 7.2 mAP). This indicates that explicitly shifting and aligning temporal features improves robustness to motion and frame-to-frame appearance variations. Similarly, removing **LSFP** reduces accuracy to 92.5% Rank-1 and 88.4% mAP on Mars (3.8 Rank-1, 3.2 mAP) and to 87.1% Rank-1 and 77.2% mAP on LS-Vid (6.5 Rank-1, 8.1 mAP). These results suggest that localized salient regions provide complementary discriminative cues that are not fully captured by global/tracklet-level aggregation alone.

LSFP parser: We further evaluate the sensitivity of LSFP to the boundary-penalty weight δ used in the body-part parsing stage. To analyze the impact

of the body-part parsing parameters, Figure 5.6 reports downstream mAP and Rank-1 while varying δ for each semantic part. We observe a clear part-dependent optimum: head and hands achieve their best accuracy at $\delta = 1.5$, whereas torso and legs peak at $\delta = 1.0$. We attribute this difference to scale. Head and hands occupy smaller regions and often have weaker contrast and more ambiguous boundaries; consequently, boundary blur or mask leakage can affect a relatively large portion of the pooled features, making sharper boundaries (larger δ) beneficial. In contrast, torso and legs cover larger areas and are less sensitive to minor boundary smoothing, while overly strong boundary penalties may over-regularize the masks near part transitions and slightly reduce discriminative content. Based on these results, we adopt a mixed setting in our main experiments to favour the more boundary-sensitive small parts while maintaining competitive performance on larger regions.

SAN: Our spatiotemporal attention network (SAN) includes a dedicated tracklet-level branch that aggregates information across the entire sequence, complementing the frame-level/local representations. Disabling **SAN** causes the largest drop in both datasets: on Mars, performance decreases to 90.4/85.2 (5.9 Rank-1, 6.4 mAP), and on LS-Vid it decreases to 81.4/74.6 (12.2 Rank-1, 10.7 mAP). This confirms that spatiotemporal modelling at the tracklet level is essential for video-based Re-ID.

Replacing sub-modules with alternative designs: To assess whether our gains come from the specific design of each component, we substitute TFS with TCSS [125] and replace SAN with an NLA-CNN variant following [140]. Using TCSS yields 95.3/89.1 on Mars and 88.1/82.4 on LS-Vid, which are consistently lower than the full model. Replacing SAN with NLA-CNN further reduces performance to 91.4/85.7 on Mars and 82.1/76.4 on LS-Vid. Overall, Table 5.3 shows that each proposed component contributes complementary improvements, and their combination produces the best performance across both benchmarks.

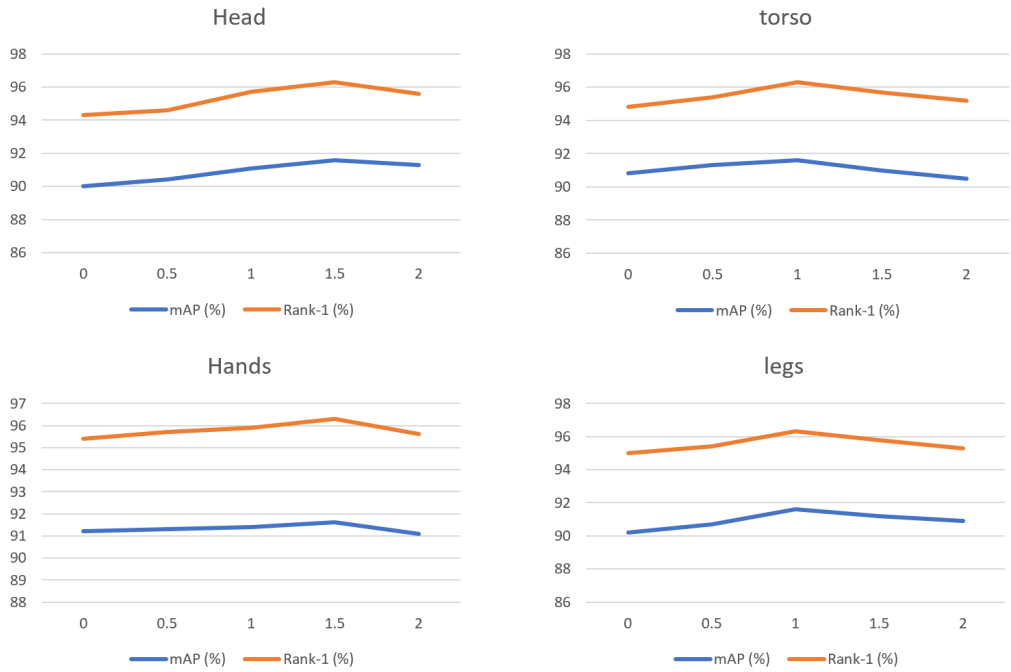


FIGURE 5.6: Effect of the boundary-penalty weight δ in the human body-part parsing module used by LSFP, measured in downstream Re-ID performance. Since LSFP pools features within the predicted part masks (head, torso, hands, and legs), δ influences the quality of the part regions and, consequently, the final tracklet descriptor. The results show a part-dependent optimum: $\delta = 1.5$ yields the best performance for the smaller regions (head and hands), whereas $\delta = 1.0$ is optimal for the larger regions (torso and legs). This behaviour is consistent with boundary sharpness being more critical for small parts, where minor boundary blur or leakage can corrupt a large fraction of the pooled features, while larger parts are less sensitive and can be over-regularized by excessively strong boundary penalties.

5.4 Conclusion

This chapter presented a supervised video-based person Re-ID framework for the accuracy-oriented deployment regime, where identity labels are available and the primary objective is to maximize retrieval performance. In contrast to the resource-constrained setting studied earlier, this chapter examined the highest-level performance achievable when full supervision and sufficient computational resources are assumed. The proposed framework combines a ViT backbone with complementary global and local representation branches

to exploit both tracklet-level spatiotemporal context and fine-grained identity cues.

The experimental results demonstrate that the proposed design consistently improves retrieval performance across multiple benchmarks. More importantly, the results reveal that global and local modelling play complementary roles in supervised video Re-ID. The spatiotemporal attention mechanism (SAN) enables robust aggregation of tracklet-level context, while the local branch (TFS and LSFP) enhances fine-grained discriminative cues that are critical for distinguishing similar visual identities. The combination of these components leads to a more structured and discriminative embedding space, as reflected in the substantial gains achieved through re-ranking. This indicates that the learned representations preserve meaningful neighbourhood relationships, allowing contextual refinement to recover hard positive matches.

More broadly, this chapter establishes that achieving high performance in supervised video Re-ID requires an explicit balance between global temporal aggregation and fine-grained local discrimination within a transformer-based framework. The proposed design demonstrates that neither component alone is sufficient and their integration leads to more robust identity representations under real-world challenges such as occlusion, viewpoint variation, and temporal inconsistency.

In the context of the overall thesis, this chapter provides an upper-bound reference for performance under full supervision, complementing the more constrained deployment regimes explored in earlier and subsequent chapters. These findings offer practical guidance for designing video Re-ID systems under varying levels of supervision and highlight the importance of aligning architectural choices with the constraints of the target deployment scenario.

Taken together, these results indicate that carefully designed transformer-based architectures, when guided by principled integration of temporal and

spatial cues, can effectively address the core challenges of video Re-ID. This insight provides a foundation for extending such designs to more constrained and realistic deployment settings explored in the remainder of this thesis.

Chapter 6

Transductive Video Re-ID

This chapter is based on our published work [10], with minor revisions for thesis coherence and additional experiments.

6.1 Introduction

This chapter focuses on the transductive setting for video-based person Re-ID, where a labelled source dataset is available, but the target dataset contains no identity annotations. In this regime, the central challenge is not label scarcity alone, but the domain gap between the source and target distributions. Differences in viewpoint, illumination, resolution, scene layout, and camera characteristics often cause models trained on one dataset to degrade when evaluated on another.

In video-based Re-ID, this difficulty is amplified by the interaction between temporal dynamics and appearance variation across surveillance environments. A model that performs well on benchmarks, such as MARS or Duke, may suffer substantial performance loss when transferred to a different dataset such as LS-VID. This makes domain adaptation essential for practical deployment, especially when target-domain labels are unavailable.

To address this setting, we introduce a Deep Transductive Learning (DTL) framework for video-based person Re-ID. The proposed method combines a transformer encoder for spatiotemporal feature extraction with an adversarial

minimax optimization strategy to encourage domain-invariant representation learning. Unlike approaches that rely only on clustering or simple domain confusion, our framework uses two complementary classifiers whose disagreement serves as an explicit signal of domain discrepancy. By minimizing this discrepancy through adversarial training, the encoder is encouraged to learn features that remain both transferable across domains and discriminative for identity matching.

A broader review of transfer learning, domain adaptation, and prior video-based Re-ID adaptation methods is provided in Chapter 2. Here, we focus on the aspects that directly motivate the proposed DTL framework. In particular, this chapter addresses the still-limited use of transformer-based backbones in domain-adaptive video Re-ID, despite their strong spatiotemporal modelling ability in the supervised setting.

This chapter makes the following contributions:

- **Novel DTL framework:** We propose an architecture that integrates Vision Transformers (ViT) with adversarial dual discriminators to explicitly reduce domain gaps while maintaining discriminative power for Re-ID.
- **Domain generalization across challenging datasets:** We conduct extensive experiments on the LS-VID dataset, which, to our knowledge, has been less explored outside the supervised setting.
- **Comprehensive evaluation:** We benchmark against state-of-the-art semi-supervised and adversarial models, perform ablation studies on the discrepancy loss and classifier design, and provide a detailed analysis of model behaviour under domain shift.

