

UNIVERSITY OF OTTAWA

DOCTORAL THESIS

Thermal-RGB Sensory Data for Reliable and Robust Perception

Author: [Wassim El Ahmar](#)

© Wassim El Ahmar, Ottawa, Canada, 2023

A Doctoral thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

Faculty of Engineering

School of Electrical Engineering and Computer Science

November 29, 2023

UNIVERSITY OF OTTAWA

Abstract

Faculty of Engineering

School of Electrical Engineering and Computer Science

Doctor of Philosophy

Thermal-RGB Sensory Data for Reliable and Robust Perception

by Wassim EL AHMAR

The significant advancements and breakthroughs achieved in Machine Learning (ML) have revolutionized the field of Computer Vision (CV), where numerous real-world applications are now utilizing state-of-the-art advancements in the field. Advanced video surveillance and analytics, entertainment, and autonomous vehicles are a few examples that rely heavily on reliable and accurate perception systems.

Deep learning usage in Computer Vision has come a long way since it sparked in 2012 with the introduction of Alexnet. Convolutional Neural Networks (CNN) have evolved to become more accurate and reliable. This is attributed to the advancements in GPU parallel processing, and to the recent availability of large scale and high quality annotated datasets that allow the training of complex models. However, ML models can only be as good as the data they train on and the data they receive in production. In real-world environments, a perception system often needs to be able to operate in different environments and conditions (weather, lighting, obstructions, etc.). As such, it is imperative for a perception system to utilize information from different types of sensors to mitigate the limitations of individual sensors.

In this dissertation, we focus on studying the efficacy of using thermal sensors to enhance the robustness of perception systems. We focus on two common vision tasks: object detection and multiple object tracking. Through our work, we prove the viability of thermal sensors as a complement, and in some scenarios a replacement, to RGB cameras. For their important applications in autonomous vehicles and surveillance, we focus our research on pedestrian and vehicle perception. We also introduce the world's first (to the best of our knowledge) large scale dataset for pedestrian detection and tracking including thermal and corresponding RGB images.

Keywords: machine learning, artificial intelligence, object detection, multiple object tracking, perception

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Robert Laganière for trusting my abilities and believing in my potential. The time I have spent under his supervision as a member of the VIVA Research lab has been extremely beneficial for me. I have learned a lot from him and gained vital research and industrial experience under his supervision.

I also thank my family, my parents and siblings, who have been supportive of me for many years, both financially and emotionally. My wife Aya, you have been an outstanding support for me throughout this journey, and I will forever be grateful.

Finally, I would like to thank all my friends and colleagues, in particular, the Dhanvin Kolhatkar, Yahya Massoud, Hamzah AlGhamdi, and Farzan Erlik Nowruzi for their help and support.

Contents

Abstract	i
Acknowledgements	iii
1 Introduction	1
1.1 RGB Sensors	3
1.2 Thermal Sensors	5
1.3 Research Objectives	9
1.4 Research Hypothesis	9
1.5 Contributions	10
1.5.1 Datasets	10
1.5.2 Comparative experiments and analysis	11
1.5.3 Novel models and algorithms	11
1.5.4 Publications	12
2 Literature Review	13
2.1 Thermal Imaging	13
2.2 Multi-Modal Sensor Fusion	15
2.3 Thermal-RGB Datasets	18
2.4 Object Detection	20
2.4.1 Feature Extractors	23
ResNet	23
ConvNext	23
2.4.2 Detection Heads	24

Task-aligned One-stage Object Detection (TOOD) . . .	24
VarifocalNet (VFNET)	24
2.5 Multiple Object Tracking	25
2.5.1 ByteTrack	31
2.5.2 Observation-Centric SORT (OC-SORT)	32
2.6 Detection and Tracking Using Thermal Sensors	33
2.7 Conclusion	34
3 RGB and Thermal sensor performance in MOT	35
3.1 Introduction	36
3.1.1 MOT Applications	36
3.1.2 Two-stage MOT models	37
3.1.3 Perception Systems and Thermal Sensors	38
3.1.4 Chapter Summary	39
3.2 Proposed Approach and Experiments	39
3.2.1 Datasets	40
City Scene RGB-Thermal MOT Dataset	40
FLIR ADAS dataset	41
3.2.2 Object Detection Experiments	41
3.2.3 MOT Experiments	43
3.2.4 Dynamic Confidence Cut-Off (DCC)	43
3.3 Results and Analysis	44
3.3.1 Precision Recall of Object Detectors	44
3.3.2 MOT Metrics	47
3.3.3 Speed	50
3.3.4 Limitations and Failure Cases	50
3.4 Conclusion	51
4 Enhanced RGB-Thermal Sensor Fusion for Robust Perception	53
4.1 Introduction	53

4.1.1	2D Object Detection	55
4.1.2	Computer Vision Systems	55
4.1.3	Chapter Contributions	56
4.2	Enhancing Our City Scene Dataset	57
4.3	RGB-T Fusion Experiments	59
4.3.1	Architectural Overview	59
4.3.2	Fusion Operators	60
4.3.3	Enhanced Sigmoid Gating	61
4.4	Experiments	63
4.4.1	Experimental Setting	64
4.4.2	Evaluation of Early Sensor Fusion	65
4.4.3	Evaluation of Sigmoid Gating	68
4.4.4	Qualitative Analysis	69
4.4.5	Inference Time and Benchmarking	70
4.5	Conclusion	72
5	Development of Thermal MOT Model	77
5.1	Introduction	77
5.1.1	Two-Stage Multiple Object Trackers	78
5.2	Implementation details	81
5.2.1	OpenMMLab	81
	MMDetection [17]	81
	MMTracking [20]	81
5.2.2	Data Collection	82
5.2.3	Training Object Detection Models for two-stage MOT	84
5.2.4	Benchmarking ByteTrack and OCSORT on the RGB-Thermal MOT dataset	85
5.2.5	Development of new MOT box association method utilizing thermal information	85

Algorithm Details	86
Algorithm Summary	87
5.3 Results and Analysis	89
5.3.1 Comparing MOT models running on RGB against Thermal images	89
5.3.2 Weighted average alpha value selection	90
5.3.3 MOT Metrics Comparison	93
5.4 Conclusion	99
6 Conclusion and Future Work	101
6.1 Key Findings and Contributions	103
6.2 Implications	104
6.3 Future Work	105
Bibliography	107

List of Figures

1.1	Images of a scene taken using an RGB camera (left) and thermal camera (right). The pedestrian and vehicle are almost entirely non visible in the RGB image, while they are clearly visible in the thermal image.	7
1.2	Images of a scene taken using an RGB camera (left) and thermal camera (right). The thermal image is not at all affected with the glare from the sun, while the RGB camera is significantly affected and the truck can barely be seen.	8
2.1	Capturing system	19
2.2	Sensor plastic enclosing	20
2.3	In object detection, the model is trained to predict the class and location of every object of interest in an image.	21
2.4	In MOT, in addition to predicting the class and location of objects of interest, the model must also maintain the same class ID for the same object across different frames. This allows the re-identification of the object as it moves throughout a video sequence. This figure shows the workflow of two stage MOT where the first step is to detect the objects, and the second step conducts box association between bounding boxes of the previous frame and bounding boxes of the current frame.	27
2.5	Variable appearance of pedestrians captured using RGB sensors.	28
2.6	Inter-class occlusion (left) and intra-class occlusion (right). . .	29

3.1	Samples from the three testing sequences of our dataset taken during different times of the day.	40
3.2	TOOD Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box. The precision/recall value pairs are calculated using detection threshold values between 0 and 1, with 0.05 increments.	45
3.3	VFNET Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box.	46
3.4	Samples showing missing annotations from the FLIR ADAS Dataset. Green rectangles represent ground truth boxes, red rectangles represent detected boxes, orange arrows point to instances where a pedestrian is detected by a model but no ground truth annotation exists.	51
4.1	Original RGB and corresponding thermal images from the City Scene RGB-Thermal MOT Dataset. The original images are un-registered and have different resolutions.	59
4.2	Matched keypoints between the RGB and thermal images using the ORB feature extractor [99]. The green lines represent the matched keypoints between the RGB and thermal images.	59
4.3	Registered Thermal and RGB images after applying the transformation homography to the RGB image. The labelled objects in red are ground truth annotations labelled on the thermal images.	60

4.4	Left: A diagram illustrating the bilaterally-guided fusion "BGF" operator [92]. Right: A diagram illustrating the multi-modal factorized bilinear pooling "MFB" fusion operator [79]. The input to the fusion operators are the raw modality stream (RGB or thermal image), and the Fused output is then processed by the feature extractor.	62
4.5	Proposed sigmoid-gating for early RGB-thermal fusion.	64
4.6	Precision-Recall curves for TOOD with ConvNeXt as a backbone network. All plots aim to contrast the detection accuracy of fusion operators versus baseline models (RGB-only and thermal-only).	73
4.7	Precision-Recall curves for TOOD with ResNet-50 as a backbone network. All plots aim to contrast the detection accuracy of fusion operators versus baseline models.	74
4.8	A day frame from the City Scene [27] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N . Confidence scores (%) of bounding boxes are positioned to the left of each box's index.	75
4.9	A night frame from City Scene [27] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N . Confidence scores (%) of bounding boxes are positioned to the left of each box's index.	76
5.1	Sample occlusion case from our collected dataset	80

5.2	RGB and corresponding thermal images of the RGB-Thermal MOT dataset.	83
5.3	MOTA and IDF1 values of our proposed box association method applied on OCSORT on the validation sequences of the RGB-Thermal MOT dataset using different values of alpha.	92
5.4	MOTA and IDF1 values of our proposed box association method applied on ByteTrack method on the validation sequences of the RGB-Thermal MOT dataset using different values of alpha.	93
5.5	Overall comparison of the important MOT metrics running standard motion box association against our proposed box association algorithm using ByteTrack.	97
5.6	Overall comparison of the important MOT metrics running standard motion box association against our proposed box association algorithm using OCSort.	99

List of Tables

2.1	FLIR ADK technical specifications	19
2.2	JAI GO 5100-C technical specifications	19
3.1	Training hyper-parameters	42
3.2	Object detection experiments conducted	42
3.3	Overall ByteTrack with dynamic cut-off results on the 3 sequences of our City Scene RGB-Thermal MOT dataset. a) First test sequence captured during daytime. b) Second test sequence captured during daytime. c) Third test sequence taken at night. d) Overall performance across all 3 testing sequences.	49
3.4	Comparison of the performance of the trackers before and after the use of a DCC on our City Scene RGB-Thermal MOT dataset.	49
3.5	Benchmarking tracker speed on the AWP machine and the Jetson Xavier module.	50
4.1	Contrasting 2D detection accuracy, i.e. average precision, of four different early fusion operators.	64
4.2	Demonstrating the performance boost that results from adding the sigmoid gated mechanism to the fusion operators. The experiments are conducted on the TOOD object detector with two different backbone networks. Experiments are compared based on their average precision (AP_{50}).	69

4.3	Inference time and benchmarking results on NVIDIA GeForce RTX 3090, Jetson Xavier, and Jetson TX2. The reported results include inference time in seconds (<i>s</i>) and the total number of frames-per-second (<i>fps</i>).	71
5.1	RGB-Thermal MOT dataset statistics.	84
5.2	Results of benchmarking Bytetrack on the RGB validation sequence using standard implementations of Bytetrack (top), and OCSORT (bottom).	90
5.3	Results of benchmarking Bytetrack on the thermal validation sequence using Bytetrack standard motion association (top), and using our novel thermal-motion box association method (bottom).	96
5.4	Results of benchmarking OCSort on the thermal validation sequence using OCSort standard motion association (top), and using our novel thermal-motion box association method (bottom).	98

List of Abbreviations

CNN	Convolutional Neural Networks
ML	Machine Learning
GPU	Graphical Processing Unit
CV	Computer Vision
ADAS	Advanced Driver Assistance Systems
RGB	Red Green Blue
MOT	Multiple Object Tracking
IR	Infra Red
NIR	Near Infra Red
SWIR	Short Wave Infra Red
MWIR	Mid Wave Infra Red
LWIR	Long Wave Infra Red
FIR	Far Infra Red
FPS	Frames Per Second
ROI	Region Of Interest
FOV	Field Of View
SOTA	State Of The Art
DCC	Dynamic Confidence Cutoff
SGD	Stochastic Gradient Descent
MOTA	Multiple Object Tracking Accuracy
IDF	IDentity F1
IOU	Intersection Over Union
CVPR	Computer Vision and Pattern Recognition

HCI	Human Computer Interaction
VR	Virtual Reality
MT	Mostly Tracked
PT	Partially Tracked
ML	Mostly Lost
FP	False Positives
FN	False Negatives
AP	Average Precision

Chapter 1

Introduction

The rapid advancement in machine learning has sparked intensive research in the field of Computer Vision (CV). It has led to a significant leap in the accuracy and reliability of machine vision models. We now rely on this advanced research in our daily lives, even if we don't realize it. For instance, the face unlock feature on our smartphones rely on advanced CV algorithms to reliably detect and identify the user face. Surveillance cameras utilize algorithms that detect humans and other objects of interest to notify the homeowners of any suspicious events. Most newer generation vehicles are equipped with Advanced Driver Assistance Systems (ADAS) with features like forward collision prevention, lane keep assist, and blind spot checking to make roads safer for drivers and pedestrians.

The sample use cases mentioned above are just a few examples of the utilization of state-of-the-art CV research in our daily lives. Significant research has been conducted, especially in the last decade with the advancement of deep learning methods, on improving the efficiency, accuracy, and reliability of CV algorithms. However, for CV systems deployed in real-world environments, there is a strong requirement that these systems function reliably in different weather, lightning, and other environmental conditions. This is why CV systems often require data from different types of sensors to ensure reliable performance.

The mentioned use cases are just a few examples of how advanced CV research impacts our daily lives. In the past decade, CV research has made significant progress, thanks to improved deep learning methods. These advancements have made CV algorithms more efficient, accurate, and reliable. However, in real-world situations, CV systems must perform well in different weather conditions, lighting, and various environments. That's why CV systems often rely on data from various sensors to ensure they work properly. These sensors include thermal cameras, LiDAR, radar, and GPS, which help CV systems adapt and perform effectively in different scenarios.

This approach is essential to make CV systems resilient in diverse conditions, from bright sunlight to foggy weather, and in various settings, from busy cities to remote areas. In essence, integrating sensor data and using complementary technologies are crucial to maximize the potential of CV in practical applications. This highlights the ongoing importance of CV research in changing how we experience and understand the world.

In this thesis, we investigate the effectiveness of using thermal sensors in two important tasks in CV: Object Detection and Multiple Object Tracking (MOT). We study the benefits of fusing RGB and thermal data to enhance robustness of perception systems, and provide detailed analysis over the strengths and weaknesses of each sensor. Our experiments will be conducted on a new dataset that we are contributing to the scientific community: the world's first dataset of thermal and corresponding RGB images annotated for multiple object detection and tracking.

The organization of this chapter is as follows. We first discuss the limitations of RGB camera perception in section 1.1. We then briefly introduce

thermal sensors in section 1.2.

1.1 RGB Sensors

An RGB sensor (color camera) is one of the most popular and common perception sensors. It is integrated into most smartphones, tablets, computers, and has many applications in our daily lives.

RGB cameras have several advantages, including:

- **Human-Centric Visual Perception:** RGB cameras excel in capturing images that closely align with human visual perception. The images they produce are familiar and visually pleasing to the human eye. This similarity to human vision facilitates easy interpretation of the context and semantic information within a color image. This makes RGB imagery highly intuitive and accessible for humans, which is especially valuable in applications where human interaction and understanding are crucial, such as in augmented reality, human-computer interaction, and content creation.
- **Color Information:** RGB cameras provide rich color information about objects present in an image. By capturing the red, green, and blue channels, they offer a comprehensive representation of the visual world. This color information is valuable in numerous applications, from computer vision and image analysis to multimedia content creation, where precise color rendering and discrimination are essential. It allows for accurate identification of objects based on their color characteristics, aiding in tasks like object recognition and tracking.
- **High Resolution:** Generally, RGB cameras are available in a wide range of resolutions, including high-resolution options. This versatility makes

them suitable for various applications, from capturing detailed scenes to identifying fine-grained features.

- **Cost Efficiency:** RGB cameras tend to be cost-effective compared to some specialized sensors like multispectral or hyperspectral cameras. This cost-efficiency makes them accessible for a broader range of applications and industries, enabling the widespread adoption of computer vision and image analysis technologies.

On the other hand, RGB sensors are known to have several limitations, some of which are given below:

- **Sensitivity to Light:** One of the most significant limitations of RGB sensors in real-world systems is their sensitivity to different lighting conditions [47, 113, 49, 124, 114]. Their performance significantly deteriorates under poor lighting conditions. This sensitivity poses a critical drawback in real-world applications, such as smartphone face unlock features, where the device must reliably identify a person even in sub-optimal lighting conditions.
- **Sensitivity to Weather Conditions:** RGB sensors are also susceptible to adverse weather conditions, including fog, rain, and snow, which can cause the camera lens to become blurry [89, 100, 83]. In such conditions, the quality of RGB image capture can degrade, impacting the accuracy and reliability of computer vision applications that rely on clear visual data.
- **Occlusion:** RGB sensors capture color information about objects in a

scene. However, they face challenges when objects are fully or partially occluded [76, 65, 82]. Extracting information about obscured objects can be complicated, and this limitation becomes particularly apparent when objects have similar color features. For instance, in scenarios where school students wear identical uniforms, distinguishing between individuals based solely on color can be challenging.

Simply put: A sensor can't perceive what it can't see, and the perception ability of an RGB sensor is significantly affected by environmental factors.

1.2 Thermal Sensors

RGB sensors function through capturing reflected light in the visual-optical spectrum. This implies, that for visual spectrum sensitive cameras, the scene must be illuminated by a source of light (either natural or artificial). This is one of the main limitations of RGB sensors as discussed in section 1.1.

All objects around us emit heat radiations, which were first discovered by Sir William Herchel in 1800 [106]. The human eye can perceive light in the range between 380 and 700nm. As heat radiations have a much higher wavelength ranging between 0.73 to 1,000 μm [106], they can't be captured by the human eye. Infrared sensors are capable of capturing the heat radiation emitted from objects. The IR spectrum is composed of several bands:

- Near Infrared (NIR): Wavelength range between 800 and 2,500 nm
- Short Wave Infrared (SWIR): Wavelength range between 0.9 and 1.7 μm
- Mid Wave Infrared (MWIR): Wavelength range between 3 and 5 μm
- Long Wave Infrared (LWIR): Wavelength range between 8 and 14 μm
- Far Infrared (FIR): Wavelength range between 15 μm and 1mm

MWIR and LWIR IR sensors are commonly referred to as thermal IR sensors, or thermal cameras. A thermal sensor is a useful sensor in autonomous systems as it does not require any light source for its operation. For critical systems like autonomous vehicles for example, where RGB sensors would poorly perceive vulnerable objects at night (pedestrians, cyclists, etc.), a thermal sensor would be able to provide better perception of the scene as it operates normally when there is no light sources.

Some of the advantages of using thermal sensors are given below:

- **Robust Performance in Challenging Lighting Conditions:** Thermal sensors operate effectively under poor lighting conditions and glare. As shown in Figures 1.1 and 1.2, thermal sensors exhibit remarkable robustness compared to RGB sensors in handling different lighting conditions and glare¹. This capability makes them invaluable for applications where consistent performance in varying light environments is essential.
- **Resilience to Environmental Factors:** Thermal sensors are less affected by environmental factors like dust, fog, and snow in comparison to RGB cameras. Their ability to maintain functionality in adverse weather conditions makes them a preferred choice for applications that demand reliability regardless of the weather.
- **Homogeneous Object Appearance:** Unlike RGB cameras, which are sensitive to object texture, color, and patterns, thermal cameras provide a homogeneous appearance for objects with consistent thermal values. This attribute is particularly advantageous in tasks such as pedestrian detection. In the visible spectrum, training convolutional

¹Image samples from the Teledyne FLIR ADAS Thermal Dataset <https://www.flir.ca/oem/adas/adas-dataset-form/>

neural networks (CNNs) to reliably detect pedestrians requires a diverse dataset encompassing variations in size, clothing, and skin color. In contrast, thermal images exhibit less variance in the appearance of humans, making detection based on the human thermal signature more consistent and reliable [15].

- **Flexible Data Representation:** Thermal data is typically captured as a 16-bit single-channel (grayscale) image, where each pixel represents a thermal intensity. This simplicity allows for the application of various color palettes to better visualize the thermal scene. This flexibility is advantageous as it enables the training of CNNs on thermal images without necessitating changes to the architecture. It simplifies the integration of thermal data into existing computer vision workflows.



FIGURE 1.1: Images of a scene taken using an RGB camera (left) and thermal camera (right). The pedestrian and vehicle are almost entirely non visible in the RGB image, while they are clearly visible in the thermal image.

On the other hand, some disadvantages of thermal sensors include:

- **Lower Resolution:** Thermal sensors typically offer lower resolution compared to RGB sensors [128]. This reduced resolution can limit the



FIGURE 1.2: Images of a scene taken using an RGB camera (left) and thermal camera (right). The thermal image is not at all affected with the glare from the sun, while the RGB camera is significantly affected and the truck can barely be seen.

level of detail and clarity in thermal images, which may be a drawback in applications where precise object identification or fine-grained analysis is required.