The remainder of this chapter is organized as follows. Section 6.2 details the proposed methodology, including encoder design, adversarial dual discriminators, and training strategy. Section 6.3 presents the experimental setup,

results, and ablation studies. Finally, Section 6.4 concludes with a summary of the findings and their implications.

6.2 Methodology

This section describes the proposed DTL framework for unsupervised domain adaptation in video-based Re-ID. The goal is to learn representations that are discriminative for identity recognition and robust to domain shifts across cameras and datasets. To achieve this, we integrate four key design elements: (i) a ViT encoder for spatiotemporal feature extraction, (ii) camera embeddings to model inter-camera variation, (iii) dual classifiers whose disagreement provides an adversarial signal, and (iv) a three-stage training strategy based on minimax optimization.

6.2.1 Problem Statement

Prior studies and empirical observations suggest that many Re-ID models are affected by overfitting. This behaviour becomes evident during cross-dataset evaluation, where models achieving over 90% accuracy when trained on a dataset often exhibit a drastic decline, dropping below 40% accuracy when evaluated on a different dataset (e.g., [72], [123]). This observation motivates our formulation of a transductive learning problem: given a labelled source dataset D_s and an unlabelled target dataset D_t, the goal is to transfer discriminative knowledge from D_s while adapting the learned features to remain effective in D_t. As such, we introduce a novel approach to address domain shifts in video-based person Re-ID. The proposed model begins by pre-training on the ImageNet dataset. Subsequently, we fine-tune the model on a labelled video-based Re-ID dataset. This dataset is then treated as the source domain for the subsequent adversarial training. We then train our adversarial network on a separate, unlabelled dataset, similar to the

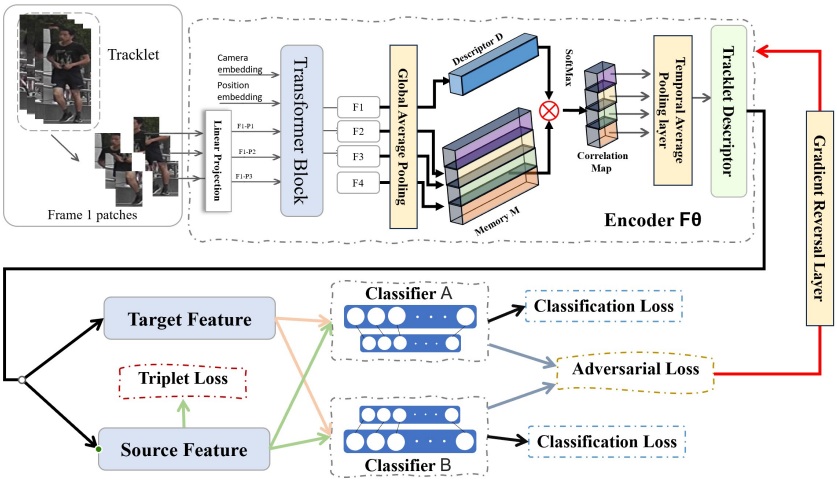


FIGURE 6.1: Illustration of DTL network model. Each frame is divided into patches, and position and camera embeddings are incorporated before the patches are input into the transformer block. The encoder $\mathcal{F}_\theta$ produces Tracklet Descriptors, and depending on their origin (source or target), classifiers ($\mathcal{A}$ and $\mathcal{B}$) select from three losses. The path indicated by the red arrow passes through the Gradient Reversal Layer and into the encoder during our third objective.

source dataset but with variations in terms of illumination, viewpoints, and background noise. This methodology encourages the model to learn features that are invariant across domains and camera perspectives.

6.2.2 Overall Idea and Model

At its core, our model consists of three primary networks: an encoder and two classifiers (Figure 6.1), which are referred to as a feature generator and discriminators in the context of GANs. The encoder is a Vision Transformer that produces tracklet-level embeddings, while the classifiers act as task-specific discriminators. Drawing inspiration from the approach outlined by Saito et al. [204] and the theorem presented by Ben-David et al. [205], we introduce dual classifiers with different initializations, unlike domain adaptation methods that rely on a single domain discriminator. The rationale is that two classifiers, trained on the same source domain, will agree on

confident samples but diverge on ambiguous or domain-shifted samples. This divergence provides a natural adversarial signal: the more the classifiers disagree, the more the encoder is encouraged to refine its features to reduce this discrepancy. This approach addresses domain shift in two complementary ways. First, it minimizes the disparity between the overall domains of the two datasets. Secondly, a unique embedding for each camera within both domains is encoded using our transformer network to emphasize camera-invariant features.

Given a source dataset domain $\mathcal{D}_s$ consisting of n_s labelled tracklets with their corresponding labels $\mathcal{Y}_s$, a target dataset domain $\mathcal{D}_t$ of n_t unlabelled tracklets, and an encoder function $\mathcal{F}_\theta$ that generates features, DTL aims to align source and target characteristics by employing task-specific classifiers as discriminators to account for the correlation between class boundaries and target features. Identifying target features distant from class support is necessary to achieve this goal. The classifier trained using the source features will likely misclassify these target features due to their proximity to the class boundaries. Moving forward, we suggest leveraging the discrepancy between the two classifiers' predictions of the target features to identify them.

In addition, we incorporate camera embeddings into the transformer input sequence. Each tracklet frame is tagged with its corresponding camera ID, which is encoded as a learnable vector. By explicitly injecting camera information, the model can differentiate between person identity features and camera-specific characteristics, such as illumination or viewpoint. This design is especially important for multi-camera datasets, where the same individual may appear drastically different across cameras.

6.2.3 Features Generator

To obtain representative feature vectors for each video tracklet, we first generate frame-level features and then aggregate them into tracklet-level features. We incorporate and encode patches to extract frame-level characteristics from our ViT encoder by dividing each frame into patches and adding position embedding before feeding them to our transformer model [16]. For each frame $I \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the frame height, width, and channels, we divide the frame into n equal-sized sections. To determine n, we use the convolutional operation's patch and stride sizes, P and S, as follows:

$$n = \left[\frac{H-P}{S} + 1\right] \times \left[\frac{W-P}{S} + 1\right] \tag{6.1}$$

Next, we flatten and map each patch $I_{p(n)}$ to D dimensions using the linear projection mapping function $\mathcal{F}_L$. Finally, the frame vector I is prepended with the learnable classifier I_c, camera $\mathcal{C}am$, and position $\mathcal{P}os$ embeddings:

$$I = \delta \cdot \mathcal{P}os + (1-\delta) \cdot \mathcal{C}am + \left[I_c; \mathcal{F}_L\left(I_{p(1)}\right); \mathcal{F}_L\left(I_{p(2)}\right); \cdots; \mathcal{F}_L\left(I_{p(n)}\right)\right], \tag{6.2}$$

where $\mathcal{P}os \in \mathbb{R}^{(n+1)\times D}$, $\mathcal{C}am \in \mathbb{R}^{(\mathcal{N}_c)\times D}$, $\mathcal{N}_c$ is the total number of cameras, and I is the encoder input vector. Finally, the transformer encoder generates the frame-level features as an output $\mathcal{T}_i$ of the i^{th} frame:

$$\mathcal{T}_i = I + MSA(LN(I)) + MLP(MSA(LN(I))) \tag{6.3}$$

After layer normalization (LN), the transformer block utilizes a multi-head self-attention (MSA) layer and a multilayer perceptron (MLP). The network aggregates the transformer encoder output to feed our Spatiotemporal Attention Network (SAN), inspired by [138], [164], [193]. This network uses a

query-memory attention block, where each frame is compressed into a descriptor $\mathcal{D}$ using Global Average Pooling (GAP), and the memory $\mathcal{M}$ contains the remaining frames' descriptors. Normalized Cross Correlation (NCC) is used to compute the correlation map for each descriptor by measuring the semantic relevance between $\mathcal{D}$ and all other descriptors in $\mathcal{M}$. To compute the final attention mask, a softmax layer is added to the correlation map [165]. Subsequently, the feature vectors used by the classifiers are generated by a Temporal Average Pooling (TAP) layer, which takes a temporal attention score and produces tracklet-level descriptors $\mathcal{D}_t$.

6.2.4 Adversarial Dual Discriminators

Given two classifiers ($\mathcal{A}$ and $\mathcal{B}$) that correctly identify labelled source samples, each with distinct characteristics and initializations, our first objective is to maximize and utilize the differences in their predictions on target samples, especially those beyond the source samples' support, which will likely have different classifications. Our second objective is to force the encoder to refrain from generating target features unsupported by the source by measuring the disagreement between these classifiers and training our encoder to minimize it. Thus, our objective becomes a minimax optimization problem.

Equation 6.4 depicts the discrepancy loss, expressed as the absolute difference between the probabilistic outputs of two classifiers:

$$d\left(p_{\mathcal{A}}, p_{\mathcal{B}}\right) = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \left| p_{\mathcal{A}_k} - p_{\mathcal{B}_k} \right|, \tag{6.4}$$

where $p_{\mathcal{A}_k}$ and $p_{\mathcal{B}_k}$ represent the probability outputs of classifiers $\mathcal{A}$ and $\mathcal{B}$ for class k in $\mathcal{K}$, respectively. Therefore, our adversarial loss can be expressed as the disparity between the predictions of label y of the two classifiers for each

tracklet x_t in $\mathcal{D}_t$, formulated as follows:

$$\ell_{adv}(\mathcal{D}_t) = \frac{1}{n_t} \sum_{t=1}^{n_t} d\left(p_{\mathcal{A}}(y \mid x_t), p_{\mathcal{B}}(y \mid x_t)\right) \tag{6.5}$$

We can further define the identification loss ℓ_{id}, which serves as the classification loss during training on the source dataset for both classifiers. This loss is computed as the cross-entropy between the probability of a tracklet x_s belonging to its own label class y_s, expressed as:

$$\ell_{id}\left(\mathcal{D}_s, \mathcal{Y}_s\right) = -\sum_{s=1}^{n_s} \log p(y_s \mid x_s) \tag{6.6}$$

During the early phase of training, we incorporate triplet loss, employing positive α and negative β pairs to optimize our feature generator. This involves drawing the tracklet x_s closer to the positive sample while simultaneously pushing it away from the negative one.

$$\ell_{Tri}\left(x_s, \alpha, \beta\right) = \max\left\{0, m + d_{x_s,\alpha} - d_{x_s,\beta}\right\}, \tag{6.7}$$

where m is the minimum margin between the two pairs.