- **Temperature Variance Requirement:** Thermal sensors capture thermal intensities in a scene. For effective thermal imaging, there needs to be a certain amount of variance in the temperatures of the objects of interest and the background in the scene. In scenarios where temperature differences are minimal or where objects have similar thermal signatures, thermal sensors may struggle to provide meaningful distinctions, potentially leading to limitations in detection and recognition tasks.
- **Higher Cost:** In comparison to RGB sensors, thermal sensors are generally more expensive to acquire and deploy [118]. The cost factor can be a significant consideration for organizations or projects with budget constraints. This increased expense can be a barrier to the widespread adoption of thermal imaging technology, especially in applications that require multiple sensors or large-scale deployments.

1.3 Research Objectives

Advanced machine vision systems are having a larger and more significant impact on our daily lives. Increasing the level of automation of our daily tasks is a core focus of emerging technology. The breakthrough that deep learning achieved in the field of machine vision allowed the field to be transformed from almost being purely theoretical to being more integrated in real-world applications. With many critical applications such as autonomous vehicles, medical imaging, video surveillance, and face identification security systems utilizing computer vision models, it becomes necessary to focus on enhancing the robustness, safety, reliability, and accuracy of such systems.

The research objectives of this thesis are summarized below:

- Study the efficacy of using thermal sensors for the tasks of object detection and MOT, and how the performance compares to that of RGB sensors.
- Gain a deeper understanding of the strengths and weaknesses of thermal sensors in different environmental conditions (temperature variance due to shade, performance during different times of the day).
- Study different approaches for fusing RGB and thermal data to leverage the strengths of both the visible and infrared spectrums, while mitigating the weaknesses.

1.4 Research Hypothesis

The use of thermal sensors in machine vision has been well studied. It is established that thermal sensors provide significant advantages over RGB

sensors in poor lighting and weather conditions.

However, till the time of writing of this thesis, there has not been to the best of our knowledge a clear research study that conclusively and thoroughly exploits the efficacy of thermal sensors in end-to-end pedestrian perception. This is partially attributed to the lack of a large-scale, comprehensive dataset annotated for detection and tracking of pedestrians in the thermal and visible spectrums.

Thereby, the central research hypothesis of this thesis is: *Leveraging the unique characteristics of thermal data, such as thermal signatures and reduced reliance on visible light conditions, will enhance the accuracy, robustness, and scalability of object detection and tracking algorithms.*

1.5 Contributions

The contributions of this thesis work can be classified into three categories: introduction of new benchmarks, comparative experiments and analysis, and novel models and algorithms.

1.5.1 Datasets

- We introduce the world’s first (to the best of our knowledge) large scale MOT dataset of RGB and corresponding thermal images for advanced, large scale research on multiple object detection and tracking for far objects (100 – 300 meter distances) in the visible and infrared spectrums. The images from both sensors are aligned to be in the same coordinate system, facilitating research in thermal-RGB sensor fusion.

- We introduce a second MOT dataset annotated for pedestrians at varying distances, taken at 5 different crossing locations in an urban environment. The dataset is expected to act as a benchmark for thermal multiple object tracking methods.

1.5.2 Comparative experiments and analysis

- We conduct a comprehensive set of experiments to analyse the strengths and weaknesses of the utilization of thermal sensors for robust object detection compared to RGB sensors. The analysis also exploits the effect of adverse weather and lighting conditions on the performance of RGB and thermal sensors.
- We conduct a comprehensive set of experiments to highlight the strength of existing sensor fusion methods for combining RGB and thermal images for object detection tasks.
- We compare the performance of MOT models operating in the thermal domain against MOT models operating in the RGB domain.

1.5.3 Novel models and algorithms

- We introduce a dynamic confidence cut-off (DCC) threshold mechanism for 2-stage object trackers that enhances the box association step by assigning a DCC for detected objects based on box area.
- We introduce a new method for RGB-thermal sensor fusion that achieves better results than existing fusion methods for the task of object detection.
- We introduce a new box association method that can be used with two-stage MOT models running in the thermal spectrum that utilizes the

unique characteristics of thermal data and combines them with motion data for robust MOT in the thermal spectrum.

1.5.4 Publications

- Multiple Object Detection and Tracking in the Thermal Spectrum [27]
- Enhanced Thermal-RGB Fusion for Robust Object Detection [26]
- Thermal-Motion Box Association for Robust Tracking in the Thermal Spectrum, paper finalized for submission to a conference.

Chapter 2

Literature Review

We utilize thermal sensors to enhance the perception ability of systems for the tasks of multiple object detection and tracking. In this chapter, we provide a literature review of the use of thermal sensors in machine vision, in addition to related work in the tasks of multiple object detection and tracking.

The remainder of this chapter is organized as follows: Section 2.1 provides an overview of the use of thermal sensors in machine vision literature. Section 2.2 elaborates on multi-modal sensor fusion. In section 2.3, we go over existing thermal-RGB datasets. We provide an overview of approaches to object detection and MOT in sections 2.4 and 2.5 respectively. Section 2.6 goes over several approaches for pedestrian detection and tracking that utilize thermal sensors.

2.1 Thermal Imaging

Image processing serves as a fundamental cornerstone in the realm of computer vision, with its overarching goal being the extraction of meaningful information from visual data, whether it's a single image or a sequence of images constituting a video stream. This information encompasses a wide array of elements, including the identification of objects, understanding the

context in which these objects exist, and even deciphering the actions or activities unfolding within the visual scene. By leveraging various techniques and algorithms, image processing endeavors to transform raw pixel data into structured knowledge, enabling machines to comprehend and interact with their visual environment effectively.

In this context, thermal sensors stand out as invaluable tools due to their ability to capture and represent data in a two-dimensional format denoted as $(width, height)$. This format closely parallels the representation commonly associated with RGB images, which consists of an array of pixels organized along two dimensions. Because of this similarity, thermal images offer a fertile ground for the application of a wide spectrum of image processing techniques, ranging from traditional computer vision methodologies to cutting-edge deep learning approaches.

In [133], Zuluaga-Gomez et al. highlight the efficacy of using thermal imaging for the reliable detection of breast cancer. The authors cite limitations such as expensive X-Ray equipment, social and cultural norms, and dense tissue formation during young age as a motive to experiment with thermography for breast cancer detection as an alternative to other techniques like ultrasound, magnetic resonance imaging, and mammographies. The authors train a CNN on a dataset of thermal images, and achieve an accuracy of 90% in predicting breast cancer. The authors also show that applying data augmentation on the thermal dataset achieves almost identical performance compared to training on a dataset with up to 50% more original images.

In [11] the authors argue that RGB hand gesture recognition is limited by lighting conditions. As a result, such systems perform poorly in scenarios where there is no or poor source of natural or artificial light. The authors

train a compact CNN model on a collected dataset of thermal images for hand gesture recognition. The authors achieve an accuracy of 99.98%, while showing that the compact CNN runs at around 12 frames-per-second (FPS) on a low power edge device.

In order to improve automation in the process of building inspections, Park et al. introduce a pipeline to detect thermal anomalies utilizing RGB and thermal sensors [91]. Heat anomalies could be a result of inadequate insulation applied, leaks, moisture, ventilation heat loss, or conductivity heat loss. The authors utilize RGB images to segment the wall region in an image using a CNN. Then, they use the segmented wall area to detect anomalies in the temperature distribution of the wall region.

2.2 Multi-Modal Sensor Fusion

Sensor fusion is an important concept in the field of data integration and computer vision, often employed to enhance the overall performance of systems that rely on information from multiple sensors [125, 77, 103, 68, 48]. While it is versatile and can accommodate various combinations of sensors, the most promising results typically emerge when data from complementary sensors are fused together. This notion hinges on the idea that different sensors can capture unique aspects of the environment, and their combined insights can offer a more comprehensive and accurate understanding of the world around us.

The fusion of sensor data can be categorized into three primary methods based on when it occurs within the network's architecture: early-fusion, mid-fusion, and late-fusion.

Early-fusion, also known as feature-level fusion, takes place at the initial stages of data processing. In this method, the raw data from multiple sensors are combined before any extensive analysis. The merged data may include information from various sensors, such as visual, thermal, or depth sensors. Early-fusion aims to create a unified representation of the environment, allowing subsequent processing stages to work with a single, integrated data stream. This approach can be advantageous when the sensors provide complementary information that can improve the accuracy and robustness of the system [66, 4, 85, 21, 109].

Mid-fusion, occurs after some initial processing has taken place on each sensor's data individually. Each sensor's output is processed independently to extract relevant features or make preliminary decisions. These intermediate results are then combined to make a final decision or inference. Mid-fusion is beneficial when sensors have distinct strengths and weaknesses, and processing their data separately can highlight their respective expertise [62, 90, 43, 121]. By fusing these intermediate decisions, the system can capitalize on the unique capabilities of each sensor while mitigating their limitations.

Late-fusion, or sensor-level fusion, is the fusion approach that happens at the final stages of processing. Here, each sensor's data is processed independently, and the results are maintained separately until the very end of the analysis pipeline. Only at this stage are the outcomes merged, typically through techniques like weighted averaging or voting, to arrive at a final decision or interpretation. Late-fusion is often employed when different sensors provide diverse perspectives or modalities of the same scene, and the system aims to maintain flexibility by postponing the fusion until it's most contextually relevant [3, 1, 32, 98, 75].

The Aggregate View Object Detection (AVOD) is a two-stage network proposed by Ku et al. [54]. The first stage uses an element-wise mean operation to fuse LiDAR and RGB features and generate 3D object proposals which are refined by the second stage. Liang et al. [64] use an element-wise add operation for their continuous [116] LiDAR-RGB fusion approach.

Massoud [79] implements learnable fusion from a LiDAR scanner and an RGB camera. The approach uses fusion methods with the LiDAR bird's eye view (BEV) representation, the LiDAR's frontal view (FV) representation, and RGB frames for the tasks of 2D, 3D, and BEV object detection. The method proposed for learnable fusion, Multi-modal Factorized Bilinear Pooling (MFB), is tested as an early and mid-fusion approach and is compared with the more straight-forward solutions of element-wise addition [64] and element-wise multiplication.

Pasandi et al. [92] expand upon this work by comparing the same set of fusion operations, but add in feature concatenation and their proposed Bilaterally-Guided Fusion (BGF). The authors compare these techniques as either early-fusion or mid-fusion in the context of the 2D object detection task on the KITTI dataset. They present a detailed analysis of performance and computational complexity.

As the work in [79, 92] shows that early fusion methods out-perform mid and late fusion approaches, we focus on early fusion in our work.

2.3 Thermal-RGB Datasets

Most existing tracking datasets that include thermal images are annotated for single-object tracking [72, 73, 61]. However, none of the existing thermal MOT datasets offer a sufficient volume of data for multi-object tracking for pedestrians. With limited sequences and annotated frames, these datasets are inadequate for training and evaluating MOT algorithms in complex real-world scenarios. Recognizing this limitation, we undertook the collection and annotation of two new, large-scale datasets containing both RGB and corresponding thermal data. This initiative addresses the need for a more extensive and diverse dataset collected in real-world scenarios, allowing for the development and assessment of MOT algorithms in challenging, real-world conditions in both the thermal and color spectrums. The first dataset is more suitable for RGB-Thermal fusion research, while the second is a large-scale dataset that can be used specifically for research on MOT in the thermal domain.

We utilize the same FLIR ADK thermal sensor and a JAI GO-5100C RGB sensor for the collection of both datasets. The specifications of both sensors are given in tables 2.1 and 2.2 respectively. A plastic mount was built to enclose and fix both sensors next to each other, and a tripod thread was installed on the plastic enclosure such that both sensors could be installed on a tripod for data collection. Figures 2.1 and 2.2 show the developed capturing setup and the sensor plastic enclosure.

Specification	Value
Horizontal FOV	75 °
Spectral Band	8-14 microns (LWIR)
Thermal Sensitivity	<50 mK
Power Consumption	4W average
Image Resolution	640 × 512

TABLE 2.1: FLIR ADK technical specifications

Specification	Value
Shutter type	Global shutter
Power Consumption	4.35W
Frame Rate	Up to 74 frames/sec

TABLE 2.2: JAI GO 5100-C technical specifications



FIGURE 2.1: Capturing system



FIGURE 2.2: Sensor plastic enclosing

2.4 Object Detection

Object detection is a field in CV that focuses on localizing objects of interest in an image. As opposed to image classification, where the model receives an input image and outputs a label corresponding to the class of the image, object detection focuses on determining where in an image an object of interest exists. Object detection allows the localization of multiple objects of possibly different classes in one image.

Training a deep-learning based object detector requires the collection and labelling of a dataset of images where for every image, the ground truth label represents the location (x,y, width, height) and class of every object present in the image (Figure 2.3).

Object detection is a core research area in computer vision for its numerous applications in real-world systems (surveillance, scene understanding, action recognition, autonomous systems, etc.). In addition, object detection is often the first step in the pipelines of Multiple Object Tracking (MOT) and re-identification models.

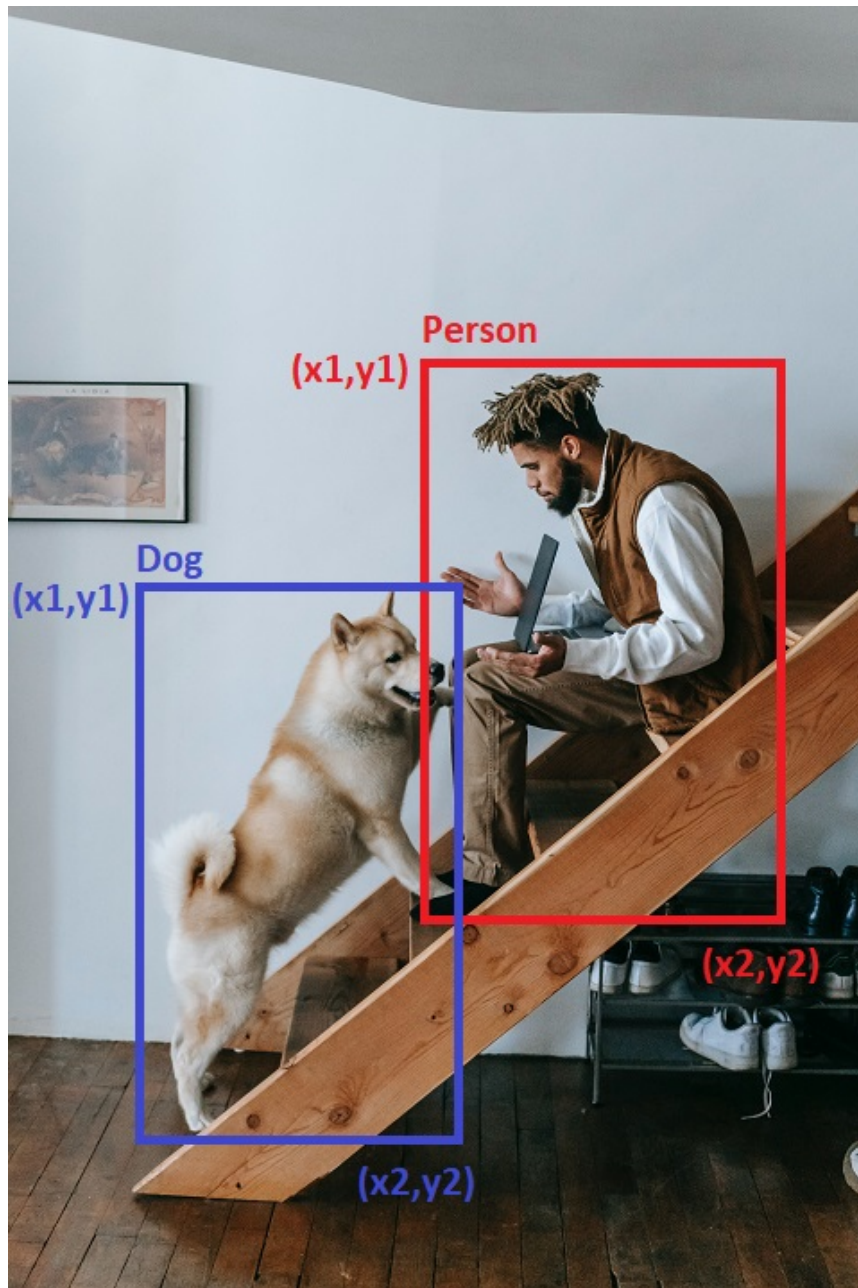


FIGURE 2.3: In object detection, the model is trained to predict the class and location of every object of interest in an image.

Object detection methods are divided into two categories: two-stage detection and one-stage detection. Methods in the former category use a first stage to generate object proposals that are refined and classified by a second stage [35, 34, 96, 12, 38, 25], while those in the one-stage category bypass the region proposal stage and generate predictions directly [95, 74, 71, 59]. One stage detectors have become the more popular and standard approach for 2D object detection as they are generally faster, and the accuracy of these models have become comparable to that of two stage approaches. Two stage detectors are generally composed of a feature extractor, and a detection head.

Two common metrics used to assess the performance of object detection models are precision and recall. Precision in object detection measures the accuracy of the identified positive instances, calculated as the ratio of true positive predictions to the total positive predictions. Recall, on the other hand, assesses the ability of the algorithm to capture all relevant positive instances by calculating the ratio of true positive predictions to the total actual positive instances in the dataset. Precision and recall are calculated using Equations 2.1 and 2.2 respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.2)$$

Where:

- **TP:** True Positives
- **FP:** False Positives
- **FN:** False Negatives

2.4.1 Feature Extractors

In general, any convolutional neural network can be used as a feature extractor for one-stage object detection models. The choice of feature extractor to be used depends on the application requirements of speed and accuracy. Larger, more sophisticated feature extractors generally extract more robust features that lead to better detection performance, at the expense of higher processing times. Some of the common feature extractors used as backbones for object detection models are described below.

ResNet

He et al. [39] introduced the residual block to enable the training of significantly deeper models than previously possible. The architecture of the residual block uses identity mappings as shortcuts around consecutive convolutional layers. The resulting residual network, or ResNet, is presented in multiple versions depending on its number of layers: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. Each version is composed of 5 stages, each composed of residual blocks of two or three convolutional layers.

ConvNext

Liu et al. [78] proposed ConvNeXt as an update of the ResNet architecture. They change training parameters (using the AdamW optimizer and new data augmentation techniques), the number of residual blocks in each stage (making the 3rd stage heavier and the other stages lighter), the structure of the first layer (using a smaller non-overlapping kernel), and how downsampling layers are implemented (between stages instead of at the end of each stage). Additionally, the authors use depth-wise convolutions, inverted bottleneck

design, and larger kernel sizes, replace ReLU activation functions with GELU and Batch Normalization with Layer Normalization, and reduce the number of each. All these design changes result in a significant increase in accuracy for multiple tasks.

2.4.2 Detection Heads

A detection head receives the processed feature map output by the convolutional feature extractor, and predicts the locations and corresponding classes of objects present in an image.

Task-aligned One-stage Object Detection (TOOD)

Feng et al. introduced the Task-aligned One-stage Object Detection (TOOD) [29] which strengthens the link between the localization and the classification tasks of object detection. This is accomplished by introducing "Task-Alignment": taking the network's outputs for each task and passing them to a network head that modifies the score and the location predictions to align their optimal anchors. Their Task Alignment Learning (TAL) also pushes the network to predict better aligned bounding boxes. TOOD achieved an AP of 51.1 on the MS-COCO dataset.

VarifocalNet (VFNET)

Zhang et al. designed VFNet using Varifocal Loss [127]. This loss is meant to maximize the IoU-aware classification score (IACS) that takes into account both the classification and the location of a prediction. The Varifocal loss also modifies focal loss by weighing positive examples more heavily than

negative ones. Additionally, VFNet uses a new nine-point deformable convolution representation for bounding boxes and a network head to refine the network's box predictions by learning an additional offset to their locations.

2.5 Multiple Object Tracking

The goal of object detection is to localize objects of interest in a scene by predicting the location of the objects' bounding boxes and the corresponding classes. The goal is to identify objects of interest in an image on a frame-by-frame basis. This means that if a video sequence is being processed, object detectors will process every frame independently and generate predictions over the locations and classes of objects present in each frame. Single object tracking aims at assigning an ID to an object, and consecutively re-identify the same object (same ID) in consecutive frames in a video sequence. MOT is considered a more complex task than object detection and single object tracking. MOT aims at assigning a unique ID for every object of interest present in a frame, and maintain the same ID in corresponding frames as the object moves in the video sequence (Figure 2.4).

In the context of MOT, a tracklet is defined as the set of detections having the same class and same unique ID. Ideally, a tracker would assign one ID for all detections of an object instance in a video sequence. However, since MOT is naturally a more difficult task than object detection, tracklet fragmentation often occurs where a tracker fails to maintain the same unique ID for an object, leading to the generation of multiple tracklets corresponding to the same object in a video sequence. MOT has important applications including trajectory prediction, video security and surveillance, virtual reality, human-computer interaction, and person re-identification.

MOT is a complex task that involves the continuous localization and association of multiple objects across video frames. While MOT has made significant strides in recent years, it still faces several challenges:

- **Prediction of Object Trajectories:** One of the primary challenges in MOT is accurately predicting the trajectory of moving objects. The motion of objects can be intricate and challenging to forecast, especially when objects exhibit abrupt changes in speed, direction, or behavior. Factors like occlusions and interactions between objects further complicate trajectory prediction.
- **Variability in Object Appearance (Figure 2.5):** Objects tracked in MOT systems often exhibit varying appearances as they move. For instance, when tracking a pedestrian, their facial features and clothing may appear quite different when facing the camera compared to when they have their back turned. Handling these appearance variations and maintaining accurate object associations is a persistent challenge in MOT.
- **Intermittent Object Occlusion (Figure 2.6):** In real-world video sequences, objects of interest can momentarily disappear from view due to occlusions, sudden obstructions, or temporary exits from the camera's field of view. These interruptions in object visibility can disrupt the tracking process and make it challenging to maintain a unique object identity over time.
- **Dependency on Detector Accuracy:** Many state-of-the-art MOT systems rely on two-stage detection-based trackers, where the first stage involves object detection followed by tracking. The performance of such trackers is heavily contingent on the accuracy of the object detector. If the detector fails to detect an object in a particular frame, the tracker will be unable to initiate tracking for that object. This dependence on the detector's precision introduces vulnerabilities in MOT

systems, especially when dealing with challenging detection scenarios, such as low-resolution images or objects with complex appearances.

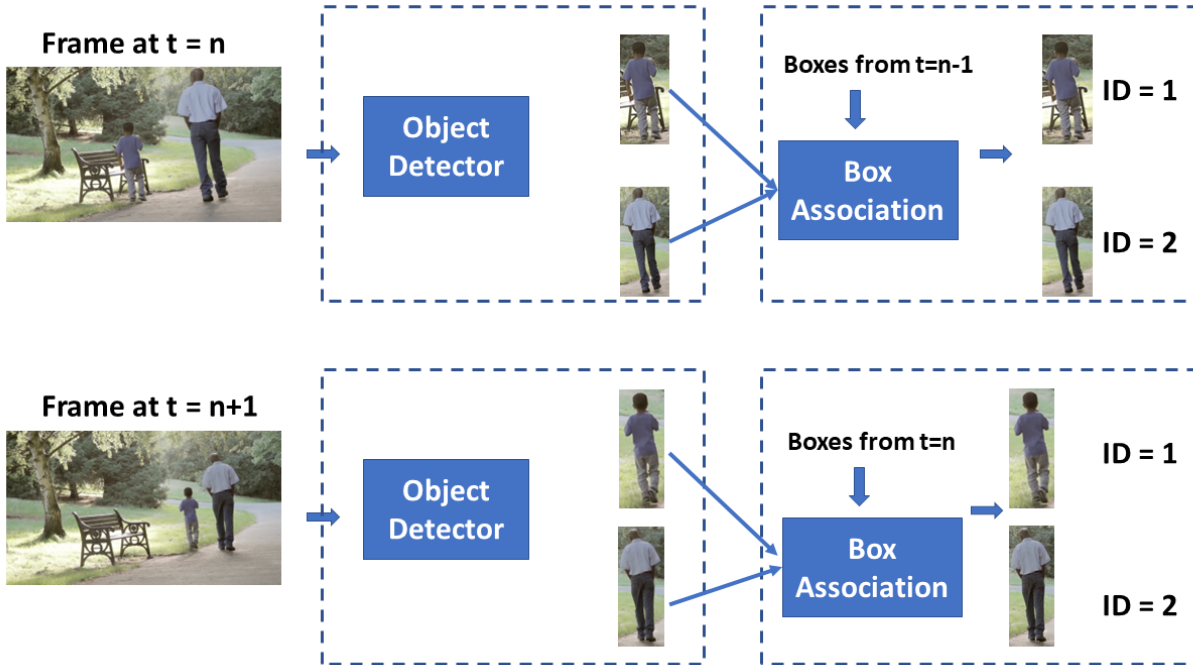


FIGURE 2.4: In MOT, in addition to predicting the class and location of objects of interest, the model must also maintain the same class ID for the same object across different frames. This allows the re-identification of the object as it moves throughout a video sequence. This figure shows the workflow of two stage MOT where the first step is to detect the objects, and the second step conducts box association between bounding boxes of the previous frame and bounding boxes of the current frame.

The tracking-by-detection paradigm is more suitable for real-world applications, where different detectors could be used as a first step in the tracking pipeline, and the training data does not necessarily need to contain tracking ground truth labels. In addition, the rapid breakthrough in deep learning has led to the emergence of faster and more accurate detectors [29, 127, 69]. This has led to more research in MOT utilizing state-of-the-art object detection models [122, 9].



FIGURE 2.5: Variable appearance of pedestrians captured using RGB sensors.

Since object detectors are not perfect, and there will always be cases where the detector predicts false positive boxes or misses true detections (false negatives), state-of-the-art MOT approaches often eliminate predicted boxes with a low confidence by setting a cut-off threshold for detector confidence. However, this inevitably leads to some true detections being ignored, especially in cases where there is occlusion.

The main two metrics we use to evaluate the reliability of MOT models are:



FIGURE 2.6: Inter-class occlusion (left) and intra-class occlusion (right).

- MOTA: stands for *Multiple Object Tracking Accuracy*, and it is a commonly used metric in the field of *Multiple Object Tracking (MOT)* for evaluating the performance of object tracking algorithms. MOTA is used to assess how well an algorithm tracks objects over time in video sequences or other tracking scenarios.

MOTA takes into account various factors, including false positives (**FP**), false negatives (**FN**), identity switches (**IDs**), and the total number of frames. It quantifies the overall accuracy of the tracking system by considering both the tracking errors (e.g., missed detections and incorrect associations) and the tracking performance (e.g., correctly tracked objects). It is possible for MOTA to be negative, indicating a tracking system that performs worse than a simple detector without tracking. A negative MOTA suggests that the tracker is introducing more errors than it is correcting. The formula for calculating MOTA is as follows:

$$\text{MOTA} = 1 - \frac{\text{FN} + \text{FP} + \text{IDs}}{\text{GT}}$$

Where:

- FN (False Negatives): Number of ground truth objects that were not correctly tracked.
- FP (False Positives): Number of objects that were falsely detected or associated with a track.
- IDs (Identity Switches): Number of times the tracking system incorrectly switched the identities of objects.
- GT (Ground Truth): Total number of objects in the ground truth (there must be at least one GT instance).

A higher MOTA value indicates better tracking accuracy, with 1 being a perfect score where there are no tracking errors. Conversely, a lower MOTA value suggests more tracking errors and less accurate tracking.

- IDF1: is a metric used in Multiple Object Tracking (MOT) to evaluate the accuracy of identity assignments or tracking of objects over time. It provides a single measure that combines both precision and recall of identity assignments.

IDF1 takes into account the following two metrics:

- IDP (Identity Precision): IDF1 includes the concept of identity precision, denoted as IDP. Identity precision measures how accurately the tracking algorithm associates identities or tracks with objects in a video sequence. It is calculated as follows:

$$\text{IDP} = \frac{\text{Number of Correct Identity Assignments}}{\text{Total Number of Identity Assignments by the Tracker}}$$

IDP assesses the proportion of identity assignments made by the tracker that are correct. A high IDP indicates that the tracker makes accurate identity assignments.

- IDR (Identity Recall): IDF1 also considers identity recall, denoted as IDR. Identity recall evaluates how well the tracking algorithm captures all the true identities in the ground truth. It is calculated as follows:

$$\text{IDR} = \frac{\text{Number of Correct Identity Assignments}}{\text{Total Number of Ground Truth Identities}}$$

IDR measures the proportion of true identities correctly tracked by the algorithm. A high IDR indicates that the tracker successfully captures most of the ground truth identities.

IDF1 combines IDP and IDR to provide an overall measure of identity assignment accuracy using the F1 score. The F1 score is calculated as the harmonic mean of IDP and IDR:

$$\text{IDF1} = \frac{2 \cdot \text{IDP} \cdot \text{IDR}}{\text{IDP} + \text{IDR}}$$

2.5.1 ByteTrack

Of the systems utilizing the tracking-by-detection paradigm, ByteTrack [130] has achieved state-of-the-art performance on the test data of the benchmark MOT17 dataset [81] with a MOTA of 80.3%. ByteTrack utilizes YOLOX [33] to generate detections. Instead of ignoring detection boxes with low confidence, ByteTrack separates the predicted boxes into a set of high-score detection boxes, and a set of low-score detection boxes.

The algorithm first predicts the locations of the tracklets in the next frame using a Kalman filter, then matches the tracklets with the high-score detection boxes by computing the IOU between the high-score detection boxes and the predicted tracklet location. Tracklets that remain unmatched through this first association are then matched with the low-score detection boxes through a second association step. At the end of this two-step association process, tracklets that remain unmatched are deemed to be lost, new tracklets are created from the unmatched high-score detection boxes, and the remaining low-score detection boxes are ignored.

2.5.2 Observation-Centric SORT (OC-SORT)

Another state-of-the-art MOT model is OC-SORT, where the authors introduce a novel approach aimed at addressing the limitations of the traditional SORT (Simple Online and Realtime Tracking) [7] algorithm. OC-SORT is specifically designed to enhance object tracking performance, particularly in scenarios involving object occlusion and non-linear motion. The core innovation centers around a paradigm shift from an estimation-centric methodology to an observation-centric one. To combat the issue of error accumulation when tracking an object that has been temporarily lost, the authors propose an Observation-centric Re-Update (ORU) strategy. ORU leverages observations from a virtual trajectory to recalibrate the parameters of the Kalman filter.

Furthermore, the authors introduce an Observation-Centric Momentum (OCM) term within the association cost function, prioritizing the use of observations over estimations to improve the accuracy of motion direction estimation. An Observation-Centric Recovery (OCR) technique, which aids in

recovering temporarily lost object tracks, such as those occluded or momentarily stationary is also introduced. Together, these methodological innovations constitute the OC-SORT algorithm, which is developed as an enhancement to ByteTrack [130].

2.6 Detection and Tracking Using Thermal Sensors

Since our work focuses on the problem of pedestrian detection and tracking, we dedicate this section to cover previous work that utilized thermal sensors for the task of people detection and tracking.

In [28], the authors aim to improve the enforcement of high occupancy vehicle lanes through thermal sensors that are installed inside a vehicle. The authors argue that thermal sensors conceal the visual characteristics of human faces and are thereby more suitable to be installed inside a vehicle to enhance privacy. A dataset is collected with images of a varying number of passengers in different vehicles, with annotations for head and torso locations. The authors use thermal images to train object detectors to count the number of passengers, achieving impressive precision-recall. A compact CNN classifier is trained to output the predicted number of passengers, achieving an accuracy of 91%.

The work in [60] utilizes both an RGB and thermal camera for the task of pedestrian detection. The two sensors are fixed to have a consistent horizontal axis, and the geometric transformation between the two sensors is derived. Data is streamed from both sensors simultaneously, and adaptive image subtraction is applied on each sensor independently to detect movement in each sensor. When movement is detected in one of the sensors, the geometric transformation between the sensors is used to predict the location of the pedestrian in both images. A series of morphological operations are

applied to remove noise, but there is no use of a classifier to clearly differentiate a pedestrian from any other moving object.

In [57], the authors also utilize an RGB sensor and a thermal sensor for accurate detection of pedestrians at night. The approach is composed of two steps: hypothesis generation and hypothesis verification. In the first step, the thermal image is segmented using the multi-threshold OTSU method [42]. In the second step, the extracted regions of interest (ROI) from the thresholding operation are classified using random forests. The features used for classification are Histogram of Oriented Gradients [88] and Histogram of Optical Features [93]. This approach a precision and recall of 97.32% and 98.32% respectively on the INO Video Analytics [44] and LITIV datasets [108].

In [58], the authors devise a method for pedestrian detection and tracking on thermal images obtained from Unmanned Aerial Vehicles for surveillance and security applications such as border control. The proposed approach is composed of two steps: ROI detection and ROI tracking. In the first step, a human detector utilizing saliency maps and contrast enhancement techniques is used alongside an SVM classifier to predict the locations of pedestrians in the image. In the second step, the motion vector from the MPEG compression algorithm are used to predict the location of the pedestrian in the consecutive frames.

2.7 Conclusion

In this chapter, we provided a literature review of the use of thermal sensors in the field of computer vision, in addition to specific research utilizing thermal sensors for the ask of pedestrian detection and tracking. We provided an overview of state-of-the-art approaches in object detection and MOT.

Chapter 3

RGB and Thermal sensor performance in MOT

In this chapter, we explain the work we have conducted on comparing the performance of RGB and thermal sensors for the tasks of multiple object detection and tracking. This work was published in a specialized workshop on Perception Beyond the Visible Spectrum, part of the Computer Vision and Pattern Recognition conference in 2022 [27].

Multiple Object Tracking (MOT) is an integral part of machine vision research. Most tracking-by-detection based MOT solutions utilize video streams from RGB cameras for their operation. However, for real-world applications, it is necessary to utilize sensors that operate in different spectrums to accommodate for varying lighting conditions. Since object detection is the first step of the tracking pipeline in tracking-by-detection approaches, we compare the performance of state-of-the-art object detectors when trained on color images to their performance when trained on thermal images. We introduce a new dataset for multiple object tracking with thermal images and corresponding RGB images and show that state-of-the-art trackers perform better on thermal images, especially in poor lighting conditions. Finally, we propose the use of a dynamic cut-off threshold for tracking-by-detection approaches that factors the size of a predicted box to enhance the tracker association.

This chapter is organized as follows. We provide a brief introduction and background of our work in 3.1. We describe our proposed approach and the experiments conducted in 3.2. The results and analysis are given in 3.3. We conclude this chapter in 3.4.

3.1 Introduction

3.1.1 MOT Applications

Multiple Object Tracking (MOT) represents a fundamental and challenging problem within the field of computer vision. At its core, MOT is the tracking of multiple objects across a continuous stream of video frames, with the primary goal of assigning and preserving unique identities for each object. This task demands not only reliable tracking algorithms but also robust methods for handling object occlusions, appearance variations, and complex interactions between objects. As a result, MOT has garnered significant attention from researchers and developers aiming to harness its potential across diverse applications.

One of the key domains where MOT has made substantial contributions is in the realm of autonomous vehicles. In this context, MOT plays a critical role in enhancing the perception capabilities of autonomous vehicles. By accurately tracking and identifying pedestrians, cyclists, and other vehicles, MOT systems enhance safety and facilitate intelligent decision-making. The ability to continuously monitor and predict the trajectories of surrounding objects is pivotal for achieving safe and efficient autonomous navigation, making MOT an indispensable component of autonomous vehicle technology.

Furthermore, MOT has important applications in the field of human-computer interaction (HCI), reshaping the way humans interact with machines. Through MOT, computer systems can track and interpret gestures, body movements, and facial expressions in real-time, enabling more intuitive and responsive interactions. This technology has far-reaching implications, spanning from natural and immersive virtual reality experiences to assistive technologies that empower individuals with disabilities to engage with digital interfaces more effectively.

In the domain of virtual reality (VR), MOT contributes to the creation of highly immersive and interactive virtual worlds. By precisely tracking the movements and positions of physical objects and users, VR systems can seamlessly integrate real-world actions into the virtual environment. This fusion of the physical and digital realms enhances the sense of presence and immersion, making VR experiences more compelling and engaging for users.

3.1.2 Two-stage MOT models

Most state of the art MOT methods [119, 130, 19, 5] utilize the tracking-by-detection paradigm, which is a two-stage process. In the first stage, a standalone detector is run on the video sequence to predict the location and class of objects present in the frames. In the second stage, a tracker processes these detections to conduct association: assigning a unique ID to detections of the same object across consecutive frames. While the performance of an object detector is usually measured by how accurate it can localize and classify objects in a single frame, the performance of a multiple object tracker also factors how well it can correctly re-identify an object across consecutive frames

of a video sequence.

While some approaches have been developed that perform one shot, end-to-end tracking [126, 63], such methods were still unable to surpass the performance of two-stage tracking methods. Zhang et al. [131] performed an empirical study that concluded that the tasks of object detection and object tracking often compete with each other during training, causing one-shot trackers to be less accurate than two-stage trackers.

3.1.3 Perception Systems and Thermal Sensors

Deep learning helped achieve significant breakthroughs in machine vision tasks (advanced perception modules, intelligent monitoring systems, and autonomous vehicles, to name just a few) [87]. Such systems rely on the fusion of information from multiple types of sensors (lidar, radar, RGB cameras, depth sensors, thermal sensors, etc.) to get a better perception of the environment. This information is processed by an artificial intelligence module to perform advanced analysis and make critical decisions. Accurate multiple object detection and tracking is an essential task as it allows the localization of objects of interest and the prediction of the trajectory of moving objects.

The use of thermal sensors in machine vision is becoming more popular [52, 23, 36, 80, 102, 132, 101] as they offer a powerful perception of the thermal identities of objects in a scene. They are also suitable for outdoor applications as thermal sensors operate normally at night and are not significantly affected by poor weather conditions [107].

3.1.4 Chapter Summary

In this chapter, we study the feasibility of using thermal sensors to conduct accurate multiple object detection and tracking. The core contributions of the work elaborated in this chapter are summarized below:

- We introduce a new dataset for multiple object tracking with images and ground truth annotations for RGB and corresponding thermal images.
- We compare the performance of two state-of-the-art object detectors (TOOD [29], VFNET [127]) when trained on thermal images to when they are trained on RGB images of the Teledyne FLIR Thermal Dataset for Advanced Driver-Assistance Systems¹.
- We study the efficacy of applying transfer learning of weights trained on a RGB dataset when training an object detector on thermal images.
- We develop a tracking-by-detection MOT method based on the current state-of-the-art approaches that operates on thermal images, and enhance the data association of the tracker by applying a dynamic cut off score for detections based on the predicted box area.

3.2 Proposed Approach and Experiments

In this section, we elaborate on the experiments we have conducted in the color and thermal domain on object detection, multiple object tracking, and our proposed dynamic confidence thresholding for MOT.

¹<https://www.flir.ca/oem/adas/adas-dataset-form/>

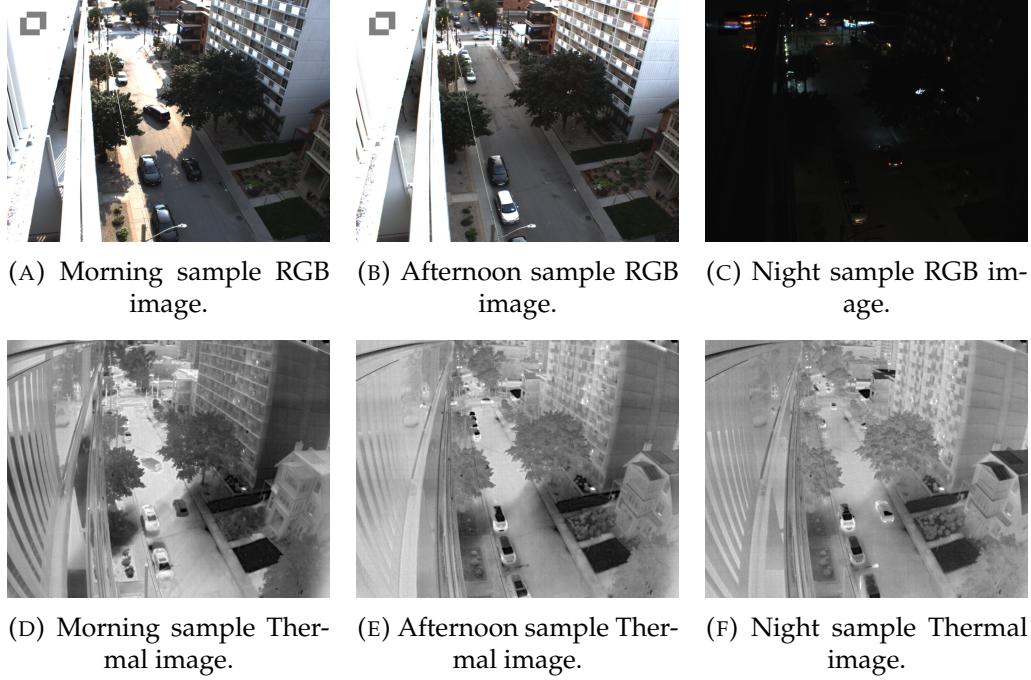


FIGURE 3.1: Samples from the three testing sequences of our dataset taken during different times of the day.

3.2.1 Datasets

City Scene RGB-Thermal MOT Dataset

For the purpose of comparing tracking methods on visible and infrared images, we collected and manually annotated a dataset using a FLIR infrared camera and a visible-light camera, both at a framerate of 10Hz. One in every 2 frames was annotated, resulting in an effective framerate of 5Hz for the dataset. During data collection, the cameras were static and aimed at a city intersection. Cars and pedestrians were annotated up to a distance of 300m and 100m respectively.

The dataset is composed of 15 sequences collected over the course of a day, for a total of 1,997 annotated frames. These sequences are divided into a training set of 12 sequences (1,591 frames) and a testing set of 3 sequences (406 frames). Samples from the three testing sequences are given in Figure

3.1.

The annotated frames contains 267 unique car instances, 25,985 car bounding boxes, 145 unique pedestrian instances, and 7,822 pedestrian bounding boxes.

In this module, the dataset was not aligned. The sequences from RGB and thermal were annotated separately.

FLIR ADAS dataset

In our experiments, we study the performance of state-of-the-art object detectors when trained on color images compared to when trained on thermal images. To that end, we use the FLIR Thermal Dataset for Advanced Driver-Assistance Systems, which is one of the largest and most comprehensive thermal datasets that also provides color images corresponding to the thermal images. It is composed of city scenes captured using a thermal sensor and an RGB camera installed on top of a car. It is manually annotated for 15 classes (person, car, bike, etc.). There is a total of 11,886 training images and 3,749 testing images.