6.2.5 Training Strategy

After pretraining the encoder on ImageNet, we address the training challenge through a three-step process (Figure 6.2):

Our first objective is to train the classifiers and the encoder to appropriately classify the source data. Ensuring that the classifiers and encoder gain task-specific discriminative features is crucial in this stage. We do this by reducing

both the identification and triplet losses:

$$\min_{\mathcal{F}_\theta, \mathcal{A}_\theta, \mathcal{B}_\theta} \rightarrow \gamma \cdot \ell_{id}\left(\mathcal{D}_s, \mathcal{Y}_s\right) + (1-\gamma) \cdot \ell_{Tri}\left(\mathcal{D}_s\right), \tag{6.8}$$

where γ is a parameter to control the weight of the two loss functions.

Subsequently, the classifiers (A and B) are trained to function as discriminators for our encoder $\mathcal{F}_\theta$ after we freeze its weights. With the classifiers trained to amplify the discrepancy between domains, they become adept at detecting target samples beyond the boundaries of the source domain's support. During this stage, we retain the classification loss for the source samples while excluding the triplet loss. Instead, we apply the adversarial loss to the target samples. Furthermore, we ensure that an equal number of source and target samples are utilized to update the model. Hence, our second objective function becomes:

$$\min_{\mathcal{A}_\theta, \mathcal{B}_\theta} \rightarrow \ell_{id}\left(\mathcal{D}_s, \mathcal{Y}_s\right) - \ell_{adv}(\mathcal{D}_t), \tag{6.9}$$

The final objective is to train the encoder $\mathcal{F}_\theta$ to minimize the discrepancy while maintaining the classifiers' weights untouched. To accomplish this goal, we employ the gradient of the adversarial loss from the previous stage, inverted, using a gradient reversal layer (GRL) [206], and subsequently propagate it backwards via the encoder. The objective of this reversal procedure is to minimize those features present in the discrepancy zone using the target dataset:

$$\min_{\mathcal{F}_\theta} \rightarrow \ell_{adv}(\mathcal{D}_t), \tag{6.10}$$

In contrast to the method employed by Saito et al. [204], we do not perform multiple iterations over the final objective. Instead, we leverage a gradient reversal layer, thereby reducing training time.

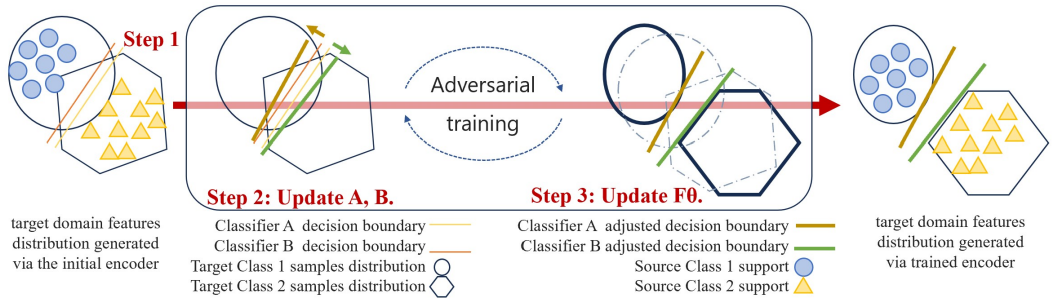


FIGURE 6.2: An overview of the adversarial training process. First, we increase the number of disagreement instances by adjusting the decision boundaries of the classifiers. Then, minimize the classifiers' disagreements via the feature encoder to find the distinguishing features between target-domain classes.

6.3 Experiments

6.3.1 Implementation Details

All models were developed using PyTorch. In accordance with the strong baseline practices outlined by Luo et al. [201], we adopted several training refinements that have proven effective for person Re-ID. In particular, their "bag of tricks" establishes a competitive baseline by systematically combining simple but powerful training strategies. While our architecture is based on a Vision Transformer rather than a CNN, many of these refinements remain relevant and were incorporated into our training pipeline.

Input video frames are resized to 256×128 and augmented by random horizontal flipping and random erasing with a probability of 0.5. Following [203], random erasing helps the model handle occlusion by masking random regions in training images, thereby improving robustness to background clutter and partial observations. Normalization is applied by dividing pixel values by $(0.5, 0.5, 0.5)$ and subtracting $(0.5, 0.5, 0.5)$ to standardize the RGB channels[16].

Batch construction follows a $P \times K$ sampling strategy. During the initial stage of training, sampling is performed only on the labelled source domain.

Specifically, each mini-batch contains 32 tracklets drawn from 8 source identities, with two tracklets per identity. In subsequent stages, each mini-batch remains of size 32 but is divided evenly between source- and target-domain tracklets. At this stage, the sampling is no longer identity-driven across both domains; instead, the emphasis is on maintaining a balanced representation of the source and target domains, since the identities in the two datasets are not shared and classifier training is not performed across domains. Following Luo et al. [201], this batch construction remains beneficial for triplet loss optimization on the labelled source samples, while the balanced inclusion of target samples supports domain adaptation within each mini-batch.

Optimization is performed with stochastic gradient descent (SGD), using a learning rate of 0.008, momentum of 0.9, weight decay of 10^{-4}, and a cosine decay schedule. While Luo et al. adopt Adam with a warm-up strategy, we found that SGD with cosine decay is more stable for ViT-based encoders. Nonetheless, the principle of gradually adapting the learning rate, as emphasized in [201], is preserved in our model.

The encoder backbone is a Vision Transformer [16] pretrained on ImageNet-21K. Training proceeds for 120 epochs. Each tracklet is composed of 6 frames, selected using a restricted random sampling (RRS) strategy to maintain temporal diversity. As positional embeddings remain constant across frames, we set $\delta = 0.25$ in the transformer input to emphasize the role of camera embeddings and to better capture cross-camera viewpoint variations.

To increase robustness, we further split each tracklet into C equal clips, following the augmentation principle in [123]. Unless otherwise specified, $C = 2$ is used. This strategy serves as data augmentation without applying transformations, effectively increasing the number of samples while providing reliable positive pairs for triplet loss.

We also integrate the insights from Luo et al. [201] regarding the balance between identification loss and triplet loss. To emphasize the discriminative

TABLE 6.1: Cross-dataset transfer performance using mINP, mAP, and Rank-1 CMC. White cells represent transductive transfer results (no target labels), while highlighted cells indicate supervised upper bounds obtained by training directly on the target dataset. The gap between these values reflects the impact of domain shift.

Source	Target mINP/ mAP / Rank-1 CMC								
	Mars			LS-Vid			Duke		
Mars	68.7	89.6	95.7	25.3	37.4	52.7	64.4	83.1	96.8
LS-Vid	43.6	69.9	85.1	41.4	55.6	67.4	61.0	80.2	99.6
Duke	41.2	68.7	82.6	26.2	38.6	53.9	77.9	89.2	99.6

role of triplet loss in early training, we assign it a weighting factor $\gamma = 0.66$. For classification, we use two classifiers, each consisting of four fully connected (FC) layers with three intermediate batch normalization layers. Random initialization of classifier weights encourages early divergence, which is necessary for effective discrepancy learning. Finally, for the discrepancy loss, we use the L1 distance, which our ablation studies (see Section 6.3.2) confirm to be more stable and effective than L2 or KL divergence.

6.3.2 Results and Comparison

The mINP, mAP, and Rank-1 CMC scores achieved by our method on three major video Re-ID datasets are shown in Table 6.1. The supervised upper bounds, obtained by training directly on the target dataset, are highlighted in the cells. For the MARS and Duke datasets, our methodology closely approaches the theoretical maximum outcomes; however, this is not the case for LS-Vid. This highlights the importance of evaluating our method on LS-Vid, because even supervised training fails to achieve the 70% accuracy threshold. This makes LS-Vid an excellent stress test for domain adaptation approaches, as it better reflects the challenges of real-world multi-camera deployments.

TABLE 6.2: Evaluation results on AG-VPReID where Source represents the source dataset and the target will be either the ground split of the dataset or the aerial split.

Source	AG-Ground		AG-Aerial	
	MAP	Rank1	MAP	Rank1
Mars	65.3	71.3	54.8	70.3
LS-ViD	66.1	72.6	55.3	71.4
AG-Aerial	62.4	70.2	-	-
AG-Ground	-	-	56.4	73.8

The results in Table 6.1 show that our method transfers well between MARS and Duke, achieving performance close to supervised upper bounds. By contrast, performance drops considerably when transferring to LS-Vid, confirming its status as a more challenging dataset with greater camera and environmental diversity. This drop can be attributed to the increased scale and diversity of LS-Vid, which introduces greater variation in camera viewpoints, motion patterns, and environmental conditions. Unlike MARS and Duke, LS-Vid contains more complex and less constrained scenarios, making domain alignment more difficult. Furthermore, the absence of target labels prevents the model from fully adapting to these variations, which limits performance.

This observation suggests that while DTL is effective under moderate domain shifts, more challenging datasets such as LS-Vid may require additional mechanisms, such as stronger temporal modelling or partial supervision, to further reduce the performance gap.

The difference between the transfer results (white cells) and the supervised upper bounds (highlighted cells) reflects the inherent difficulty of cross-domain generalization in video-based Re-ID. In the supervised setting, the model has access to identity labels from the target domain, allowing it to learn domain-specific discriminative features. In contrast, the transductive setting relies solely on labelled source data and unlabelled target data, making it significantly more challenging.

Therefore, the observed performance gap does not indicate a failure of the proposed domain adaptation method but rather highlights the extent of the domain shift. The role of DTL is to reduce this gap as much as possible, and the results demonstrate that it consistently improves transfer performance compared to prior methods, even though a gap to fully supervised training remains.

Table 6.2 extends our evaluation to the AG-VPReID cross-platform setting, where the domain shift is induced by viewpoint and platform changes (ground vs. aerial) rather than conventional camera-to-camera variation. The results show that our model maintains stable Rank-1 performance when transferring between these two splits, with only a modest drop in mAP, indicating that the learned representation remains largely discriminative under severe viewpoint changes. This experiment complements Table 6.1 by demonstrating that the proposed discrepancy-based adaptation is not limited to standard benchmark transfers but also generalizes to heterogeneous capture conditions that better reflect practical deployments.

Our DTL is also evaluated against state-of-the-art algorithms, including those utilizing one-shot annotation, clustering, association, and adversarial techniques, on two prominent video Re-ID datasets: Duke-VideoReID and MARS. The comparative analysis is presented in Table 6.3. Our results show improvements in Rank-1 accuracy over existing methods, exceeding prior approaches by more than 3% on Duke and by more than 1% on MARS. While some clustering-based methods yield competitive mAP, our approach consistently delivers higher Rank-1 accuracy. This indicates that discrepancy-driven adaptation improves reliability on the most critical evaluation metric in practice.

In Table 6.3, DTL achieves the highest Rank-1 accuracy among the compared methods on both MARS and Duke. However, DCCAL achieves higher mAP on both datasets, indicating stronger overall retrieval consistency across

TABLE 6.3: Comparison with state-of-the-art unsupervised and semi-supervised domain adaptation methods on MARS and Duke.

Model	year	Type	MARS		Duke	
			mAP	Rank-1	mAP	Rank-1
Baseline	-	Supervised	15.45	36.16	49.5	48
EUG [72]	2018	One shot	42.4	62.6	63.2	72.7
SRC [103]	2022	Clustering	40.5	62.7	76.5	83.0
TASTC [137]	2023	association	47.2	65.6	68.2	76.8
MPC [108]	2023	Clustering	71.4	81.6	87.3	89.3
CAWCL [123]	2023	Adv.	44.8	62.2	-	-
DCCAL [124]	2024	Adv.	**74.5**	84.0	**91.4**	93.0
DTL (Ours)	2025	Adv.	69.9	**85.1**	83.1	**96.8**

ranks. This suggests that while DCCAL is more effective at optimizing global ranking quality, DTL is particularly strong at correctly retrieving the top match, which is often the most critical requirement in practical Re-ID systems.

This trade-off highlights a key difference in behaviour: discrepancy-driven adaptation in DTL prioritizes confident top-rank predictions, whereas clustering-based approaches such as DCCAL may better capture broader ranking distributions. Therefore, the two methods can be seen as complementary, with DTL excelling in Rank-1 reliability and DCCAL providing stronger mAP performance.

To further analyze the retrieval behaviour of the proposed DTL model, Figure 6.3 presents qualitative results and compares them with earlier models from Chapters 4 and 5.

In case (d), where the query originates from Camera 6, the model successfully retrieves correct matches from other cameras within the top ranks. This behaviour contrasts with the supervised model in Chapter 5 (Figure 5.5), where tracklets from Camera 6 were rarely observed in higher ranks, indicating that the proposed transductive framework improves cross-camera

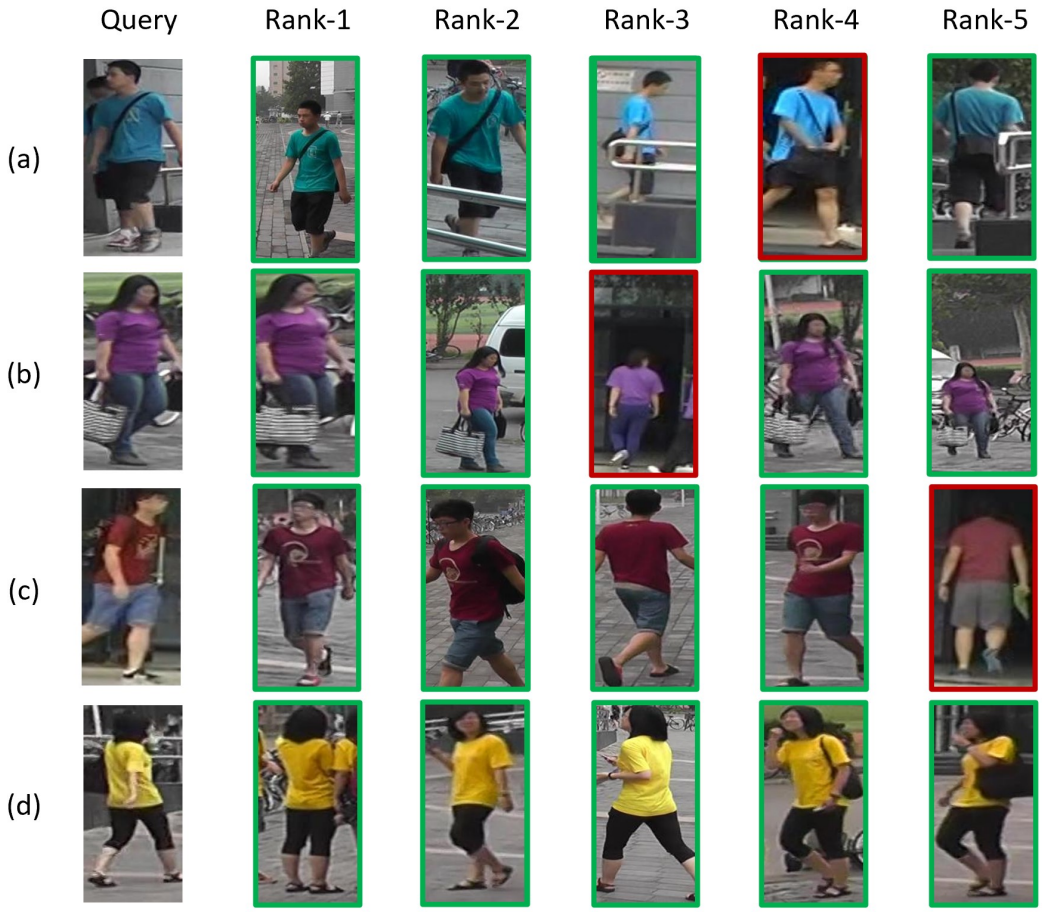


FIGURE 6.3: Qualitative retrieval results of the proposed DTL model on MARS. Green boxes indicate correct matches, and red boxes indicate incorrect matches. Compared to earlier models (Figs. 4.3 and 5.5), the DTL framework demonstrates improved cross-camera retrieval (e.g., Camera 6 in (d)), while still exhibiting confusion in fine-grained appearance discrimination in some cases (a–c).

generalization.

However, in cases (a)–(c), limitations in fine-grained discrimination are evident. For instance, in case (a), both a correct match (Rank-3) and an incorrect match (Rank-4) are retrieved from Camera 6, suggesting ambiguity in appearance similarity. A similar pattern is observed in case (b), where Rank-3 corresponds to a false match from the same camera, and in case (c), where a false match appears at Rank-5.

These observations indicate that, although the DTL model improves domain and camera invariance, it does not consistently preserve fine-grained

identity discrimination. In contrast, the supervised model in Chapter 5 benefits from part-based representations, which better capture local discriminative features and reduce such ambiguities. This highlights a trade-off between cross-domain robustness and fine-grained feature discrimination across the three proposed models.

6.3.3 Ablation Study

Table 6.4 compares different choices for the discrepancy loss. L1-norm consistently yields the best results, outperforming L2 and KL divergence on both datasets. This suggests that L1 provides a stronger gradient for minimax optimization and prevents vanishing gradients that can arise with L2. This behaviour is particularly beneficial in Re-ID, where emphasizing hard-to-separate identity boundaries is more important than averaging discrepancies across classes. These results confirm that the L1 distance provides a more stable and informative discrepancy, resulting in better alignment between the source and target domains. Consequently, this design choice is central to the robustness of the proposed DTL method.

It is important to note that the results reported in Table 6.4 are not directly comparable to those in Tables 6.1 and 6.3. While Tables 6.1 and 6.3 report cross-dataset transfer performance under fixed experimental settings, Table 6.4 presents an ablation study evaluating individual components under controlled configurations. In particular, the discrepancy-loss variants are assessed in isolation, without necessarily enforcing the exact same training schedule, data splits, or hyperparameter tuning used in the full model evaluation.

As a result, the best-performing configuration in Table 6.4 may yield slightly different absolute values than those in Tables 6.1 and 6.3. Therefore, the purpose of Table 6.4 is to analyze relative trends between design choices rather than to provide directly comparable performance benchmarks.

TABLE 6.4: Ablation study comparing different discrepancy loss functions (L1, L2, KL) on MARS and LS-Vid.