3.2.2 Object Detection Experiments

The object detector is an integral part of tracking-by-detection approaches. The performance of the tracker is significantly influenced by the performance of the object detector. In our experiments on object detection, we address the following two matters:

- We study the efficacy of applying transfer learning of weights trained on the Imagenet dataset [24] (color images) to train an object detector on thermal images.

Parameter	Value	
	Toood	VFNet
Learning Rate	0.01	0.005
Batch size	6	4
Input image size	300×300	300×300
Optimizer	SGD	SGS

TABLE 3.1: Training hyper-parameters

Experiment #	Weight Initializatoin	Trained / Tested on	Object Detector
1	Imagenet	ADAS Thermal Train/Test Sets	TOOD
2	Random	ADAS Thermal Train/Test Sets	TOOD
3	Imagenet	ADAS RGB Train/Test Sets	TOOD
4	Random	ADAS RGB Train/Test Sets	TOOD
5	Imagenet	ADAS Thermal Train/Test Sets	VFNET
6	Random	ADAS Thermal Train/Test Sets	VFNET
7	Imagenet	ADAS RGB Train/Test Sets	VFNET
8	Random	ADAS RGB Train/Test Sets	VFNET

TABLE 3.2: Object detection experiments conducted

- We compare the performance of object detectors when trained on thermal images against when they are trained on color images. As both the FLIR ADAS dataset and our City Scene RGB-Thermal MOT dataset contain images taken at night, this experiment is important to examine the efficacy of thermal images under poor lighting conditions.

Table 3.2 shows a summary of the object detection experiments conducted. We use the MMDetection toolbox [17] for all object detection training experiments we perform. The tuning of training parameters was guided by previous experience, established default values, in addition to experiment results. Those parameters are listed in table 3.1.

We use Resnet50 [40] as the backbone of the detectors in all experiments. We use focal loss [70] for box classification. We apply data augmentation (resizing to random smaller dimensions preserving aspect ratio, and flipping) to enrich the dataset.

3.2.3 MOT Experiments

For our experiments on MOT, we use the state-of-the-art ByteTrack [130] approach. ByteTrack’s simple design that conducts data association based on motion similarity make it ideal for our experiments, as the tracker does not require any domain-specific training. We fine tune the trained object detectors on the tracking dataset we collected as follows:

- Fine tune TOOD and VFNET on the thermal images of our City Scene RGB-Thermal MOT dataset with weights initialized from the trained detectors on the ADAS Thermal dataset.
- Fine tune TOOD and VFNET on the RGB images of our City Scene RGB-Thermal MOT dataset with weights initialized from the trained detectors on the ADAS RGB dataset.

3.2.4 Dynamic Confidence Cut-Off (DCC)

In the original implementation of ByteTrack, the authors set a minimum detection threshold $\mathcal{T}_{min}$ (0.1) and a threshold for high score detection boxes $\mathcal{T}_{high}$ (0.5). Bounding boxes that have a confidence below $\mathcal{T}_{min}$ are ignored, boxes that have a confidence between $\mathcal{T}_{min}$ and $\mathcal{T}_{high}$ are treated as low score detection boxes, and boxes that have a confidence higher than $\mathcal{T}_{high}$ are treated as high score detection boxes as explained in Section 2.5.

A drawback of this implementation is that it treats all objects of all sizes in the same manner strictly based on which range the confidence value falls in. This inevitably leads to false and missed detections as it does not take into account several factors like very small detections that are far from the camera, significant overlap and occlusion, especially as the size of the objects gets smaller as they move away from the camera (as is the case in some sequences

in our City Scene RGB-Thermal MOT dataset, where the size of a tracked car for example becomes smaller as it moves further from the camera).

To address this issue, we propose the use of a dynamic confidence cut-off score for $\mathcal{T}_{high}$ in the implementation of ByteTrack inspired by the work conducted by Stalder et al. [105], where the authors enhance detector prediction confidences by incorporating spatial and geometrical context . Our results show that the dynamic confidence cut-off significantly improves the performance of the trackers, especially for objects with a smaller area. We elaborate more on the findings in Section 3.3.2.

3.3 Results and Analysis

In this section, we elaborate and analyze the results we achieved on object detection and MOT.

3.3.1 Precision Recall of Object Detectors

To study the performance of TOOD and VFNET on the thermal and RGB variants of the ADAS dataset, we plot the precision-recall curves of all the models from Table 3.2. The results are given in Figure 3.2 for TOOD and Figure 3.3 for VFNET.

From the analysis of the precision-recall curve of TOOD, we see that the detectors trained on thermal images perform significantly better than the ones trained on RGB images. This could be attributed to the fact that the ADAS dataset contains data captured at night, where the RGB images would contain little to no features of the objects of interest present in the frames. This explains why the maximum recall achieved by the detectors trained on RGB images is 79%, while the maximum recall achieved by the detectors

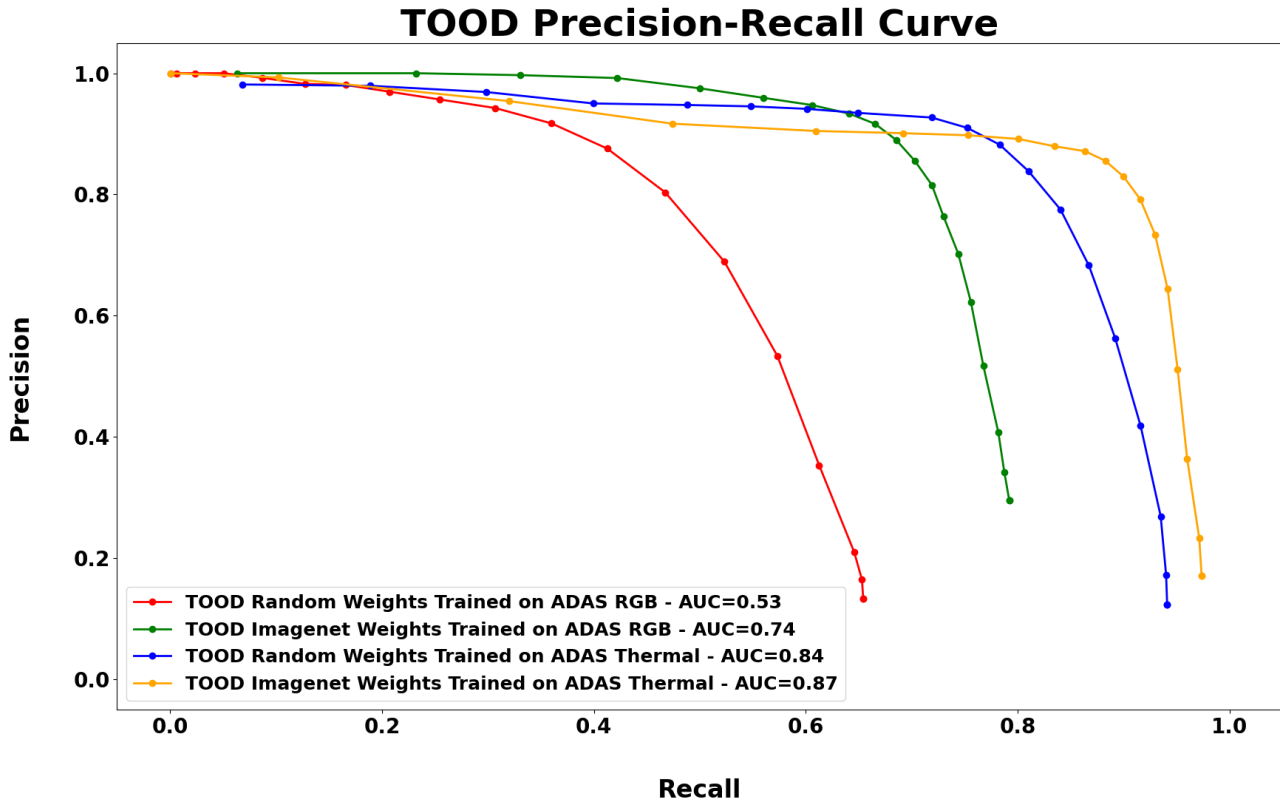


FIGURE 3.2: TOOD Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box. The precision/recall value pairs are calculated using detection threshold values between 0 and 1, with 0.05 increments.

trained on thermal images is 97%: even with a very low detection confidence threshold, the detector still can't localize objects in the dark in the color spectrum. As for precision, it increases with the increase of the detection confidence threshold, as expected. This shows the superiority of detectors trained on thermal images, especially under poor lighting conditions. The ADAS dataset does not split the frames captured during the day from those captured during the night, so we were unable to conduct experiments to show the performance of the detectors exclusively on the frames taken at night. However, in the dataset that we collected, we split the sequences taken during the day from those taken at night, allowing us to compare the performance of the trackers in both settings.

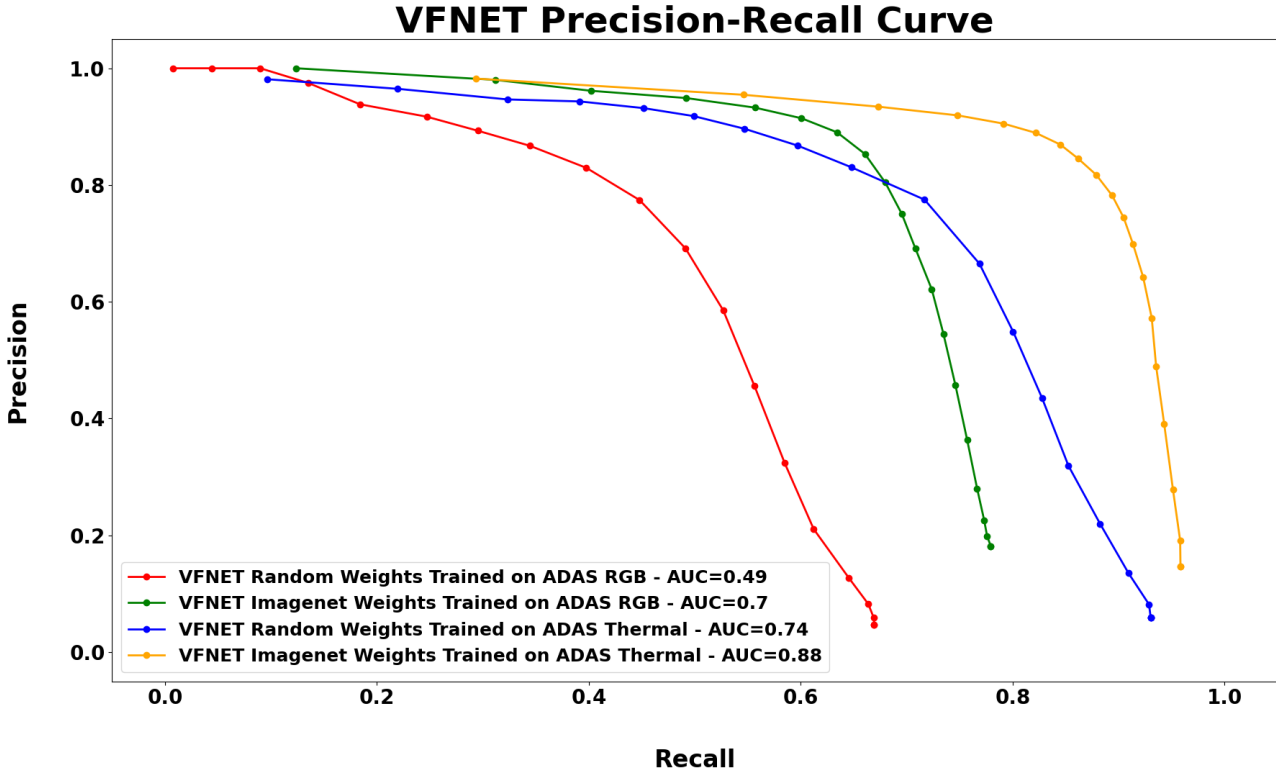


FIGURE 3.3: VFNET Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box.

The results also show that using transfer learning when training the detectors improves the overall performance of the detector. However, the results show that the effect of transfer learning was more significant on the detectors trained on the RGB images of the ADAS dataset. This could be explained by the fact that pre-trained weights are also a result of training RGB images (the Imagenet dataset [24]). While transfer learning did improve the overall performance of the detector trained on thermal images, the improvement was not as significant. This shows that transfer learning is most effective when the initial weights and the dataset on which the detector is being trained on both belong to the same spectrum.

The analysis of the precision-recall curve of VFNET confirms the above

findings. The maximum recall achieved from the detectors trained on RGB images is 78%, while the maximum recall achieved by the detectors trained on thermal images is 96%. Applying transfer learning also improved the performance of detectors both in the color and thermal domains, but we notice that the improvement in the color domain is more significant.

3.3.2 MOT Metrics

To study the effectiveness of the trained detectors in the task of MOT, we calculate the standard MOT metrics [6, 97] of ByteTrack with DCC when using each trained detector on our City Scene RGB-Thermal MOT dataset. We retrain the detectors on our City Scene RGB-Thermal MOT dataset with the weights initialized from the weights of the detectors trained on the ADAS dataset. We study the performance of the tracker across three testing sequences, taken in the morning, afternoon, and at night to study the performance of the tracker under different lighting conditions. We use the standard MOT metrics for comparison and focus mostly on the two key metrics MOTA (Multiple Object Tracking Accuracy) and IDF1 (Identity F1 score). The results are given in Table 3.3.

The results on the sequence taken in the morning show that the trackers operating on RGB images are more effective than trackers operating on thermal images. This could be explained through analysis of the sunlight distribution in that sequence. From Figure 3.1, we notice that the field of view of the thermal camera covers an area that is partially in the shade. Since the dataset was collected in a hot day in the month of July, there is a considerable difference in temperature between the areas in the shade and the areas in direct sunlight. This could be noticed by comparing the appearance of the car present in the shade to the appearance of the car present in the sunlight.

This difference makes it more difficult for the detector to operate on thermal images of this sequence.

In the second sequence taken in the afternoon, we can see from Figure 3.1 that the entire field of view of the thermal camera is almost entirely in the shade, so there is no significant variation in temperature values across different parts of the field of view. This results in the tracker performing better on thermal images, even having a higher MOTA than the tracker running on RGB images.

In the sequence taken at night, we notice that the trackers operating on thermal images perform significantly better than the trackers operating on RGB images (34% higher MOTA when using TOOD, 48% higher MOTA when using VFNET). This is further proof that the detectors perform better on thermal images when there is not a significant variation in temperature across different parts of the field of view. In addition, the trackers operating on RGB images are expected to struggle in detecting objects at night as they are not clearly visible.

To study the effect of using a dynamic cut-off confidence, we compare the performance of the trackers with DCC against the performance of the trackers when using a fixed confidence for high-score detection boxes as in the original implementation of ByteTrack does. The results are given in Table 3.4. It can be seen that applying a DCC noticeably improves the MOTA of the trackers, especially since there are lots of objects with a small area in the dataset (far from the camera).

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.35	0.57	2	1	6	63	403
TOOD RGB	0.86	0.92	6	4	0	51	51
VFNET Thermal	0.33	0.55	2	1	6	56	417
VFNET RGB	0.66	0.84	6	3	1	185	61

(A) Results on the test sequence taken in the morning.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.77	0.85	6	1	3	83	14
TOOD RGB	0.7	0.82	5	2	3	46	14
VFNET Thermal	0.58	0.79	6	0	4	153	0
VFNET RGB	0.57	0.75	5	1	4	175	0

(B) Results on the test sequence taken in the afternoon.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.85	0.92	6	4	1	40	81
TOOD RGB	0.51	0.67	3	4	3	126	268
VFNET Thermal	0.82	0.91	8	2	1	43	106
VFNET RGB	0.34	0.63	5	3	2	337	189

(C) Results on the test sequence taken at night.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.67	0.81	14	6	10	186	608
TOOD RGB	0.69	0.81	14	10	6	223	585
VFNET Thermal	0.59	0.77	16	3	11	252	748
VFNET RGB	0.52	0.74	16	7	7	697	533

(D) Overall results across the three testing sequences.

TABLE 3.3: Overall ByteTrack with dynamic cut-off results on the 3 sequences of our City Scene RGB-Thermal MOT dataset. a) First test sequence captured during daytime. b) Second test sequence captured during daytime. c) Third test sequence taken at night. d) Overall performance across all 3 testing sequences.

Detector	Overall MOTA without DCC	Overall MOTA with DCC	Improvement
TOOD Thermal	0.58	0.67	9%
TOOD RGB	0.62	0.69	7%
VFNET Thermal	0.51	0.59	8%
VFNET RGB	0.46	0.52	6%

TABLE 3.4: Comparison of the performance of the trackers before and after the use of a DCC on our City Scene RGB-Thermal MOT dataset.

	AWP		Xavier	
	Tood + DCC	VFNET + DCC	Tood + DCC	VFNET + DCC
Detection Time (ms)	0.05783	0.06197	0.54908	0.58723
Tracking Time (ms)	0.00085	0.00085	0.00998	0.00998
Total (ms)	0.05868	0.06282	0.55906	0.59721
Total Frames Per Second	17.04	15.92	1.79	1.67

TABLE 3.5: Benchmarking tracker speed on the AWP machine and the Jetson Xavier module.

3.3.3 Speed

We benchmark the speed of the proposed trackers in two environments:

- A powerful machine with an NVIDIA RTX 3090 GPU, and an Intel Core i9 - 10900X 3.70 GHz Processor. (Referred to as AWP).
- A lower power edge device, NVIDIA Jetson Xavier, with a 512-core Volta GPU, and a 8-core ARM v8.2 64-bit Processor. (Referred to as Xavier).

The results are given in Table 3.5. Since the tracking process is independent from the detection process, and the tracking association runs on CPU while the detector inference runs on the GPU, the latency of the tracking process is constant.

3.3.4 Limitations and Failure Cases

As elaborated in Section 3.3.2, the thermal trackers perform poorly under conditions where there are different intensities of sunlight in the thermal sensor's field of view. We also discuss the limitations of the RGB trackers at night. In addition, we notice that there is a noticeable number of incorrect annotations in the FLIR ADAS dataset, on which we heavily rely for the training of our detectors. There are several instances where pedestrians present in an image are not annotated. This would cause a problem during training when a large number of objects predicted as pedestrians by the model do not



FIGURE 3.4: Samples showing missing annotations from the FLIR ADAS Dataset. Green rectangles represent ground truth boxes, red rectangles represent detected boxes, orange arrows point to instances where a pedestrian is detected by a model but no ground truth annotation exists.

have corresponding ground truth annotations, causing the model to train on considering them negatives when they are actually true positives. Similarly, when evaluating the model, several false positives are actually true positives, which would negatively affect the precision and recall values. We randomly sample 100 testing images from the ADAS Thermal dataset, and found that 6 of them have annotation faults (2 of which are shown in Figure 3.4. This should also be taken into consideration when analyzing the precision-recall curves of the models.

Since our dataset covers an urban scene using a fixed sensor setup, and captures far away objects, it is mostly suited for urban surveillance tasks. The system is not expected to perform reliably in setups that are significantly different such as self-driving cars or pedestrian crossings.

3.4 Conclusion

In this chapter, we conducted empirical studies to analyze the feasibility of using thermal sensors for multiple object detection and tracking. We train two state-of-the-art object detectors on the RGB and thermal variations of

the FLIR ADAS dataset, and study the efficacy of transfer learning when applied to training a detector on a dataset from a different spectrum than the initial weights.

We show the superiority of detectors trained on thermal images compared to those trained on RGB images, especially under poor lighting conditions. We also show that transfer learning is more effective when used to train a detector on a dataset from the same spectrum as the initial weights. We introduce the use of a dynamic confidence cut-off, which factors the size of a predicted box, as an enhancement to the motion similarity association of tracking-by-detection based MOT, and show that it improves the accuracy of the tracker.

We have highlighted the limitations of trackers operating on RGB images under poor lighting conditions, and the limitations of trackers operating on thermal images when the thermal sensor field of view is covering areas of significantly different sunlight intensities.

Our experiments and analysis clearly highlight the importance of sensor fusion, especially in critical systems like ADAS. Each type of sensor is optimal under certain conditions and has limitations under other conditions. Being able to combine data sources from different spectrums significantly enhances the perception ability of an autonomous system.

Chapter 4

Enhanced RGB-Thermal Sensor Fusion for Robust Perception

In this chapter, we provide details about the work we have conducted on comparing different methods for RGB-Thermal sensor fusion. We elaborate on our work on introducing a novel sigmoid-activated gating mechanism to enhance RGB-Thermal early fusion methods. This work was published in a specialized workshop on Perception Beyond the Visible Spectrum, part of the Computer Vision and Pattern Recognition conference in 2023.

4.1 Introduction

The rapid advancement of machine learning models in the realm of computer vision has ignited growing interest and concerted efforts in deploying these systems in practical, real-world settings. Ensuring these systems deliver consistent and dependable performance across a wide range of lighting and weather conditions is essential for successful deployment. To address this challenge, machine vision systems have increasingly started incorporating various types of sensors to mitigate the limitations of individual sensors.

In recent years, thermal imaging has taken center stage in terms of development and innovation. The heightened interest in thermal imaging can be attributed to several key factors. Thermal imaging demonstrates remarkable resilience when faced with diverse and challenging weather and lighting conditions. Unlike traditional visible light cameras, which may struggle in low-light settings, be affected by harsh glare, or become ineffective in total darkness, thermal imaging systems remain highly functional and reliable under such adverse conditions. This resilience is particularly critical for applications that demand continuous surveillance and object tracking, such as security, search and rescue operations [104, 50, 22, 51, 94], and autonomous navigation systems.