Target	Mars		LS-Vid	
divergence	mAP	Rank1	mAP	Rank1
L1-norm	**69.9**	**85.1**	**37.4**	**52.7**
L2-norm	57.6	79.4	24.5	36.8
KL	66.5	82.9	29.0	46.1

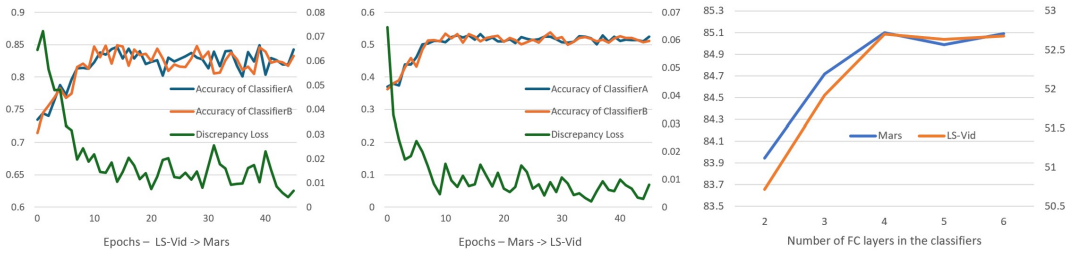


FIGURE 6.4: The relationship between discrepancy loss (green) and accuracy (blue and orange) during training for both Mars and LS-Vid datasets. As the discrepancy loss decreases, accuracy increases for both classifiers. The rightmost chart illustrates the impact of the number of fully connected layers on the accuracy for both classifiers.

TABLE 6.5: Performance of the features generator.

Target Dataset	**Mars**		**LS-Vid**	
Feature Embedding	**mAP**	**Rank1**	**mAP**	**Rank1**
Tracklet descriptor	**69.9**	**85.1**	**37.4**	**52.7**
AvrPooling of FD	57.7	82.3	24.8	36.7
MaxPooling of FD	57.1	81.5	22.7	28.5
classifier token	69.7	83	28.3	45.7
w/o camera emb.	42.4	69.1	11.2	22.7
Number of clips per tracklet				
C =1	69.8	83.3	28.9	46.1
C= 2	**69.9**	**85.1**	**37.4**	**52.7**
C= 3	58	84.8	36.7	52
C= 4	57.6	84.5	36.5	51.8

Figure 6.4 illustrates the training dynamics. As discrepancy decreases, accuracy consistently improves, demonstrating the effectiveness of the minimax procedure. The right panel shows that deeper classifiers provide diminishing returns, indicating that moderate classifier depth is sufficient and avoids unnecessary model complexity.

Table 6.5 further analyzes the feature generator. Using SAN descriptors outperforms simple pooling, and camera embeddings prove critical. Splitting tracklets into two clips yields the best results, while larger splits fragment sequences and degrade performance. These observations confirm that both temporal modelling (via SAN) and explicit handling of camera variation (via embeddings) are indispensable for achieving strong performance across datasets.

The first section of Table 6.5 shows the impact of our encoder components on the efficacy of our adversarial video-based person Re-ID task. Our

analysis begins by investigating the SAN tracklet descriptors, revealing a significant performance decline when using only max pooling or average pooling on frame descriptors for tracklet descriptor generation, rather than SAN. Likewise, using the classifier token instead of frame descriptors, as in ViT implementations, results in a slight performance drop. Finally, the absence of camera embeddings results in the poorest performance of our investigation. Including camera embeddings is crucial, as it enables the encoder to learn and account for inter-camera variations in illumination and viewpoints, thereby significantly improving the model's performance.

Since the video-based person Re-ID has access to each frame of the whole tracklet, it is possible to split it into clips. We observe in the second part of Table 6.5 that dividing each tracklet into clips improves overall performance, with splitting tracklets into two clips ($C = 2$) yielding the best trade-off between data augmentation and preserving temporal consistency since larger splits ($C = 3$ or $C = 4$) fragment the tracklet too aggressively, resulting in noisy pairs and reduced performance. This behaviour can be explained by two main factors. First, the number of samples in the dataset increases, effectively serving as data augmentation, but without transformation. Second, it generates strong, reliable positive pairs, providing a supervised indicator even with unlabelled data.

Overall, these experiments show that the key components of the proposed DTL framework collectively improve transferability under domain shift. The combination of discrepancy-based training, SAN descriptors, and camera embeddings results in a model that remains both discriminative and robust across domains.

6.4 Conclusion

This chapter presented the proposed Deep Transductive Learning (DTL) framework for domain-adaptive video-based person Re-ID. By combining a Vision Transformer encoder, camera embeddings, and dual-classifier adversarial discrepancy learning, the framework is designed to improve cross-domain generalization when labelled target data are unavailable.

The experimental results demonstrate that DTL transfers effectively across challenging video Re-ID benchmarks. On Duke and MARS, the method approaches the performance of supervised methods, while on LS-VID, it establishes a strong baseline under domain shift and highlights the value of testing adaptation methods on larger and more diverse datasets. The ablation studies further confirm that discrepancy-based training, SAN-based tracklet descriptors, and camera embeddings each contribute to improving cross-domain performance under the evaluated settings.

6.4.1 Broader Implications

Although this chapter focuses on video-based person Re-ID, the main design ideas of DTL have broader relevance to video representation learning under distribution shift. In particular, camera embeddings, clip-splitting as a form of temporal augmentation, and discrepancy-driven adaptation may also be useful in related problems such as action recognition, gait analysis, and cross-camera video retrieval. In this sense, the contribution of DTL extends beyond Re-ID to more general settings where robust feature learning is required across heterogeneous domains.

Limitations. Despite these improvements, the proposed DTL framework has several limitations. The adversarial training process introduces additional optimization complexity and may be sensitive to hyperparameter

selection. Furthermore, the approach depends on the quality and diversity of the source dataset; a significant mismatch between source and target domains may still degrade performance. Finally, the dual-classifier design increases computational overhead, potentially limiting real-time deployment in resource-constrained environments.

6.4.2 Future Directions

Several promising research directions remain:

- **Multi-source domain adaptation:** Extending DTL to leverage multiple labelled source datasets may further improve generalization to more diverse target environments.
- **Lightweight architectures:** Investigating hybrid CNN–Transformer models or efficient ViT variants may enable real-time deployment in resource-constrained surveillance systems.
- **Semi-supervised extensions:** Incorporating small amounts of labelled target data could help bridge the gap between unsupervised and fully supervised adaptation.

Overall, this chapter shows that adversarial discrepancy minimization can be used to learn video representations that are both discriminative and more robust under domain shift. Together with the previous chapters, these results support the broader thesis objective of developing a video-based person Re-ID models that remain practical, scalable, and adaptable in real-world surveillance settings.

Chapter 7

Conclusion and Future Work

To provide a consolidated view of the trade-offs studied throughout the thesis, Table 7.1 summarizes the retrieval performance of the three thesis configurations across MARS, LS-ViD, and the AG-VPReID ground/aerial splits, while Table 7.2 reports their computational characteristics in terms of model size, throughput, and FLOPs. Taken together, these two tables highlight that the three configurations are not competing for the same deployment objective but rather occupy different operating points along the accuracy-efficiency spectrum.

The supervised configuration delivers the strongest overall retrieval accuracy on the large-scale benchmarks, achieving 91.6 mAP / 96.3 Rank-1 on MARS and 85.3 mAP / 93.6 Rank-1 on LS-ViD, while also obtaining the best AG-Aerial performance with 58.6 mAP / 76.1 Rank-1. This confirms that, when labelled data and sufficient computational resources are available, fully supervised spatiotemporal learning remains the most effective approach for

TABLE 7.1: Summary comparison of the three thesis configurations across MARS, LS-ViD, and AG-VPReID (ground/aerial) using mAP and Rank-1 (R1).

Model	MARS		LS-ViD		AG-Ground		AG-Aerial	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
one-Shot	55.5	72.2	63.4	72.8	61.5	70.0	54.9	65.7
DTL	69.9	85.1	81.2	89.4	**66.1**	**72.6**	56.4	73.8
Supervised	**91.6**	**96.3**	**85.3**	**93.6**	65.7	72.2	**58.6**	**76.1**

TABLE 7.2: Computational comparison of the three thesis configurations in terms of the number of parameters, throughput in frames per second (FPS) estimate, and floating-point operations (FLOPs).

Model	Parameters	FPS	FLOPs
one-Shot	4.7M	475	0.31G
DTL	23.5M	204	1.8G
Supervised	27.7M	117	4.5G

maximizing retrieval performance. However, Table 7.2 shows that this gain comes at the highest computational cost, with 27.7M parameters, an estimated throughput of 117 FPS, and 4.5G FLOPs.

The discrepancy-driven transductive configuration provides the best balance when target-domain labels are unavailable. As shown in Table 7.1, it substantially improves over the one-shot setting on all datasets and achieves the best performance on AG-Ground with 66.1 mAP / 72.6 Rank-1, highlighting the value of adaptation under cross-platform and cross-view domain shift. At the same time, its computational profile remains moderate relative to the supervised model, requiring 23.5M parameters, 204 FPS, and 1.8G FLOPs, as reported in Table 7.2. This makes it a practical compromise between robustness and efficiency when deployment must generalize beyond the training domain.

By contrast, the one-shot configuration is the most computationally efficient model in the thesis. Although its retrieval performance is lower than that of the other two configurations, it still achieves competitive results in a lightweight setting, including 55.5 mAP/72.2 Rank-1 on MARS and 63.4 mAP/72.8 Rank-1 on LS-ViD. More importantly, it requires only 4.7M parameters, runs at an estimated 475 FPS, and uses just 0.31G FLOPs. These characteristics make it particularly attractive for early-stage deployment, rapid prototyping, and resource-constrained camera networks where minimizing computational overhead is more important than extracting the last increment

of accuracy.