Another significant driver behind the rapid progress of thermal imaging is the substantial reduction in production costs. Recent technological advancements, coupled with economies of scale, have resulted in more affordable thermal imaging solutions. This increased cost-effectiveness has expanded the accessibility of thermal cameras to a broader range of applications and industries, including industrial automation, firefighting, wildlife monitoring, and automotive safety.

The rapid evolution of machine learning models, computer vision techniques, and sensor technology has increased interest in robust and versatile machine vision systems. To achieve consistent performance in diverse environmental conditions, the integration of different sensor modalities has become a standard approach. Among these modalities, thermal imaging has emerged as a prominent choice due to its ability to excel in challenging scenarios. This development represents a significant milestone in the evolution of machine vision systems tailored for real-world applications.

In this chapter, we study the performance of different RGB-Thermal fusion methods in the task of object detection, and introduce a new RGB-Thermal fusion approach that enhances the performance in object detection. We conduct our experiments on an enhanced version of the City Scene RGB-Thermal MOT Dataset where we register the RGB and corresponding thermal images in order to conduct fusion experiments. The improved version of the dataset, our trained models, and source code were made publicly available.

4.1.1 2D Object Detection

Deep CNNs have achieved impressive results in 2D object detection [12, 38, 71, 59, 25, 111, 112] - the task of classifying and predicting the location of objects in a scene - which has been used for applications such as face detection [41] and autonomous driving [18, 117, 10, 37]. In addition, object detection is often the first step in other computer vision tasks like MOT [120, 129]. Deep learning-based object detection models are most often trained to detect objects on frames from RGB cameras rather than other types of sensors. Thermal sensors prove to be more robust than RGB cameras in some scenarios [27, 106], particularly in low-light settings such as night-time assisted driving. The performance of systems relying on thermal and RGB sensors can be improved through sensor fusion, which should increase a network's performance by improving the quality of the input information, leveraging the important features from both sensors.

4.1.2 Computer Vision Systems

Deep learning has played a pivotal role in driving significant advancements in various machine vision applications, ranging from the development of advanced perception modules to the creation of intelligent monitoring systems

and the realization of autonomous vehicles [87]. These groundbreaking systems heavily rely on the seamless integration of data from a diverse array of sensors, such as lidar, radar, RGB cameras, depth sensors, thermal sensors, and more, to achieve a comprehensive and detailed understanding of the surrounding environment. The fusion of data from these sensors is then processed by artificial intelligence modules, enabling them to carry out sophisticated analyses and make pivotal decisions.

One of the key and indispensable tasks within this context is the precise detection and tracking of multiple objects. This capability is paramount as it enables the system to pinpoint and identify objects of interest, while also facilitating the anticipation and projection of the trajectories of moving objects. Through this robust combination of sensory data integration, artificial intelligence, and object detection and tracking, deep learning has propelled machine vision systems to new heights of accuracy and effectiveness, thereby enhancing their potential for a wide range of applications, including autonomous transportation [86, 16, 84, 115, 123], surveillance [31, 55], and more.

The use of thermal sensors in machine vision is becoming more popular [52, 23] as they offer a powerful perception of the thermal identities of objects in a scene. They are also suitable for outdoor applications as thermal sensors operate normally at night and are not significantly affected by poor weather conditions [107].

4.1.3 Chapter Contributions

Due to the aforementioned importance of object detection as a core computer vision task, we study in this chapter the efficacy of thermal-RGB fusion. The

main contributions of the work discussed in this chapter are summarized below:

- We improve the City Scene RGB-Thermal MOT Dataset by registering the RGB and thermal images so that it can be used for RGB-thermal fusion experiments.
- We compare the performance of object detectors using different backbones, with and without the utilization of existing fusion operations.
- We introduce a novel fusion approach for RGB and thermal images that shows better performance than different existing fusion approaches for the task of 2D object detection.

4.2 Enhancing Our City Scene Dataset

We conduct our experiments on the City Scene RGB-Thermal MOT Dataset [27]. The dataset is composed of thermal and RGB images collected through two different sensors mounted on a fixed support. During data collection, the sensors were static and aimed at a city intersection. Cars and pedestrians were annotated up to a distance of 300m and 100m respectively.

The collected data contains sequences taken during different times of the day (morning, afternoon, and night) making it ideal for experimenting the efficacy of different sensor fusion approaches under different conditions. However, even though the sensors are fixed close to each other, the captured images were not registered and had different resolutions. In order to conduct fusion experiments, the collected RGB and corresponding thermal images needed to be aligned to be in the same coordinate system.

To achieve this, we utilize an ORB feature extractor [99] to extract 5000 features from the thermal image and the corresponding RGB image (after being converted to grayscale). We then take the top 90% matched features (with the lowest hamming distance) and find the homography between the two images using the RANSAC algorithm [30] running for 10000 iterations. Calculating the homography matrix is not guaranteed to be accurate from one pair of thermal-RGB images. However, since the cameras and support are fixed, it is enough for the homography to be accurate for only one pair of thermal-RGB images, and then this calculated homography can be applied to all images in the dataset.

It is important to note that registering the thermal/RGB pair using this method is possible because all of the annotated objects (vehicles and pedestrians) are all rather far from the capturing system. This means that one homography can accurately transform the object coordinates from the RGB domain to the thermal domain. If the annotated objects were at considerably different distances from the capturing system, it would have been difficult for one homography to be reliable for all objects. The accuracy of the homography was visually verified on all the RGB sequences of the captured dataset.

As a result of this transformation, both thermal and RGB images in the enhanced dataset have the same resolution of 500×425 , and share the same annotation file. This contribution allows for thermal-RGB sensor fusion experiments to be conducted and simplifies multi-modal training of object detection and MOT algorithms.

Figures 4.1, 4.2, and 4.3 respectively show the original unregistered dataset samples, the matched keypoints using ORB [99], and the resulting registered dataset samples after applying the transformation homography.



FIGURE 4.1: Original RGB and corresponding thermal images from the City Scene RGB-Thermal MOT Dataset. The original images are unregistered and have different resolutions.

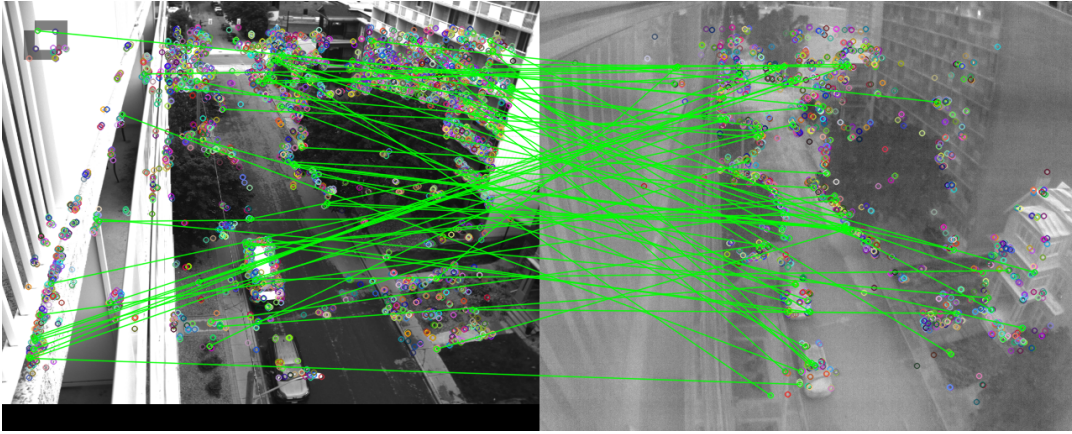


FIGURE 4.2: Matched keypoints between the RGB and thermal images using the ORB feature extractor [99]. The green lines represent the matched keypoints between the RGB and thermal images.

4.3 RGB-T Fusion Experiments

4.3.1 Architectural Overview

The architecture of our RGB-T fusion framework consists of (1) a data preprocessing module, (2) a backbone network, and (3) an object detector. The data preprocessing module loads the registered frames from the City Scene dataset. Each frame contains a pair of images: an RGB image and a thermal image. We compare the performance of fusion experiments to two different

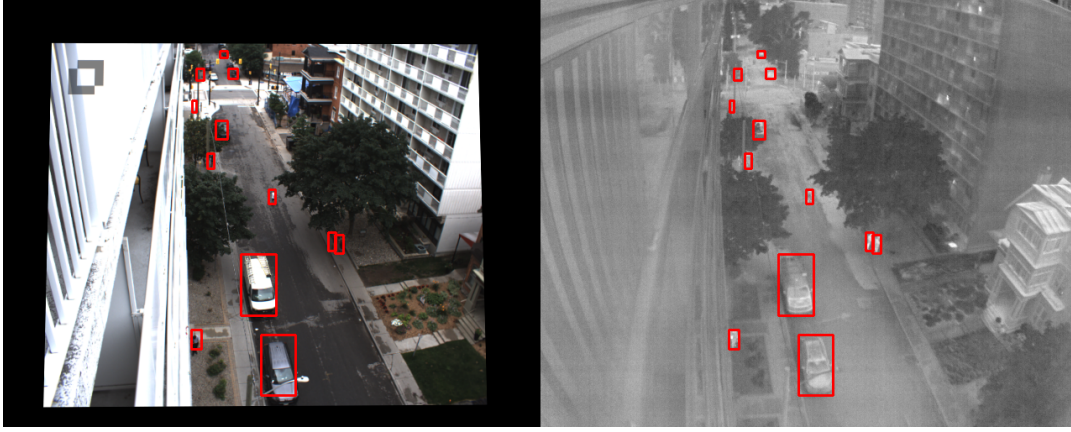


FIGURE 4.3: Registered Thermal and RGB images after applying the transformation homography to the RGB image. The labelled objects in red are ground truth annotations labelled on the thermal images.

baseline models. The first baseline model uses RGB-only input. The second baseline uses thermal-only input. Other experiments perform a fusion operation on both modalities before proceeding to the next step. This type of fusion is denoted as early fusion and is applied before feature extraction. Moreover, the data preprocessing module standardizes the input images before feeding them to the backbone network.

The second component is the backbone network, which performs feature extraction on either its uni-modal or multi-modal input. We experiment using ResNet-50 [39] and ConvNeXt [78]. Both backbone networks are pre-trained on the ImageNet [24] dataset. The final component is the detector. We use TOOD [29] as our single-stage 2D object detector.

4.3.2 Fusion Operators

We aim experimented with our multi-modal detector with two categories of fusion operators: (1) arithmetic-based and (2) learnable fusion. Arithmetic-based operations are advantageous in terms of processing speed since they do not include any learnable weights. Meanwhile, learnable fusion operators

have processing overhead due to the layers of trainable parameters. That being said, learnable fusion operators provide an edge over arithmetic-based fusion when it comes to their learning capacity. This is due to their ability to capture and learn sophisticated interactions between features from different input modalities.

Following recent work [79, 92], we use two arithmetic-based operators: (1) element-wise addition and (2) element-wise multiplication, presented in Equation 4.1 and Equation 4.2 respectively. Where F represents the fused feature map, and I represents an input image. To further enrich our set of experiments, we use two learnable fusion operators: (1) multi-modal factorized bilinear pooling [79] and (2) bilaterally-guided fusion [92], illustrated in Figure 4.4. We implement the aforementioned fusion operators using convolutional layers in order to keep our architecture fully convolutional. We use ReLU as the main activation function, a padding value of 1, a kernel size of 3 in all of our convolution operations, and batch normalization. Euclidean normalization is applied to the fused feature map before it runs through the feature extractor.

$$\mathcal{F}_{add} = \mathcal{I}_{rgb} + \mathcal{I}_{thermal} \quad (4.1)$$

$$\mathcal{F}_{mul} = \mathcal{I}_{rgb} \cdot \mathcal{I}_{thermal} \quad (4.2)$$

4.3.3 Enhanced Sigmoid Gating

In order to enhance the robustness of fusion operators in multi-modal training, we introduced a new fusion gating mechanism that relies on the sigmoid function. The gating mechanism is shown in Equation 4.3. The sigmoid-based gating mechanism, which is simple yet effective, is designed to learn

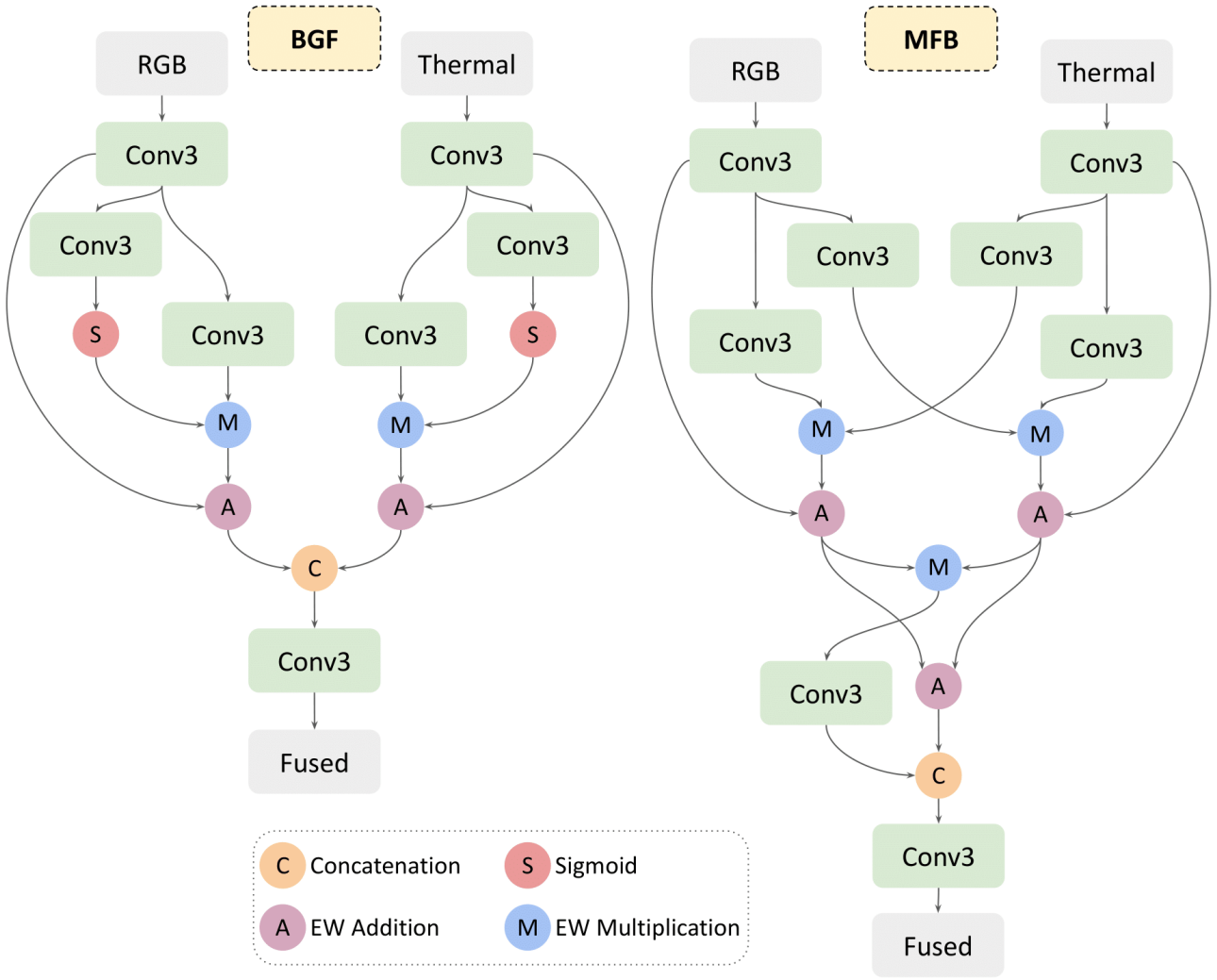


FIGURE 4.4: Left: A diagram illustrating the bilaterally-guided fusion "BGF" operator [92]. Right: A diagram illustrating the multi-modal factorized bilinear pooling "MFB" fusion operator [79]. The input to the fusion operators are the raw modality stream (RGB or thermal image), and the Fused output is then processed by the feature extractor.

the significance of each modality on a per-pixel basis before performing the multi-modal fusion operation.

$$\begin{aligned}\mathcal{G}_{Sigmoid} &= \text{Conv}_{c \rightarrow c}(\text{Sigmoid}(\mathcal{I})) \\ \mathcal{I} &= \mathcal{I} \cdot \mathcal{G}_{Sigmoid}\end{aligned}\tag{4.3}$$

We illustrate our fusion method through Figure 4.5. The gating mechanism is performed by feeding an input modality $\mathcal{I}$, either an RGB or thermal

input, into a sigmoid-activated point-wise convolution operation while preserving the number of channels of the input modality. The resulting sigmoid gate produces a probability value for each pixel, indicating its importance. The gate $\mathcal{G}_\sigma$ is comprised of probability values ranging from 0 to 1, where a value of 1 indicates a significant pixel and a value of 0 indicates insignificance. The higher the pixel probability, the higher its significance. Finally, the probability gate $\mathcal{G}_\sigma$ is multiplied by the original input via point-wise operation. The multiplication will result in keeping the values of significant pixels, but more importantly, diminishing the values of insignificant pixels.

For RGB-thermal fusion, this gating mechanism proves quite effective as thermal and RGB images are complimentary to each other (Thermal sensors perceive well at night, while RGB sensors do not. RGB sensors perceive well when there is little variance between foreground and background, while thermal sensors do not. etc.). In addition, since there is a 1 : 1 mapping between RGB pixels and corresponding thermal pixels, the gating mechanism can reliably learn the significance of each pixel in each modality.

4.4 Experiments

In this section, we present the results of applying both fixed and learnable fusion operators as an early fusion technique for RGB-T data to perform the task of 2D object detection. We first provide our detailed experimental setting in 4.4.1, then we provide a quantitative analysis of early sensor fusion in 4.4.2, followed by reporting results of an enhanced fusion method in 4.4.3. Moreover, we qualitatively analyze in 4.4.4 the performance of our trained detectors on a set of chosen scenarios from the City Scene dataset. Lastly, we provide an analysis of inference time in 4.4.5 by benchmarking our detectors on three different computing platforms.

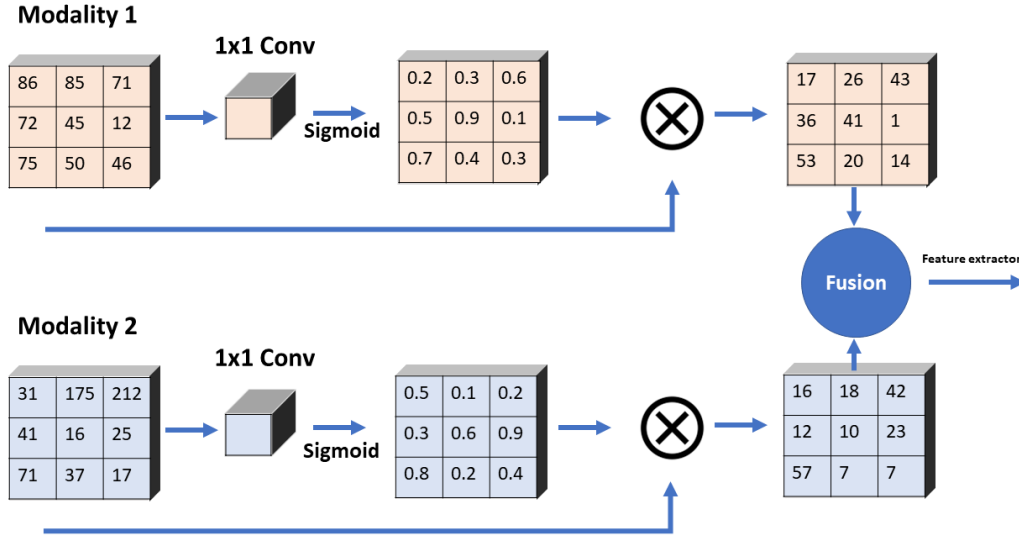


FIGURE 4.5: Proposed sigmoid-gating for early RGB-thermal fusion.

Object Detector	Backbone Network	Fusion Operation	Average Precision Car (AP_{50})	Δ from Thermal-only	Δ from RGB-only
TOOD	ResNet50	RGB	44.1%	-28.6%	-
		Thermal	72.7%	-	+28.6%
		ADD	74.1%	+1.9%	+30.0%
		MUL	75.4%	+2.7%	+31.3%
		MFB	66.2%	-6.5%	+22.1%
		BGF	74.8%	+2.1%	+30.7%
TOOD	ConvNeXt	RGB	30.1%	-38.5%	-
		Thermal	68.6%	-	+38.5%
		ADD	64.0%	-4.6%	+33.9%
		MUL	67.7%	-0.9%	+37.6%
		MFB	71.9%	+3.3%	+41.8%
		BGF	70.8%	+2.2%	+40.7%

TABLE 4.1: Contrasting 2D detection accuracy, i.e. average precision, of four different early fusion operators.

4.4.1 Experimental Setting

To assess the performance of employing early sensor fusion mechanisms, we train and evaluate two different models on the City Scene RGB-T dataset on the task of 2D object detection with a special focus on the 'Car' class. We have chosen to concentrate on the 'Car' class since pedestrian instances

are scarce in the dataset, and they have much smaller bounding box sizes compared to cars. The primary objective of this phase of our research is to assess the impact of thermal-RGB fusion on enhancing detector performance. Therefore, we have decided to emphasize the more frequently occurring class with relatively larger bounding box sizes. This decision aims to mitigate the challenges associated with the detector learning multiple classes and allows us to specifically examine the enhancements achieved through thermal-RGB fusion. The City Scene dataset contains a total of 1968 samples: each sample corresponds to a registered pair of RGB and thermal images along with their corresponding 2D annotations. We use a 75/25 split on the dataset, where 1562 samples are used for training and 406 are used for validation.

We perform our experiments using two variants of the TOOD [29] object detector, each variant contains a different backbone network, namely, ResNet-50 [39] and ConvNeXt [78]. Backbone networks are pre-trained on ImageNet [24]. We freeze the first stage of ResNet-50 while keeping the remaining three stages trainable. We do not freeze any stage in ConvNeXt. Both variants of TOOD are trained using the SGD optimizer, with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. Models are trained for 12 epochs, and we report the best-performing checkpoint out of the 12. We use average precision (AP%) with 50% intersection-over-union as a quantitative metric to assess detection accuracy, along with providing precision-recall curves. Moreover, we provide a qualitative analysis of scenes chosen from the City Scene dataset.

4.4.2 Evaluation of Early Sensor Fusion

In Table 4.1, we report the results of training two variants of the TOOD 2D object detector with four different fusion operators. We focus our results

analysis on the average precision (AP) as the standard metrics for comparing object detection models. A detection is considered to be correct if it has 50% IOU with a ground truth, it is considered a false positive otherwise. For extra reference, we provide the precision-recall curves of the detection experiments of TOOD with ConvNext and Resnet50 backbones in figures 4.6 and 4.7 respectively. These fusion operators are grouped into two subcategories: (1) arithmetic-based fusion operators, and (2) learnable fusion operators.

The arithmetic-based operators are (1) element-wise addition (ADD) and (2) element-wise multiplication (MUL). The learnable operators are (1) multi-model factorized bilinear pooling (MFB) and (2) bilaterally-guided fusion (BGF). We specify two baselines in our comparison: (1) RGB-only and (2) thermal-only. In each of these baseline models, only one modality is fed to the detector. The choice of our baseline models helps us find the delta of performance between uni- and multi-modal settings.

We first assess our baselines: RGB-only and thermal-only detectors, which represent the uni-modal setting. As observed in Table 4.1, thermal-only consistently outperforms RGB-only with a considerable margin when using both ResNet-50 and ConvNeXt with a delta performance of (+28.6%) and (+38.5%), respectively. This finding renders thermal-only a more suitable modality in the 2D object detection task as thermal data is independent of light and is prone to extreme light conditions as opposed to RGB sensors. This result is expected as one-third of the validation set (one of the three validation sequences) of the City Scene dataset is taken in poor lighting conditions.

We aim to enhance the performance of the detector by integrating the data from both modalities. In the first set of fusion experiments, we train and

evaluate a TOOD [29] 2D object detector with a ResNet-50 [39] backbone network. Applying early-fusion results in a large performance boost compared to the RGB-only baseline, with a maximum of (+31.3%) increase for element-wise multiplication, and a minimum of (+22.1%) increase for multi-modal factorized bilinear pooling. Furthermore, both arithmetic-based fusion operators outperform the thermal-only baseline, with a (+1.9%) increase for element-wise addition and (+2.7%) increase for element-wise multiplication.

Fusing RGB with thermal with the multi-modal factorized bilinear pooling results in a (−6.5%) decrease in the average-precision metric, while bilateral-guided fusion results in a (+2.1%) increase. With the exception of multi-modal factorized bilinear pooling, RGB data can contribute to the increase of the overall performance of an object detector. The best-performing detector fused both modalities by applying element-wise multiplication to reach an average precision of 75.4%.

In our second set of fusion experiments, we train and evaluate a TOOD [29] 2D object detector with a ConvNeXt [78] backbone network. Compared to the RGB-only baseline, all fusion operators are showing large performance boosts following the same trend of training with ResNet-50. Both arithmetic-based fusion operators resulted in decreased performance when compared to the thermal-only baseline. However, multi-modal bilinear fusion and bilateral guided fusion both showed a performance increase of (+3.3%) and (+2.2%), respectively, showing the superiority of learning-based fusion operators. Multi-modal bilinear fusion helped the best-performing detector to achieve 71.7% average precision.

In Figures 4.6 and 4.7, we plot the precision-recall curves for all fusion operators, when using ConvNext and Resnet50 respectively as backbones

for the TOOD detector. Consistent trends are observed with the superiority of fusion operators to both baselines in the majority of experiments.

4.4.3 Evaluation of Sigmoid Gating

In Table 4.2, we demonstrate the performance of the enhanced sigmoid gates that are incorporated with the already proposed fusion operators. Sigmoid gates show significant improvements when added to the TOOD ResNet-50 experiments, as both learnable fusion operators, namely multi-modal factorized bilinear pooling and bilateral guided fusion, with staggering (+9%) and (+5%) increases in detection accuracy, respectively.

Meanwhile, the sigmoid gates have shown less impact on arithmetic-based fusion operators, with a (-2%) decrease for element-wise addition, and a (+0.5%) increase for element-wise multiplication. After adding the sigmoid gates to TOOD with the ResNet-50 backbone network, the best-performing model has changed from element-wise multiplication which scored 75.4% AP₅₀ to bilateral-guided fusion, scoring 79.8%, achieving a (+4.4%) increase.

Furthermore, we add the sigmoid gates to the TOOD detector with ConvNeXt backbone network. From Table 4.2, we observe an increase of average precision for both arithmetic-based fusion operators, addition and multiplication, of (+7.6%) and (+1.8%) respectively. When incorporated with learnable fusion operators, sigmoid gates results in a slight decrease in average precision for both multi-modal factorized pooling (-0.2%) and bilateral-guided fusion (-0.7%). To conclude this analysis, we note the importance of carefully selecting the appropriate fusion operators when using sigmoid gates in multi-modal sensor fusion. The optimal fusion operator will depend on the type of feature extractor used, its complexity and learning capacity,

the type of perception task (classification, detection, segmentation, etc.), and the nature of the data being processed.

Backbone Network	Fusion Op.	Sig. Gates $\times$	Sig. Gates $\checkmark$	Δ
ResNet50	ADD	74.1%	72.1%	-2.0%
	MUL	75.4%	75.9%	+0.5%
	MFB	66.2%	75.2%	+9.0%
	BGF	74.8%	79.8%	+5.0%
ConvNeXt	ADD	64.0%	71.6%	+7.6%
	MUL	67.7%	69.5%	+1.8%
	MFB	71.9%	71.7%	-0.2%
	BGF	70.8%	70.1%	-0.7%

TABLE 4.2: Demonstrating the performance boost that results from adding the sigmoid gated mechanism to the fusion operators. The experiments are conducted on the TOOD object detector with two different backbone networks. Experiments are compared based on their average precision (AP_{50}).

4.4.4 Qualitative Analysis

In this subsection, we aim to showcase and qualitatively analyze the performance of the fusion operators. We put RGB and thermal images side-by-side and visualize the bounding boxes on top of them. We visualize groundtruth bounding boxes with a green color and the predictions with a red color. We filter all bounding boxes with a minimum area of 150 pixels to simplify the visualization. Also, we apply a minimum detection confidence score of 40% for all detectors.

In Figure 4.8, we choose a sample day frame from our City Scene [27] dataset. Our first baseline, RGB-only, does not perform well on this frame, predicting only two true positives out of five cars, as well as predicting one false positive. This performance reflects the detection accuracy in Table 4.1 and Figure 4.6. Our second baseline, thermal-only, performs slightly better, by predicting three out of five cars. Fusing both modalities using the bilinear-guided fusion results in detecting all five cars in this frame. Moreover, when

observing the confidence scores of all three detectors, we find that the fusion-based detector produces the most confident bounding boxes, as four out of the five bounding boxes have a confidence over 90%.

We also choose a night frame from the City Scene dataset and visualize the results in Figure 4.9. Our first baseline, RGB-only, does not perform well by predicting three cars out of six in addition to one false positive. Our second baseline, thermal-only, detects all six, but with relatively low confidence scores. The use of bilinear-guided fusion results in the detection of all six cars in this frame, with all detections having confidence scores above 90%. This highlights the effectiveness of multi-modal training to leverage the strengths of multiple sensors which produces more accurate and confident detections.

4.4.5 Inference Time and Benchmarking

In order to study the computational cost of the different fusion operations and the proposed enhanced gating method, we benchmark the inference speed of TOOD with ResNet50 and ConvNext backbones when applying the different fusion operations. The benchmarking experiments were conducted on three different computing platforms:

- Goose: Powerful computing machine equipped with an NVIDIA RTX 3090 GPU (10496 cores).
- Xavier: NVIDIA Jetson Xavier edge computing platform (512-Core Volta GPU).
- TX2: NVIDIA Jetson TX2 edge computing platform (256-Core Pascal GPU).

The benchmarking results are given in table 4.3. The results confirm the statement that learnable fusion operators (MFB and BGF) have a higher computational requirement and thus higher inference latency than arithmetic fusion operators. However, the results also show that the utilization of our proposed enhanced sigmoid gating adds negligible overhead to the fusion operators.

Object Detector	Input Modality	Goose		Xavier		TX2	
		<i>s</i>	<i>fps</i>	<i>s</i>	<i>fps</i>	<i>s</i>	<i>fps</i>
TOOD w/ ResNet50	RGB-only	0.06	17.46	0.77	1.29	1.34	0.74
	Thermal-only	0.06	17.46	0.77	1.29	1.34	0.74
	ADD	0.06	17.57	0.78	1.28	1.34	0.74
	ADD w/ Gating	0.06	17.34	0.78	1.28	1.35	0.73
	MUL	0.06	17.65	0.77	1.29	1.34	0.74
	MUL w/ Gating	0.06	17.34	0.78	1.28	1.35	0.73
	MFB	0.06	16.54	0.82	1.22	1.44	0.69
	MFB w/ Gating	0.09	16.95	0.84	1.20	1.45	0.70
	BGF	0.06	16.20	0.95	1.05	1.78	0.56
	BGF w/ Gating	0.06	16.05	0.96	1.14	1.81	0.55
TOOD w/ ConvNeXt	RGB-only	0.06	16.18	0.89	1.12	1.62	0.62
	Thermal-only	0.06	16.18	0.89	1.12	1.62	0.62
	ADD	0.06	16.29	0.89	1.12	1.62	0.62
	ADD w/ Gating	0.06	15.53	0.90	1.11	1.65	0.60
	MUL	0.06	15.88	0.89	1.12	1.63	0.61
	MUL w/ Gating	0.06	15.53	0.90	1.11	1.63	0.61
	MFB	0.06	15.51	0.94	1.06	1.73	0.58
	MFB w/ Gating	0.07	14.99	0.95	1.05	1.75	0.57
	BGF	0.07	14.94	1.07	0.94	2.06	0.49
	BGF w/ Gating	0.07	14.43	1.08	0.93	2.09	0.48

TABLE 4.3: Inference time and benchmarking results on NVIDIA GeForce RTX 3090, Jetson Xavier, and Jetson TX2. The reported results include inference time in seconds (*s*) and the total number of frames-per-second (*fps*).

4.5 Conclusion

In this chapter, we presented an investigation into RGB-Thermal sensor fusion for the task of 2D object detection. We highlighted the significance of thermal imaging in machine vision, particularly in challenging environmental conditions, and highlighted the increased affordability of thermal sensors. Our experiments were conducted on the enhanced City Scene RGB-Thermal MOT Dataset, where RGB and thermal images were registered to enable fusion experiments.

We evaluated various fusion methods, including arithmetic-based and learnable fusion operators, using two different backbone networks, ResNet-50 and ConvNeXt. The results showed that thermal-only detectors consistently outperformed RGB-only detectors, emphasizing the robustness of thermal data in diverse lighting conditions. Furthermore, we demonstrated that early fusion of RGB and thermal data can lead to substantial performance improvements, with the best-performing fusion method achieving an average precision of 79.8%.

To enhance the fusion process, we introduced an innovative sigmoid-activated gating mechanism that learns the importance of every pixel in each modality. This mechanism proved effective when combined with different fusion operators.

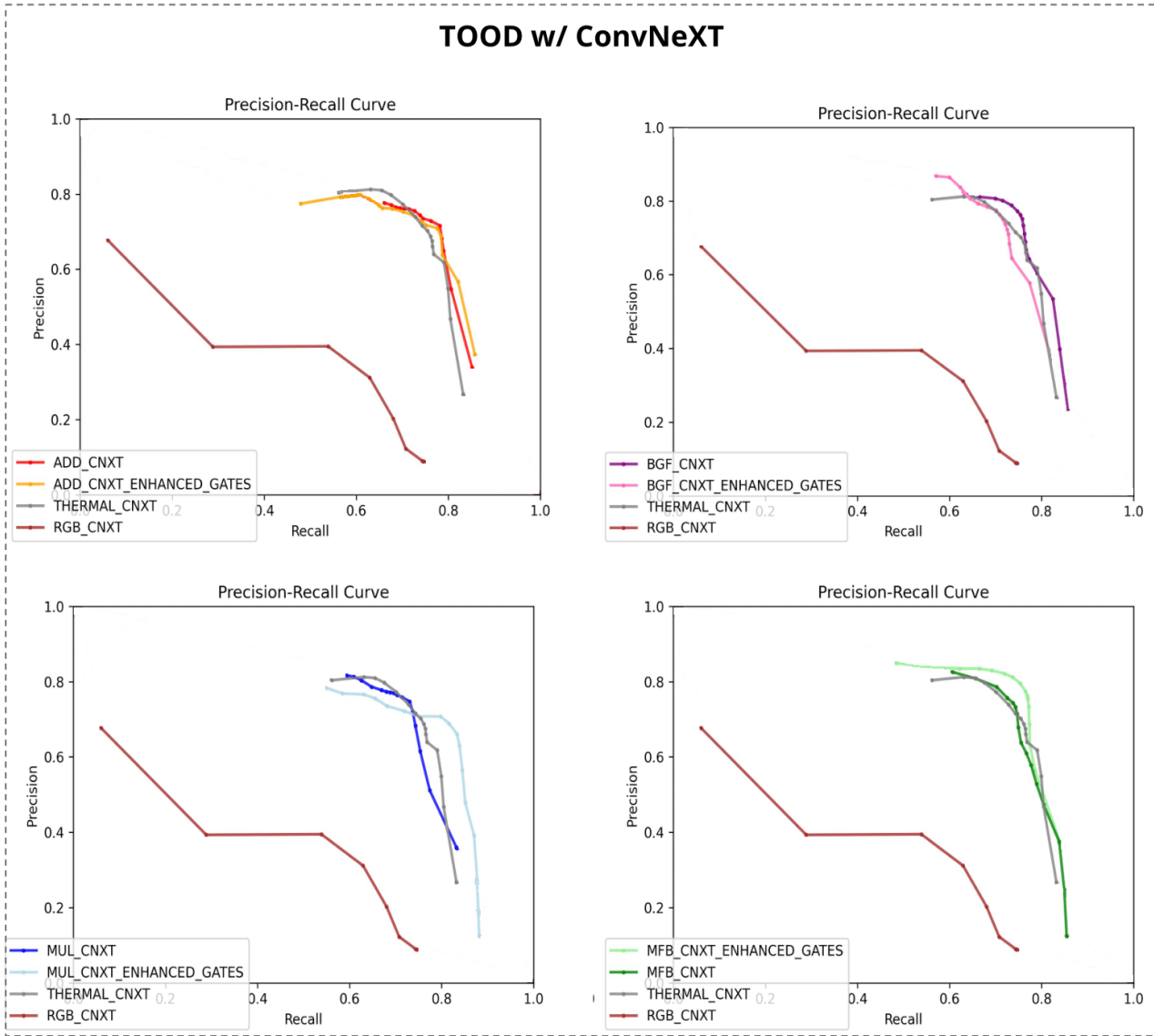


FIGURE 4.6: Precision-Recall curves for TOOD with ConvNeXt as a backbone network. All plots aim to contrast the detection accuracy of fusion operators versus baseline models (RGB-only and thermal-only).

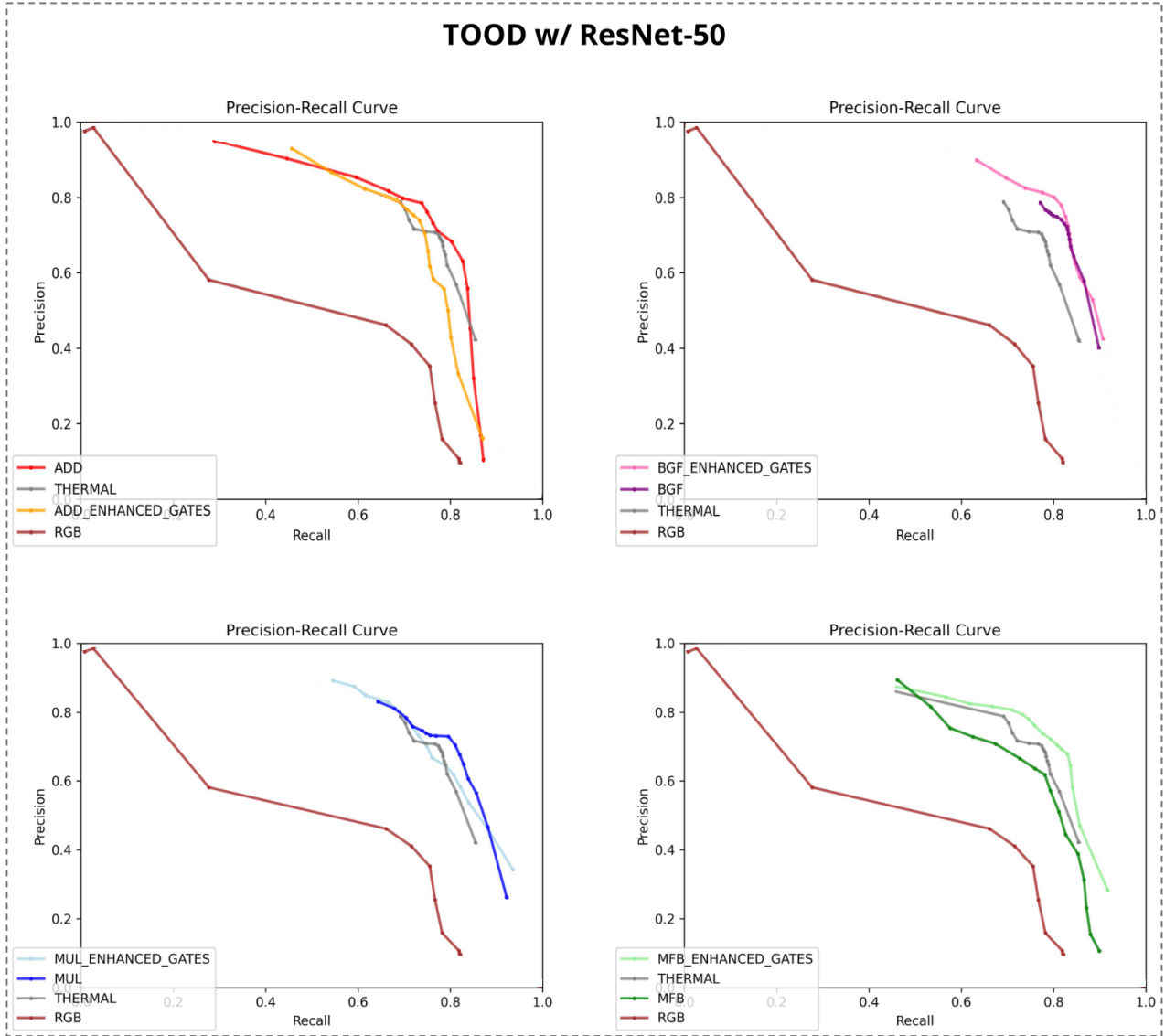


FIGURE 4.7: Precision-Recall curves for TOOD with ResNet-50 as a backbone network. All plots aim to contrast the detection accuracy of fusion operators versus baseline models.