Beyond the quantitative results, a qualitative comparison across Chapters 4–6 reveals important differences in the behaviour of the three configurations. The one-shot model, while highly efficient, relies on limited supervision and therefore exhibits weaker identity discrimination, particularly under viewpoint variation and cross-camera conditions. It often retrieves correct identities within the top ranks, but its representations remain sensitive to appearance ambiguity and domain shift.

The supervised model addresses these limitations by learning strong spatiotemporal and part-aware representations, resulting in the most consistent fine-grained discrimination across challenging scenarios such as occlusion, pose variation, and background clutter. This is reflected not only in its superior benchmark performance but also in its ability to reduce visually similar false matches through more precise localized feature modelling.

In contrast, the discrepancy-driven transductive model emphasizes domain invariance rather than purely discriminative power. As observed in Chapter 6, it improves cross-camera and cross-domain retrieval, particularly in settings where the target distribution differs significantly from the source. However, this robustness comes with a trade-off: the model may exhibit reduced sensitivity to fine-grained identity cues, leading to occasional confusion between visually similar individuals.

Taken together, these observations highlight a fundamental trade-off across the three regimes: efficiency, discrimination, and generalization. The one-shot model prioritizes computational efficiency, the supervised model maximizes discriminative accuracy, and the transductive model balances adaptation and robustness under domain shift. This qualitative analysis reinforces the quantitative findings and supports the central thesis argument that video-based Re-ID should be treated as a set of deployment-specific solutions rather than a single universal model.

Overall, the comparison confirms the central thesis argument that no single Re-ID configuration is universally optimal across all surveillance scenarios. Instead, the most suitable solution depends on the availability of annotations, the degree of domain shift, and the computational budget of the target platform. The one-shot model is the most suitable option for lightweight, cost-sensitive deployments; the discrepancy-driven model is the most appropriate when adaptation to an unlabelled target domain is required; and the supervised model is the preferred choice when the primary objective is maximum retrieval accuracy.

7.1 Limitations and Future Work

Despite the advances presented in this thesis, several limitations remain and motivate directions for future work.

First, most of the experiments in this thesis follow the common Re-ID assumption that identity is preserved through appearance consistency. In real deployments, however, a person's appearance can change over time due to changes in clothing, carried objects, seasonal variation, or partial occlusion patterns. This reduces the validity of models trained under the appearance constancy hypothesis and limits long-term identity persistence. Future work should relax this assumption by incorporating long-term appearance variation during training, for example, through stronger augmentation policies, synthetic appearance changes, and learning objectives that emphasize identity cues that are less sensitive to clothing.

Second, while benchmark datasets are essential for reproducibility, they do not fully capture the noise characteristics of real surveillance pipelines. Practical camera networks introduce imperfect detections, tracker fragmentation, compression artifacts, and substantial variation in resolution and illumination. These factors can degrade tracklet quality and increase the

difficulty of reassociation. Future work should evaluate the proposed methods under more realistic end-to-end conditions and consider training strategies that explicitly model noisy tracklets, including uncertainty-aware pseudo-labelling, memory-based stabilization, and consistency constraints across time and cameras.

Third, computational cost remains a practical constraint for video-based Re-ID. Training and inference on tracklets are more demanding than image-based settings, and the higher-capacity supervised and adaptation-based configurations can be expensive for small installations. Although this thesis includes a resource-aware configuration and explores lightweight backbones, further efficiency improvements are needed for real-time deployment at scale. Future directions include model compression and distillation from high-capacity models into compact deployment models, more efficient temporal aggregation, and adaptive frame sampling to reduce redundant computation.

Finally, domain shift is only partially addressed by existing adaptation strategies. Even when viewpoint and illumination shifts are mitigated, additional changes occur in real deployments, such as camera replacements, seasonal drift, and environmental modifications. Future work should explore continual, safe online adaptation mechanisms that update the model over time without catastrophic forgetting or drifting due to noisy pseudo-labels.

Overall, these limitations highlight that robust Re-ID requires both strong representation learning and system-level reliability. Extending evaluation to larger, more diverse camera networks, incorporating stronger robustness mechanisms, and improving efficiency will increase the practicality of Re-ID in real-world surveillance applications.

Bibliography

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013, ISSN: 0167-8655.

[2] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19, New York, NY, USA: Association for Computing Machinery, Oct. 15, 2019, pp. 365–373, ISBN: 978-1-4503-6889-6.

[3] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5098–5107.

[4] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Jul. 19, 2019, pp. 7183–7192.

[5] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1179–1188.

[6] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," presented at the

Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3765–3773.

[7] Y. Wang et al., "Resource aware person re-identification across multiple resolutions," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8042–8051.

[8] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4516–4524.

[9] S. Bak, S. Zaidenberg, B. Boulay, and F. Brémond, "Improving person re-identification by viewpoint cues," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2014, pp. 175–180.

[10] H. Alghamdi, W. El Ahmar, and R. Laganière, "Deep Transductive Learning for Person Re-Identification," in *Image Analysis and Processing – ICIAP 2025*, E. Rodolà, F. Galasso, and I. Masi, Eds., Cham: Springer Nature Switzerland, 2026, pp. 512–524, ISBN: 978-3-032-10192-1.

[11] B. Bordelon, L. Noci, M. B. Li, B. Hanin, and C. Pehlevan, "Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit," *International Conference on Learning Representations (ICLR)*, 2023, arXiv preprint arXiv:2309.16620.

[12] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–41, Jan. 31, 2022, ISSN: 0360-0300, 1557-7341. arXiv: 2101.01169[cs].

[13] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[17] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 15, 1997, ISSN: 0899-7667.

[19] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person ReID," *arXiv preprint arXiv:1805.02104*, 2018. arXiv: 1805.02104 [cs.CV].

[20] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1476–1485.

[21] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds., Cham: Springer International Publishing, 2018, pp. 270–279, ISBN: 978-3-030-01424-7.

[22] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv preprint arXiv:1412.4446*, 2015. arXiv: 1412.4446 [cs.LG].

[23] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 37, 2015, pp. 1180–1189.

[24] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, " Meta-Learning in Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 09, pp. 5149–5169, Sep. 2022, ISSN: 1939-3539.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3652–3661, ISSN: 1063-6919.

[26] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 2528–2535, ISSN: 2380-7504.

[27] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency weighted features for person re-identification," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 191–208, ISBN: 978-3-319-16199-0.

[28] L. An, X. Chen, S. Liu, Y. Lei, and S. Yang, "Integrating appearance features and soft biometrics for person re-identification," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 12 117–12 131, May 1, 2017, ISSN: 1573-7721.

[29] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2011, pp. 501–512, ISBN: 978-3-642-19282-1.

[30] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 3402–3408.

[31] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 8, pp. 1447–1460, Aug. 2016, ISSN: 1558-2205, Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*, Aug. 2014, pp. 34–39, ISSN: 1051-4651.

[33] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[34] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2008, pp. 262–275, ISBN: 978-3-540-88682-2.

[35] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local

features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2360–2367, ISSN: 1063-6919.

[36] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.

[37] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 144–151.

[38] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1288–1296.

[39] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1367–1376.

[40] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5399–5408.

[41] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.

[42] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 719–728.

[43] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 608–617.

[44] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 135–153, ISBN: 978-3-319-46478-7.

[45] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, Feb. 1, 2019, ISSN: 0031-3203.

[46] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.

[47] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 365–381.

[48] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9637–9646.

[49] J. Si et al., "Dual attention matching network for context-aware feature sequence based person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5363–5372.

[50] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5735–5744.

[51] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8040–8049.

[52] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8649–8658.

[53] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4976–4985.

[54] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 402–419.

[55] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3219–3228.

[56] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.

[57] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.

[58] H. Zhao et al., "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.

[59] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.

[60] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2119–2128.

[61] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 667–676.

[62] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3642–3651.

[63] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2288–2295, ISSN: 1063-6919.

[64] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR 2011*, Jun. 2011, pp. 649–656, ISSN: 1063-6919.

[65] X. Liu, H. Wang, Y. Wu, J. Yang, and M.-H. Yang, "An ensemble color model for human re-identification," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan. 2015, pp. 868–875, ISSN: 1550-5790.

[66] B. J. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *British Machine Vision Conference*, 2010.

[67] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1278–1287.

[68] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, 13:1–13:20, Dec. 13, 2017, ISSN: 1551-6857.

[69] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3800–3808.

[70] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5157–5166.

[71] M. Tian et al., "Eliminating background-bias for robust person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5794–5803.

[72] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5177–5186.

[73] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3754–3762.

[74] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 598–607.

[75] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 994–1003.

[76] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 748–756.

[77] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 51–58, Apr. 1, 2019, ISSN: 1047-3203.

[78] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020, ISSN: 1941-0077, Conference Name: IEEE Transactions on Multimedia.

[79] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" In *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[80] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, IEEE Computer Society, 2006, pp. 1735–1742.

[81] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. arXiv: 1703.07737 [cs.CV].

[82] H. Shi et al., "Embedding deep metric for person re-identification: A study against large variations," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 732–748, ISBN: 978-3-319-46448-0.

[83] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, Dec. 2015, pp. 4678–4686, ISBN: 978-1-4673-8391-2.

[84] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 542–551.

[85] S. Zhou et al., "Deep self-paced learning for person re-identification," *Pattern Recognition*, vol. 76, pp. 739–751, Apr. 1, 2018, ISSN: 0031-3203.