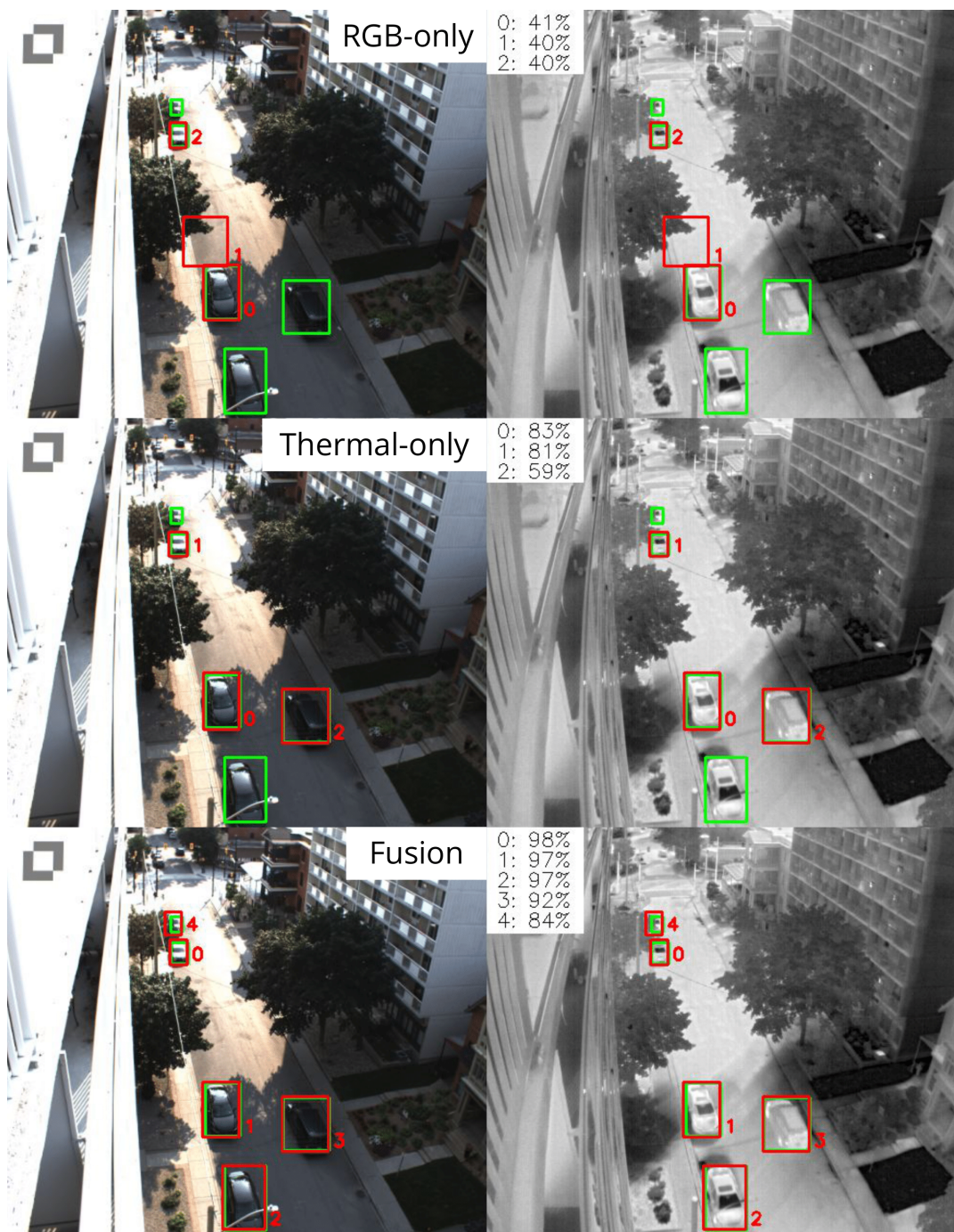


FIGURE 4.8: A day frame from the City Scene [27] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N . Confidence scores (%) of bounding boxes are positioned to the left of each box's index.

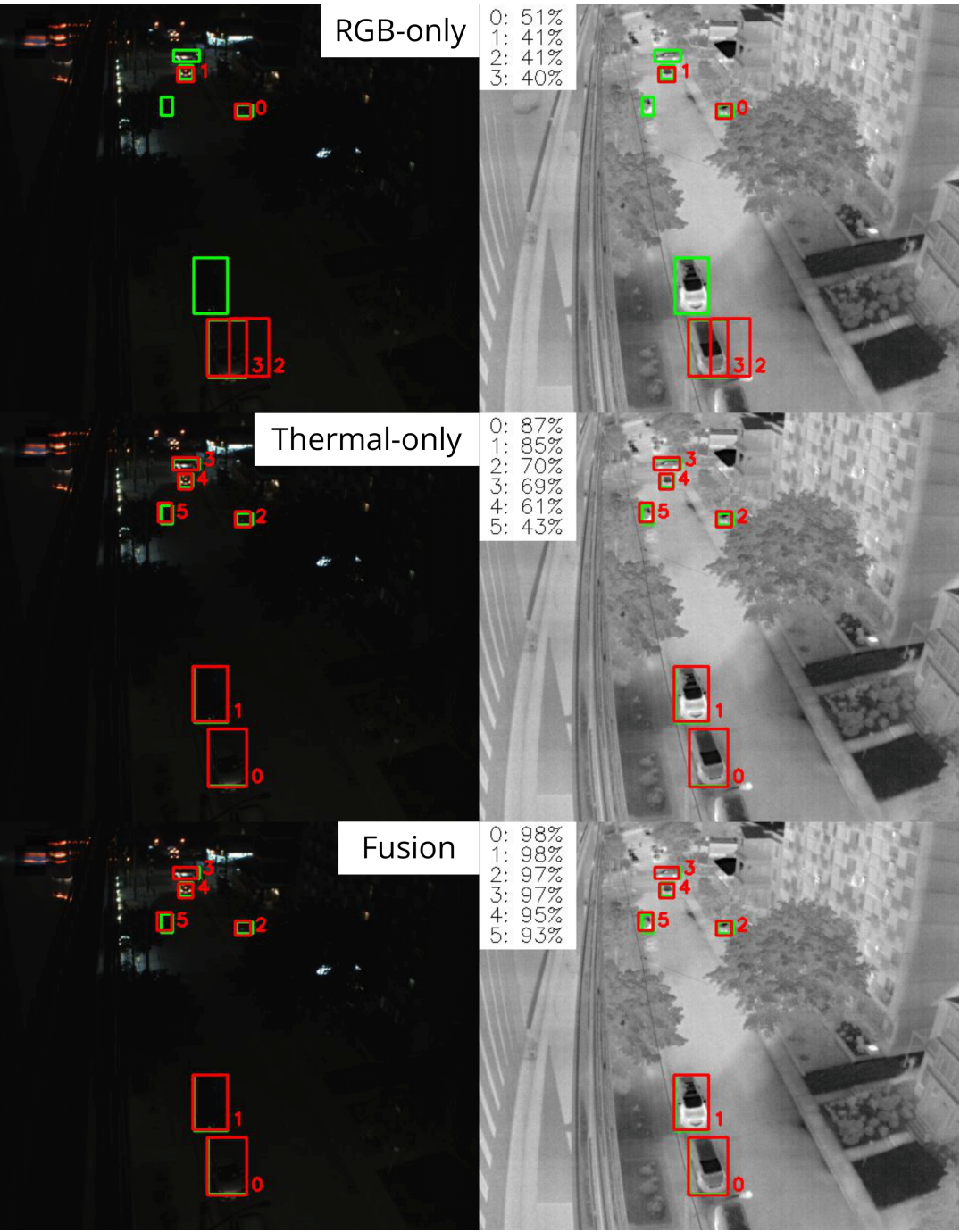


FIGURE 4.9: A night frame from City Scene [27] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N . Confidence scores (%) of bounding boxes are positioned to the left of each box's index.

Chapter 5

Development of Thermal MOT Model

In this chapter, we discuss the work we have conducted on enhancing the performance of MOT models that operate in the thermal spectrum by utilizing the unique characteristics of thermal images. One of the main contributions of this chapter is the introduction of a new dataset of RGB and corresponding thermal images annotated for MOT. This differs from the dataset from previous chapters in that this dataset is larger in size, with more sophisticated scenes. We believe this dataset will prove to be a valuable resource for researchers working in the fields of MOT and thermal vision perception.

5.1 Introduction

Thermal cameras have proven to be robust perception sensors that operate reliably under different weather and lighting conditions for various tasks in computer vision such as detection [67, 53, 2, 13] and tracking [110, 46, 56]. This characteristic of thermal cameras allows vision systems utilizing thermal sensors to utilize the unique thermal characteristics of objects for more robust and reliable performance.

For example, all humans visible in a scene will have roughly the same skin temperature, and consequently similar thermal appearance, regardless of their skin color, their body shapes, and postures. The thermal signatures captured by thermal cameras also allow perception systems to exploit the natural physical characteristics of objects. For example, wood naturally captures and preserves less heat than metal, and this natural difference could be captured using thermal sensors.

This chapter elaborates on the work we have done for utilizing the unique features and characteristics of thermal data for enhanced and more reliable tracking. While existing two-stage tracking models [130, 14] are capable of running on thermal images, they do not utilize the thermal identity of the objects. They operate on detected bounding boxes (predicted through a detector trained on thermal images), while utilizing the same box-association methods utilized in RGB tracking.

5.1.1 Two-Stage Multiple Object Trackers

Two-stage MOT models have proven to be more versatile and accurate than one-stage trackers [130, 14]. Two-stage trackers split the tracking task to two separate components:

- **First component - Box detection:** Use an object detector to localize objects of interest in a frame. The object detector can be deep-learning based, or a traditional CV algorithm as long as it is capable of outputting bounding boxes, confidence scores, and corresponding classes in a specific format. At this step, the object detector operates strictly on a per-frame bases where it is not responsible for predicting and maintaining identities of the detected objects across frames.

- **Second component - Box association:** In this step, a certain algorithm/model is utilized to match boxes predicted in a frame to boxes predicted in the previous frame to maintain object IDs across consecutive frames. This component receives as input the set of detections output by the object detector for the current frame, and matches them with "tracklets" - series of detected bounding boxes of objects detected in previous frames. For example, if we are doing MOT of pedestrians, every pedestrian will ideally have one tracklet, which is the set of bounding boxes predicted by the object detector for this particular pedestrian across all frames where this pedestrian was visible.

The two-stage approach means that any object detector can be used with any box association algorithm to conduct MOT. This makes the two-stage approaches more versatile and suitable for real-world applications. This is a challenge in one-stage trackers that utilize deep-learning models as the model would have to tune its weights to optimize its performance for both detection and association, leading to a degradation in performance [130].

Most two-stage MOT solutions utilize motion association as the main criteria when conducting box association where a Kalman filter [45] is used to predict the locations of objects in the next frame, and then calculate the IOU between the detected bounding boxes and the Kalman-filter predicted boxes to match the boxes across frames. One main advantage of motion association is that the algorithm can be utilized in systems using any type of sensor (visible and non-visible). Regardless of the sensor modality, as long as it is possible to predict bounding boxes, it would be possible to conduct tracking.

However, strictly relying on motion association does have an important drawback that is not utilizing the unique characteristics of sensory data. For example, pixel proximities and distance information could be valuable for

conducting box association in lidar or 3D camera data. In our work, we show that utilizing the thermal identity of the objects in the tracklet association step in two-stage trackers can lead to more accurate box association. Instead of strictly relying on motion association, we devise a box association algorithm that utilizes both motion information and thermal identity of objects.

The utilization of thermal signatures within the cost matrix will also assist in dealing with occlusion - one of the most challenging problems in MOT. When a pedestrian is barely visible in an occluded scene, they are less likely to present visual features that allow detectors operating in the RGB spectrum to localize them. Whereas in the thermal spectrum, even a few pixels of a human head or face would present strong thermal features and temperature variance, allowing for a more reliable detection. Figure 5.1 shows an example of how intra-class occlusion is presented in thermal images.



FIGURE 5.1: Sample occlusion case from our collected dataset

5.2 Implementation details

We rely on the use of two tools developed by the Multimedia Laboratory, the Chinese University of Hong Kong under the OpenMMLab project.

5.2.1 OpenMMLab

OpenMMLab is a comprehensive, open-source project that aims to simplify the development and implementation of computer vision research tools. It is a project initiated by the Multimedia Laboratory, the Chinese University of Hong Kong, aiming to provide a platform where researchers and developers can access high-quality, well-maintained codebases for various tasks such as object detection and tracking. The ecosystem is popular among researchers for its flexibility, extensibility, and user-friendly interface, allowing researchers to integrate and deploy state-of-the-art models seamlessly.

MMDetection [17]

MMDetection is one of the most well-known projects under the OpenMMLab initiative. It is a python-based open-source toolbox designed to accelerate the development of object detection models. It contains an extensive collection of pre-trained models, frameworks, and algorithms such as Faster R-CNN [96], Mask R-CNN [38], and YOLOv4 [111]. MMDetection is characterized by its modular design, facilitating easier adaptation and customization. Researchers and developers can leverage this toolbox to conduct experiments with high reproducibility, simplifying the process of implementing, training, and testing detection models.

MMTracking [20]

MMTracking, another project within the OpenMMLab ecosystem, focuses on multiple object tracking (MOT). It provides a rich set of resources and

tools aimed at enhancing the efficiency and reliability of developing and implementing tracking algorithms. MMTracking supports various tracking paradigms such as single object tracking (SOT) and multi-object tracking (MOT), offering a versatile platform for researchers exploring different tracking domains.

5.2.2 Data Collection

We utilize the same FLIR ADK thermal sensor and a JAI GO-5100C RGB sensor for the collection of the new dataset. The specifications of both sensors were given in tables 2.1 and 2.2 respectively.

A dataset was collected on 5 different intersections around the University of Ottawa campus in public spaces (Figure 5.2). The dataset was then annotated for multiple object detection and tracking. We refer to this dataset as the RGB-Thermal MOT dataset. As the dataset contains instances of pedestrians at varying distances from the capturing system, it was not possible to register them. As such, each modality was annotated separately.



FIGURE 5.2: RGB and corresponding thermal images of the RGB-Thermal MOT dataset.

	Thermal	RGB
Total number of annotated sequences	30 (1 minute each)	30 (1 minute each)
Capture frame rate	5	5
Total frames per sequence	300	300
Total number of annotations	58,590	50,400
Average annotations per image	6.51	5.6
Train/Test split	Train: 24 Test: 6	Train: 24 Test: 6
Total unique tracks Train/Test	Train: 313 Test: 126	Train: 284 Test: 116

TABLE 5.1: RGB-Thermal MOT dataset statistics.

To the best of our knowledge, this is the world’s first large-scale dataset of RGB and corresponding thermal images annotated for MOT. We believe that this dataset will prove to be a valuable resource for research and development for both thermal and RGB MOT research. Some relevant statistics related to the dataset are given in table 5.1.

5.2.3 Training Object Detection Models for two-stage MOT

As we did in chapter 3, we fine tune the TOOD object detector [29] on images from the newly collected RGB-Thermal MOT dataset. We use the OpenMMLab mmdetection framework [17] for training the object detector. The outcome of this training is two models:

- TOOD model initialized with Imagenet weights and fine tuned on RGB images of the RGB-Thermal MOT dataset.
- TOOD model initialized with Imagenet weights and fine tuned on Thermal images of the RGB-Thermal MOT dataset.

5.2.4 Benchmarking ByteTrack and OCSORT on the RGB-Thermal MOT dataset

We use the OpenMMLab mmtracking framework [20] to benchmark SOTA MOT models [130, 14] on both the RGB and Thermal sequences of the RGB-Thermal MOT dataset. The corresponding Thermal/RGB detectors trained in the previous step (5.2.3) were used as backbones for the trackers.

We optimize the hyperparameters of both ByteTrack and OCSORT to maximize their MOT metrics (specifically MOTA and IDF1) when using their standard implementations. This step is crucial to ensure a fair comparison of our innovative box association approach against the highest achievable performance of each of those two trackers.

Furthermore, we conduct a performance evaluation of ByteTrack and OCSORT on the RGB sequences within the RGB-Thermal MOT dataset. This assessment serves as a valuable reference for future users of this recently acquired dataset.

5.2.5 Development of new MOT box association method utilizing thermal information

The standard MOT models do not utilize information specific to thermal images in their operation as they mostly rely on motion similarity. We devise a novel thermal box-association method that utilizes both motion similarity and thermal identity of detected objects. Our approach is detailed below.

Our novel motion-thermal similarity box association is introduced as a new method for enhancing thermal object tracking. It performs pairwise histogram similarity analysis between tracking bounding boxes (track-bboxes)

and detection bounding boxes (det-bboxes) within a given thermal image (thermal-img). The goal is to assess the similarity between thermal object representations in consecutive frames. The function's output, a thermal similarity matrix, is then combined with the standard motion-based similarity scores, creating a comprehensive similarity matrix that utilizes both thermal and motion similarity.

Algorithm Details

The detailed steps involved in our algorithm are detailed below:

- **Initialization:** To start the process, we convert PyTorch tensors, track-bboxes, and det-bboxes, into NumPy arrays. These arrays contain the spatial coordinates of tracking and detection bounding boxes, respectively.
- **Similarity Matrix:** We create an empty similarity matrix called thermal-dists. This matrix will store the pairwise similarity values between tracking and detection bounding boxes, capturing the thermal aspect of the objects.
- **Pairwise Histogram Comparison:** The core functionality of this function involves comparing the thermal signatures of objects. For each pair of track and detection bounding boxes, the following steps are executed:
 - **Region of Interest Extraction:** This function isolates the regions of interest (ROIs) within the thermal image (thermal-img) based on the coordinates of the bounding boxes. These ROIs encapsulate the unique thermal characteristics of the tracked and detected objects.

- **Histogram Calculation:** Histograms of these ROIs are generated using the OpenCV function `cv2.calcHist`. These histograms serve as concise representations of the thermal intensity distributions within the regions.
 - **Histogram Normalization:** To ensure that comparisons are robust and invariant to variations in overall intensity levels, the histograms are normalized to have values between 0 and 1.
 - **Histogram Similarity:** The similarity between the normalized histograms is computed using the Bhattacharyya coefficient [8]. This coefficient quantifies the degree of overlap between the thermal intensity distributions of the objects. A higher value indicates greater similarity.
 - **Normalization of Similarity:** The similarity score is rescaled to fall within the range of [0, 1]. This ensures that it is easily interpretable, with 0 indicating no similarity and 1 indicating perfect similarity.
 - **Updating Similarity Matrix:** The calculated similarity value is placed in the thermal-dists matrix, representing the degree of similarity between the corresponding tracking and detection bounding boxes in the thermal domain.
- **Integration with Motion Similarity:** A weighted average of the motion similarity matrix produced using the standard MOT algorithms and the calculated thermal similarity matrix to generate the final comprehensive distance matrix.

Algorithm Summary

A summarized explanation of the algorithm is given below:

Let $T = \{t_1, t_2, \dots, t_m\}$ represent the set of track-bboxes predicted by the tracker Kalman filter, and $D = \{d_1, d_2, \dots, d_n\}$ represent the set of det-bboxes predicted by the object detector within a given thermal image, denoted as $I_{thermal}$.

Initialization

$$T, D \leftarrow \text{numpy}(T, D) \quad // \text{ Convert to NumPy arrays} \quad (5.1)$$

Similarity Matrix

Define $S_{thermal}$ as an $m \times n$ matrix, where each element s_{ij} represents the thermal similarity between t_i and d_j .

$$S_{thermal} \leftarrow \text{Zeroes}(m, n) \quad (5.2)$$

Pairwise Histogram Comparison

For each t_i in T and d_j in D , the regions of interest, denoted as ROI_{t_i} and ROI_{d_j} , are extracted from $I_{thermal}$ using the x, y, width, and height of every t_i and d_j . Histograms, H_{t_i} and H_{d_j} , are then calculated and normalized to $\hat{H}_{t_i}$ and $\hat{H}_{d_j}$.

Integration with Motion Similarity

Let S_{motion} be the motion-based similarity matrix between track-bboxes and det-bboxes obtained using standard MOT algorithms, then the final comprehensive distance matrix, S_{comp} , is calculated as:

$$S_{comp} = \alpha \cdot S_{motion} + (1 - \alpha) \cdot S_{thermal} \quad (5.3)$$

where α is a weighting factor balancing motion and thermal similarities. The utilization of a weighted sum allows us to study the effect of the incremental integration of thermal signatures in the similarity matrix, as shown

in Figures 5.3 and 5.4.

The pseudo-code of the function is given in Algorithm 1

Algorithm 1: Novel thermal box-association method leveraging both motion and thermal similarities.

Data: $T, D, I_{thermal}$

Result: S_{comp}

```

1 Convert  $T, D$  to NumPy arrays
2 Initialize  $S_{thermal}$  as Zeroes(m, n)
3 for each  $t_i$  in  $T$  do
4     for each  $d_j$  in  $D$  do
5         Extract  $ROI_{t_i}, ROI_{d_j}$  from  $I_{thermal}$  using bounding box
           coordinates
6         Calculate and normalize histograms  $\hat{H}_{t_i}, \hat{H}_{d_j}$ 
7         Compute  $s_{ij}$  using cv2.HISTCMP_BHATTACHARYYA
           method and update  $S_{thermal}[i, j]$  with  $s_{ij}$ 
8     end
9 end
10 Compute  $S_{comp} = \alpha \cdot S_{motion} + (1 - \alpha) \cdot S_{thermal}$ 

```

5.3 Results and Analysis

5.3.1 Comparing MOT models running on RGB against Thermal images

The benchmarking results of running the standard implementations of ByteTrack and OCSORT are given in table 5.2.

	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq54_rgb	67.8%	69.3%	66.4%	75.4%	78.7%	53.1%	0.133
seq17_rgb	64.6%	71.8%	58.8%	78.1%	95.3%	68.9%	0.150
seq2_rgb	56.8%	67.1%	49.3%	68.6%	93.4%	62.1%	0.177
seq66_rgb	71.1%	72.5%	69.8%	92.6%	96.2%	88.0%	0.129
seq47_rgb	58.7%	72.6%	49.3%	64.6%	95.1%	59.6%	0.163
seq22_rgb	87.2%	93.8%	81.4%	82.2%	94.8%	77.0%	0.115
OVERALL	63.6%	72.5%	56.7%	72.4%	92.7%	64.7%	0.152
	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq54_rgb	58.3%	55.7%	61.0%	78.0%	71.2%	42.9%	0.140
seq17_rgb	55.2%	55.1%	55.3%	84.5%	84.1%	58.3%	0.161
seq2_rgb	46.6%	49.7%	43.9%	70.3%	79.6%	48.9%	0.179
seq66_rgb	67.0%	59.6%	76.5%	93.5%	72.9%	57.4%	0.131
seq47_rgb	53.6%	60.1%	48.3%	69.6%	86.7%	53.9%	0.176
seq22_rgb	82.5%	82.8%	82.2%	86.8%	87.4%	70.4%	0.125
OVERALL	56.8%	58.6%	55.1%	76.3%	81.0%	53.7%	0.160

TABLE 5.2: Results of benchmarking Bytetrack on the RGB validation sequence using standard implementations of Bytetrack (top), and OCSORT (bottom).