[86] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *Pattern Recognition*, vol. 98, p. 107 036, Feb. 1, 2020, ISSN: 0031-3203.

[87] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3391–3402, Sep. 2021, ISSN: 1051-8215, 1558-2205.

[88] R. R. Varior, G. Wang, and J. Lu, *Learning invariant color features for person re-identification*, Oct. 9, 2014. arXiv: 1410.1035[cs].

[89] Y. Huang, Z.-J. Zha, X. Fu, R. Hong, and L. Li, *Real-world person re-identification via degradation invariance learning*, Apr. 10, 2020. arXiv: 2004.04933[cs].

[90] Y.-J. Cho and K.-J. Yoon, "PaMM: Pose-aware multi-shot matching for improving person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3739–3752, Aug. 2018, ISSN: 1057-7149, 1941-0042. arXiv: 1705.06011[cs].

[91] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 1233–1245, 2020, ISSN: 1057-7149, 1941-0042. arXiv: 1903.12395[cs].

[92] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3390–3399.

[93] C. Zhao et al., "Maximal granularity structure and generalized multi-view discriminant analysis for person re-identification," *Pattern Recognition*, vol. 79, pp. 79–96, Jul. 1, 2018, ISSN: 0031-3203.

[94] J. Meng, A. Wu, and W.-S. Zheng, "Deep asymmetric video-based person re-identification," *Pattern Recognition*, vol. 93, pp. 430–441, Sep. 1, 2019, ISSN: 0031-3203.

[95] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2429–2438.

[96] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019, ISSN: 1941-0042, Conference Name: IEEE Transactions on Image Processing.

[97] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 170–186.

[98] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6210–6219.

[99] G. Ding, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Towards better validity: Dispersion based clustering for unsupervised person re-identification," *arXiv:1906.01308 [cs]*, Jun. 4, 2019. arXiv: 1906.01308.

[100] D. S. Raychaudhuri and A. K. Roy-Chowdhury, "Exploiting temporal coherence for self-supervised one-shot video re-identification," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M.

Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 258–274, ISBN: 978-3-030-58583-9.

[101] B. Pang, D. Zhai, J. Jiang, and X. Liu, "Fully unsupervised person re-identification via selective contrastive learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 2, 64:1–64:15, Feb. 16, 2022, ISSN: 1551-6857.

[102] X. Wang, M. Liu, D. S. Raychaudhuri, S. Paul, Y. Wang, and A. K. Roy-Chowdhury, "Learning person re-identification models from videos with weak supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 3017–3028, 2021, ISSN: 1941-0042, Conference Name: IEEE Transactions on Image Processing.

[103] P. Xie, X. Xu, Z. Wang, and T. Yamasaki, "Sampling and re-weighting: Towards diverse frame aware unsupervised video person re-identification," *IEEE Transactions on Multimedia*, vol. 24, pp. 4250–4261, 2022, ISSN: 1941-0077, Conference Name: IEEE Transactions on Multimedia.

[104] L. Lan, X. Teng, J. Zhang, X. Zhang, and D. Tao, "Learning to Purification for Unsupervised Person Re-Identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 3338–3353, 2023, ISSN: 1941-0042.

[105] T. Liu, S. Cheng, and A. Du, "Multi-view similarity aggregation and multi-level gap optimization for unsupervised person re-identification," *Expert Systems with Applications*, vol. 256, p. 124 924, Dec. 2024, ISSN: 0957-4174.

[106] H. Sun and S. Ma, "Pro-ReID: Producing reliable pseudo labels for unsupervised person re-identification," *Image and Vision Computing*, vol. 150, p. 105 244, Oct. 2024, ISSN: 0262-8856.

[107] Q. Wang, Z. Huang, H. Fan, S. Fu, and Y. Tang, "Unsupervised person re-identification based on adaptive information supplementation and foreground enhancement," *IET Image Processing*, vol. 18, no. 14, pp. 4680–4694, 2024, ISSN: 1751-9667.

[108] X. Li, Q. Li, F. Liang, and W. Wang, "Multi-granularity pseudo-label collaboration for unsupervised person re-identification," *Computer Vision and Image Understanding*, vol. 227, p. 103 616, Jan. 1, 2023, ISSN: 1077-3142.

[109] X. Qu, L. Liu, L. Zhu, L. Nie, and H. Zhang, "Source-free style-diversity adversarial domain adaptation with privacy-preservation for person re-identification," *Knowledge-Based Systems*, vol. 283, p. 111 150, Jan. 11, 2024, ISSN: 0950-7051.

[110] H. Rami, J. H. Giraldo, N. Winckler, and S. Lathuilière, "Source-guided similarity preservation for online person re-identification," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1711–1720.

[111] J. Peng, J. Yu, C. Wang, H. Wang, and X. Fu, "Adapt only once: Fast unsupervised person re-identification via relevance-aware guidance," *Pattern Recognition*, vol. 150, p. 110 360, Jun. 2024, ISSN: 0031-3203.

[112] M. Liu, Y. Bian, Q. Liu, X. Wang, and Y. Wang, "Weakly Supervised Tracklet Association Learning With Video Labels for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3595–3607, May 2024, ISSN: 1939-3539.

[113] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 8079–8088, ISBN: 978-1-7281-4803-8.

[114] Y. Fu et al., "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 6111–6120, ISBN: 978-1-7281-4803-8.

[115] Y. Ge, D. Chen, and H. Li, *Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification*, Jan. 30, 2020. arXiv: 2001.01526[cs].

[116] Y. Ge, F. Zhu, D. Chen, R. Zhao, and h. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 11 309–11 321.

[117] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, "Person re-identification: A retrospective on domain specific open challenges and future trends," *Pattern Recognition*, vol. 142, p. 109 669, Oct. 1, 2023, ISSN: 0031-3203.

[118] L. Zheng et al., "MARS: A video benchmark for large-scale person re-identification," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 868–884, ISBN: 978-3-319-46466-4.

[119] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, *Jointly attentive spatial-temporal pooling networks for video-based person re-identification*, Sep. 29, 2017. arXiv: 1708.02286[cs,stat].

[120] J. Meng, S. Wu, and W.-S. Zheng, "Weakly supervised person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 760–769.

[121] M. Liu, L. Qu, L. Nie, M. Liu, L. Duan, and B. Chen, "Iterative local-global collaboration learning towards one-shot video person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 9360–9372, 2020, ISSN: 1057-7149, 1941-0042.

[122] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1770–1782, 2019.

[123] D. Mekhazni, M. Dufau, C. Desrosiers, M. Pedersoli, and E. Granger, "Camera alignment and weighted contrastive learning for domain adaptation in video person ReID," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 1624–1633, ISBN: 978-1-6654-9346-8.

[124] F. Zhang, F. Chen, Z. Su, and J. Wei, "Unsupervised domain adaptation via dynamic clustering and co-segment attentive learning for video-based person re-identification," *IEEE Access*, vol. 12, pp. 29 583–29 595, 2024, ISSN: 2169-3536.

[125] A. Alsehaim and T. P. Breckon, "Vid-trans-reid: Enhanced video transformers for person re-identification," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2022, Paper 342.

[126] H. Ma, C. Zhang, E. Ning, and C. W. Chuah, "Temporal motion and spatial enhanced appearance with transformer for video-based person ReID," *Knowledge-Based Systems*, vol. 317, p. 113 461, May 2025, ISSN: 09507051.

[127] N. Dong, S. Yan, H. Tang, J. Tang, and L. Zhang, "Multi-view information integration and propagation for occluded person re-identification," *Information Fusion*, vol. 104, p. 102 201, Apr. 2024, ISSN: 15662535.

[128] K. Khaldi, V. D. Nguyen, P. Mantini, and S. Shah, "Unsupervised person re-identification in aerial imagery," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 260–269.

[129] V. D. Nguyen, K. Khaldi, D. Nguyen, P. Mantini, and S. Shah, "Contrastive viewpoint-aware shape learning for long-term person re-identification," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1041–1049.

[130] Y. Ma et al., "Grade-skewed domain adaptation via asymmetric bi-classifier discrepancy minimization for diabetic retinopathy grading," *IEEE Transactions on Medical Imaging*, vol. 44, no. 3, pp. 1115–1126, Mar. 2025, ISSN: 1558-254X.

[131] A. Chouchane, M. Bessaoudi, H. Kheddar, A. Ouamane, T. Vieira, and M. Hassaballah, "Multilinear subspace learning for person re-identification based fusion of high order tensor features," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107 521, Feb. 1, 2024, ISSN: 0952-1976.

[132] A. A. Gharbi, A. Chouchane, A. Ouamane, E. O. Belabbaci, Y. Himeur, and S. Bourennane, "A hybrid multilinear-linear subspace learning approach for enhanced person re-identification in camera networks," *Expert Systems with Applications*, vol. 257, p. 125 044, Dec. 2024, ISSN: 09574174.

[133] X. Wang, M. Liu, F. Wang, J. Dai, A.-A. Liu, and Y. Wang, "Relation-Preserving Feature Embedding for Unsupervised Person Re-Identification," *IEEE Transactions on Multimedia*, vol. 26, pp. 714–723, 2024, ISSN: 1941-0077.

[134] C. Zhang, Y. Su, N. Wang, Y. Lan, T. Wang, and A. Li, "Dual representation modeling and progressive contrastive learning for unsupervised video person re-identification," en, *Neurocomputing*, vol. 645, p. 130 467, Sep. 2025, ISSN: 09252312.

[135] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3415–3424.

[136] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," *arXiv:1808.07301 [cs]*, Aug. 22, 2018. arXiv: 1808.07301.

[137] Y. Yang, L. Li, H. Dong, G. Liu, X. Sun, and Z. Liu, "Progressive unsupervised video person re-identification with accumulative motion and tracklet spatial–temporal correlation," *Future Generation Computer Systems*, vol. 142, pp. 90–100, May 1, 2023, ISSN: 0167-739X.

[138] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 562–572.

[139] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 2014–2023, ISBN: 978-1-6654-4509-2.

[140] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[141] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10 012–10 022.

[142] V. Fortuin, "Priors in bayesian deep learning: A review," *International Statistical Review*, vol. 90, no. 3, pp. 563–591, 2022, ISSN: 1751-5823, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12502.

[143] B.-H. Tran, S. Rossi, D. Milios, and M. Filippone, "All you need is a good functional prior for bayesian deep learning," *Journal of Machine Learning Research*, vol. 23, no. 74, pp. 1–56, 2022.

[144] D. Ulmer, C. Hardmeier, and J. Frellsen, *Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation*, Mar. 7, 2023. arXiv: 2110.03051[cs].

[145] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 1, 2021, pp. 10 347–10 357, ISSN: 2640-3498.

[146] J. Yang et al., *Focal self-attention for local-global interactions in vision transformers*, Jul. 1, 2021. arXiv: 2107.00641[cs].

[147] X. Zang, G. Li, and W. Gao, "Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8776–8785, Dec. 2022, ISSN: 1941-0050, Conference Name: IEEE Transactions on Industrial Informatics.

[148] X. Liu, P. Zhang, C. Yu, X. Qian, X. Yang, and H. Lu, "A Video Is Worth Three Views: Trigeminal Transformers for Video-Based Person Re-Identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 12 818–12 828, Sep. 2024, ISSN: 1558-0016.

[149] X. Liu, P. Zhang, and H. Lu, "Video-Based Person Re-Identification with Long Short-Term Representation Learning," en, in *Image and Graphics*, H. Lu et al., Eds., Cham: Springer Nature Switzerland, 2023, pp. 55–67, ISBN: 978-3-031-46305-1.

[150] X. Liu, C. Yu, P. Zhang, and H. Lu, *Deeply-coupled convolution-transformer with spatial-temporal complementary learning for video-based person re-identification*, 2023. arXiv: 2304.14122 [cs.CV].

[151] P. Wu, L. Wang, S. Zhou, G. Hua, and C. Sun, "Temporal correlation vision transformer for video person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 6083–6091, 2024.

[152] H. Nguyen, K. Nguyen, A. Pemasiri, F. Liu, S. Sridharan, and C. Fookes, "Ag-vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[153] K. Nguyen et al., *Ag-vpreid 2025: Aerial-ground video-based person re-identification challenge results*, 2025. arXiv: 2506.22843 [cs.CV].

[154] D. Fu et al., "Large-scale pre-training for person re-identification with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[155] C. Yu et al., "Learning text-free clip for video-based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[156] I. H. Kim et al., *Pose-dive: Pose-diversified augmentation with diffusion model for person re-identification*, 2024. arXiv: 2406.16042 [cs.CV].

[157] K. Niu, H. Yu, X. Qian, T. Fu, B. Li, and X. Xue, *Synthesizing efficient data with diffusion models for person re-identification pre-training*, 2024. arXiv: 2406.06045 [cs.CV].

[158] T. Mamedov, D. Kuplyakov, and A. Konushin, "Approaches to improve the quality of person re-identification for practical use," *Sensors*, vol. 23, no. 17, p. 7382, 2023.

[159] F. Ke et al., *Explain before you answer: A survey on compositional visual reasoning*, 2025. arXiv: 2508.17298 [cs.CV].

[160] F. Lin, *Vision language models: A survey of 26k papers*, 2025. arXiv: 2510.09586 [cs.CV].

[161] Y. Yan et al., "Learning multi-granular hypergraphs for video-based person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 2896–2905, ISBN: 978-1-7281-7168-5.

[162] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification," *arXiv:2003.12224 [cs]*, Mar. 26, 2020. arXiv: 2003.12224.

[163] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 3286–3296, ISBN: 978-1-7281-7168-5.

[164] G. Chen, Y. Rao, J. Lu, and J. Zhou, "Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?" In *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm,

Eds., vol. 12353, Cham: Springer International Publishing, 2020, pp. 660–676, ISBN: 978-3-030-58598-3.

[165] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," *arXiv:2007.09357 [cs]*, Jul. 18, 2020. arXiv: 2007.09357.

[166] Z. Chen, A. Li, S. Jiang, and Y. Wang, "Attribute-aware identity-hard triplet loss for video-based person re-identification," *arXiv:2006.07597 [cs]*, Jun. 13, 2020. arXiv: 2006.07597.

[167] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 13 329–13 338, ISBN: 978-1-6654-4509-2.

[168] J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun, "Spatial-temporal correlation and topology learning for person re-identification in videos," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 4368–4377, ISBN: 978-1-6654-4509-2.

[169] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," *arXiv:2107.11878 [cs]*, Aug. 14, 2021. arXiv: 2107.11878.

[170] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Dense interaction learning for video-based person re-identification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 1490–1501. arXiv: 2103.09013.

[171] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, "Pyramid spatial-temporal aggregation for video-based person re-identification," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12 026–12 035.

[172] J. Wu et al., "CAViT: Contextual alignment vision transformer for video object re-identification," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 549–566, ISBN: 978-3-031-19781-9.

[173] S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen, "Salient-to-broad transition for video person re-identification," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7339–7348.

[174] K. Lee, I.-S. Jang, K.-J. Kim, and P.-K. Kim, "REET: Region-enhanced transformer for person re-identification," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov. 2022, pp. 1–8.

[175] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[176] Y. DENG, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, ser. MM '14, New York, NY, USA: Association for Computing Machinery, Nov. 3, 2014, pp. 789–792, ISBN: 978-1-4503-3063-3.

[177] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds.,

ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 17–35, ISBN: 978-3-319-48881-3.

[178] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 688–703, ISBN: 978-3-319-10593-2.

[179] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018. arXiv: 1711.08766.

[180] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3958–3967.

[181] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *2007 IEEE 11th International Conference on Computer Vision*, Oct. 2007, pp. 1–8, ISSN: 2380-7504.

[182] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3701–3711. arXiv: 1905.00953.

[183] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3749–3758. arXiv: 1903.09776.

[184] S. Mittal, "A survey on optimized implementation of deep learning models on the NVIDIA jetson platform," *Journal of Systems Architecture*, vol. 97, pp. 428–442, Aug. 1, 2019, ISSN: 1383-7621.

[185] J. Li, C. Feng, Y. Li, W.-l. Fan, L. Qin, and Q. Zhao, "Width-resolution multiplier lightweight network for person re-identification," in *International Conference on Graphic and Image Processing*, 2023.

[186] X. Zhang et al., "A lightweight approach to optimizing computational efficiency in multi-source domain adaptation for pedestrian re-identification," *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms*, 2024.

[187] Q. Liu et al., "A lightweight multi-scale feature enhancement network for person re-id," *Expert Systems*, vol. 42, no. 12, e70165, 2025.

[188] W. Yin et al., "Cvnet: Lightweight cross-view vehicle reid with multi-scale localization," *Sensors*, vol. 25, no. 9, p. 2809, 2025.

[189] G. M. M. E Elahi and Y.-H. Yang, "Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition," *Pattern Recognition*, vol. 122, p. 108 273, 2022, ISSN: 0031-3203.

[190] K. Muhammad, T. Hussain, J. Del Ser, V. Palade, and V. H. C. de Albuquerque, "DeepReS: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 5938–5947, Sep. 2020, ISSN: 1941-0050, Conference Name: IEEE Transactions on Industrial Informatics.

[191] A. Gutierrez-Torre et al., "Automatic distributed deep learning using resource-constrained edge devices," *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 15 018–15 029, 2022.

[192] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," *arXiv:1704.03373 [cs]*, Apr. 11, 2017. arXiv: 1704.03373.

[193] L. Chen, H. Yang, and Z. Gao, "Joint attentive spatial-temporal feature aggregation for video-based person re-identification," *IEEE Access*, vol. 7, pp. 41 230–41 240, 2019, ISSN: 2169-3536, Conference Name: IEEE Access.

[194] P. Ma, Y. Zhou, Y. Lu, and W. Zhang, *Learning efficient video representation with video shuffle networks*, Nov. 2019. arXiv: 1911.11319 [cs.CV].

[195] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4373–4382.

[196] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, *TransReID: Transformer-based object re-identification*, Mar. 26, 2021. arXiv: 2102.04378[cs].

[197] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "AlignedReID++: Dynamically matching local information for person re-identification," *Pattern Recognition*, vol. 94, pp. 53–61, Oct. 1, 2019, ISSN: 0031-3203.

[198] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, Oct. 15, 2018, pp. 274–282. arXiv: 1804.01438[cs].

[199] Y. Liu, C. Wang, M. Lu, J. Yang, J. Gui, and S. Zhang, "From Simple to Complex Scenes: Learning Robust Feature Representations for Accurate Human Parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5449–5462, Aug. 2024, ISSN: 1939-3539.

[200] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[201] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2019. arXiv: 1903.07071.

[202] S. Bai, P. Tang, P. H. S. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 740–749.

[203] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13 001–13 008, Apr. 3, 2020, ISSN: 2374-3468, Number: 07.

[204] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 3723–3732, ISBN: 978-1-5386-6420-9.

[205] 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, no. 1, pp. 151–175, May 1, 2010, ISSN: 1573-0565.

[206] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.