5.3.2 Weighted average alpha value selection

In order to accurately find the best alpha-value (Equation 1), that corresponds to the weight of motion distance matrix and the corresponding weight of the thermal distance matrix in the comprehensive distance matrix generated through our approach, we calculate the MOTA and IDF1 generated through our approach on the validation sequences of the RGB-Thermal MOT dataset. The results are given in figures 5.3 and 5.4.

The selected alpha-value for each model should ideally take into account the best trade-off between MOTA and IDF1. Analyzing figures 5.3 and 5.4, we can deduce the following:

- The overall performance of ByteTrack is proven to be better than OCSORT, with the maximum overall MOTA value achieved using ByteTrack reaching 66.4%, while the maximum overall MOTA achieved using OCSORT is 56.4%.
- A similar trend is observable through analyzing the IDF1 values of both models, as the maximum IDF1 value achieved through ByteTrack is 64.2% while the maximum overall IDF1 value achieved through OCSORT is 58.6%.
- For ByteTrack, the alpha value with the best trade-off between MOTA and IDF1 is 0.3, meaning the weight of motion association contribution to the comprehensive distance matrix is 30%, while that of the thermal distance matrix is 70%. This shows that ByteTrack benefits significantly from the thermal similarity matrix.
- For OCSORT, the alpha value with the best trade-off between MOTA and IDF1 is 0.8, meaning the weight of motion association contribution to the comprehensive distance matrix is 80%, while that of the thermal distance matrix is 20%. This shows that OCSORT also benefits from the thermal similarity matrix.
- For an alpha value of 0, meaning that only the thermal distance matrix is used for box association, ByteTrack achieves impressive results with 63.4% overall MOTA and 55.1% overall IDF1, surpassing the best performance of OCSORT. This shows the clear benefit of using thermal similarity for conducting box association, as it performs reasonably well even without utilizing motion association at all.

- The choice of which MOT model to use would depend on the nature of the tracking task and the objects of interest.

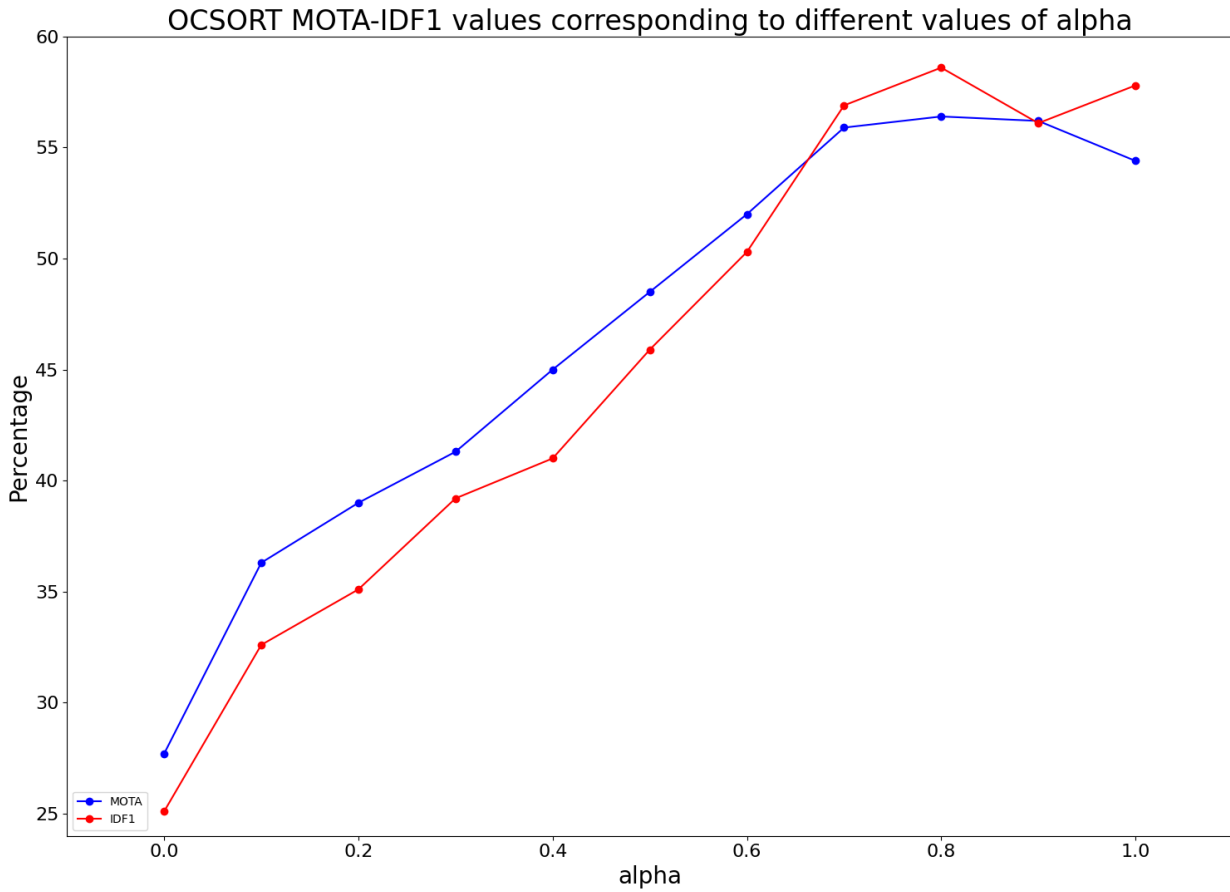


FIGURE 5.3: MOTA and IDF1 values of our proposed box association method applied on OCSORT on the validation sequences of the RGB-Thermal MOT dataset using different values of alpha.

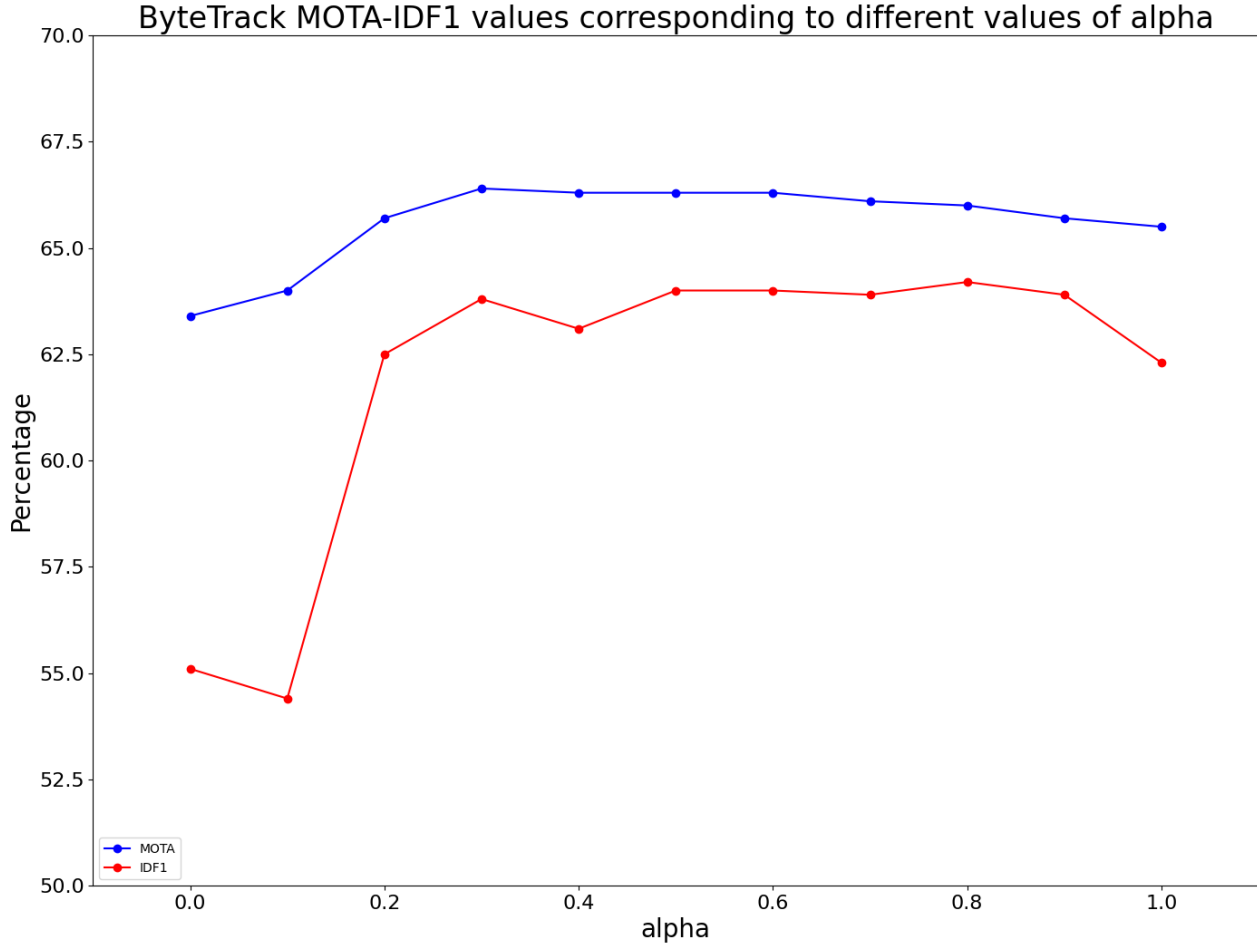


FIGURE 5.4: MOTA and IDF1 values of our proposed box association method applied on ByteTrack method on the validation sequences of the RGB-Thermal MOT dataset using different values of alpha.

5.3.3 MOT Metrics Comparison

For evaluating the feasibility of our suggested box association method, we consider two state-of-the-art MOT models Bytetrack [130] and OCSort [14]. As explained in 2.5, both Bytetrack and OCSort utilize motion association for the box-association step. The detailed results are given in tables 5.3 and 5.4, and summarized metrics are given in figures 5.5 and 5.6.

Analyzing the tables, the following conclusions can be made:

- ByteTrack benefits from utilizing thermal information when conducting box association. With the overall MOTA and IDF1 for the original implementation of ByteTrack being 65.5% and 62.6% respectively, and those values increasing to 66.4% and 63.8% respectively using our proposed approach.
- A similar trend can be seen with OCSORT, with the overall MOTA and IDF1 values increasing from 54.4% and 57.8% respectively in the original implementation of OCSORT to 56.4% and 58.6% respectively using our proposed box association method.

The results validate the effectiveness of our proposed model on enhancing tracking performance, where utilizing thermal similarity proves beneficial for box association when combined with motion similarity.

It is important to note the following when assessing the improvement in MOT metrics that our method introduces:

- We did not use the pre-set parameters for OCSORT or ByteTrack. Instead, we fine tuned the parameters of each MOT model to get the best MOT metrics for a fair comparison.
- It is common practice in MOT model implementations to fine-tune the model hyper-parameters for every validation sequence to achieve the highest results. In our work, we use one-set of hyper-parameters for all validation sequences in order to provide a clear insight on the generalizability of our approach.

The proposed approach combines thermal and motion-based similarity scores through a weighted average. This fusion method effectively integrates two sources of information, allowing the system to make decisions based on

both thermal and motion aspects.

By considering both thermal and motion aspects, the tracking system becomes more robust and adaptable. When thermal data indicates a strong match between objects with similar thermal signatures, the system can prioritize thermal information. Conversely, when motion cues are reliable, they can take precedence. This adaptability makes the tracking system more resistant to false positives and negatives. In addition, thermal data can assist in handling occlusions, a common challenge in MOT. When one object obscures another, thermal signatures may still be distinguishable, allowing the system to maintain the identity of both objects.

	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq66_thermal	67.9%	75.5%	61.7%	78.5%	95.9%	74.4%	0.154
seq2_thermal	51.2%	69.3%	40.6%	56.7%	96.7%	52.5%	0.184
seq47_thermal	59.1%	62.2%	56.4%	81.5%	89.9%	70.1%	0.170
seq17_thermal	71.5%	79.1%	65.2%	75.8%	92.0%	65.7%	0.173
seq54_thermal	63.8%	65.7%	62.0%	77.8%	82.4%	59.2%	0.176
seq22_thermal	75.4%	84.9%	67.9%	74.4%	93.0%	67.3%	0.163
OVERALL	62.6%	69.4%	57.0%	74.9%	91.1%	65.5%	0.170
	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq66_thermal	68.0%	74.3%	62.6%	80.2%	95.1%	75.3%	0.157
seq2_thermal	49.3%	62.4%	40.8%	62.0%	94.9%	56.3%	0.194
seq47_thermal	60.7%	62.1%	59.4%	83.7%	87.6%	69.6%	0.173
seq17_thermal	71.3%	77.0%	66.4%	77.5%	89.8%	64.9%	0.177
seq54_thermal	71.4%	71.5%	71.3%	81.2%	81.4%	60.9%	0.183
seq22_thermal	75.9%	83.4%	69.7%	77.1%	92.3%	69.3%	0.166
OVERALL	63.8%	68.6%	59.6%	77.7%	89.4%	66.4%	0.175

TABLE 5.3: Results of benchmarking Bytetrack on the thermal validation sequence using Bytetrack standard motion association (top), and using our novel thermal-motion box association method (bottom).

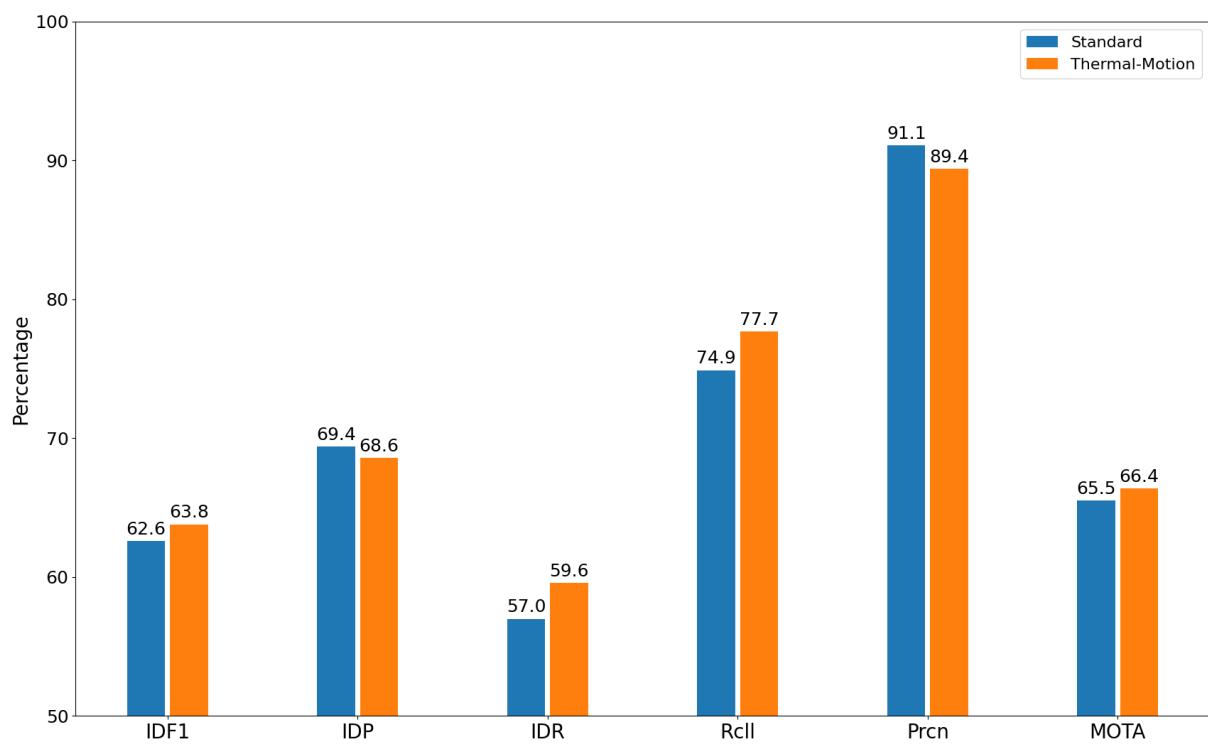


FIGURE 5.5: Overall comparison of the important MOT metrics running standard motion box association against our proposed box association algorithm using ByteTrack.

	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq66_thermal	74.3%	77.4%	71.5%	82.8%	89.6%	70.8%	0.162
seq2_thermal	51.7%	60.4%	45.2%	66.3%	88.8%	51.3%	0.197
seq47_thermal	53.8%	50.0%	58.2%	87.1%	74.8%	53.2%	0.178
eq17 thermal	60.0%	59.3%	60.7%	82.2%	80.3%	55.3%	0.185
seq54_thermal	52.6%	47.4%	59.1%	84.3%	67.7%	40.1%	0.187
seq22_thermal	64.8%	64.7%	65.0%	81.2%	80.9%	57.5%	0.174
OVERALL	57.8%	56.8%	58.7%	81.3%	78.7%	54.4%	0.180
	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
seq66_thermal	76.0%	79.1%	73.1%	82.8%	89.6%	71.7%	0.162
seq2_thermal	42.4%	49.6%	37.1%	66.4%	88.9%	54.1%	0.198
seq47_thermal	56.0%	52.1%	60.6%	87.1%	74.8%	54.5%	0.178
g17 thermal	63.2%	62.5%	64.0%	82.4%	80.4%	58.1%	0.186
seq54_thermal	55.2%	49.8%	61.9%	84.3%	67.7%	42.4%	0.187
seq22_thermal	71.1%	71.0%	71.2%	81.2%	80.9%	60.3%	0.175
OVERALL	58.6%	57.7%	59.6%	81.4%	78.7%	56.4%	0.180

TABLE 5.4: Results of benchmarking OCSort on the thermal validation sequence using OCSort standard motion association (top), and using our novel thermal-motion box association method (bottom).

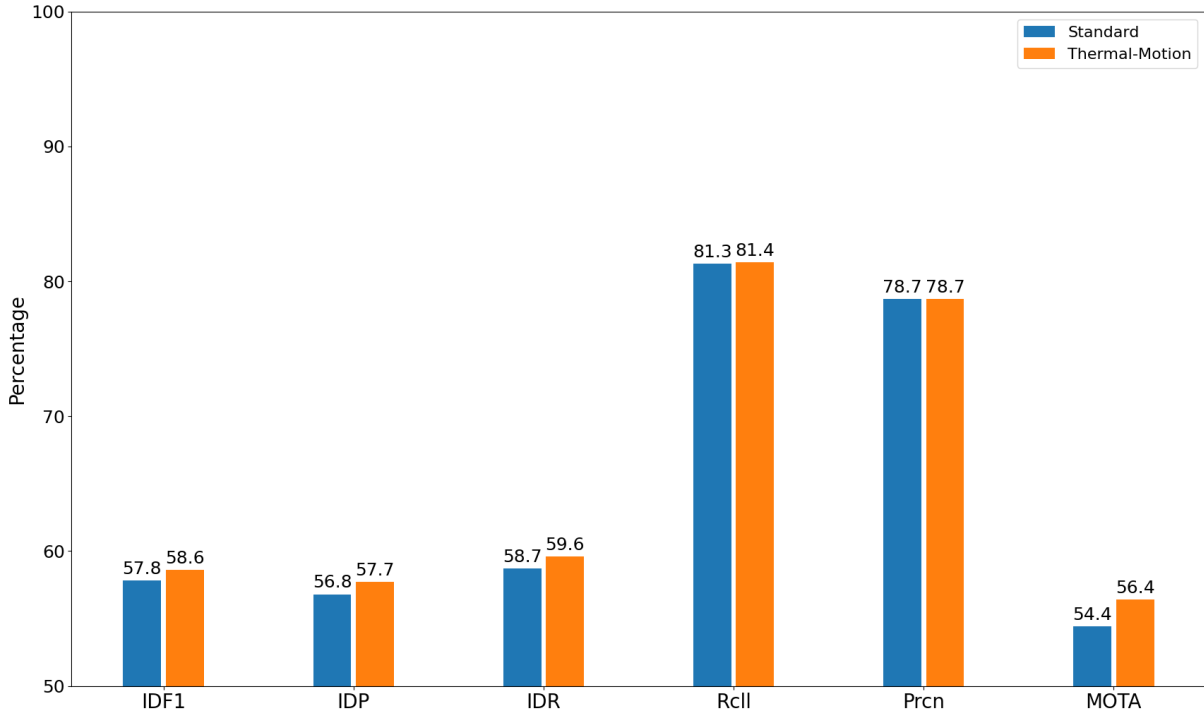


FIGURE 5.6: Overall comparison of the important MOT metrics running standard motion box association against our proposed box association algorithm using OCSort.

5.4 Conclusion

In this chapter, we have focused on enhancing the performance of MOT models operating in the thermal spectrum. Our key contribution lies in the introduction of a new box association mechanism that harnesses both motion similarity and thermal object identity. This innovative approach enhances tracking accuracy and robustness by considering not just how objects move but also their distinct thermal signatures.

In addition, we introduce a new large-scale dataset comprising both RGB and corresponding thermal images, annotated for pedestrian MOT. We anticipate that this dataset will be an invaluable resource for researchers in the fields of MOT and thermal vision perception.

Thermal cameras have emerged as robust perception sensors, unaffected by diverse weather and lighting conditions. Their unique ability to capture the inherent thermal characteristics of objects lends them an edge in computer vision tasks such as detection and tracking. We have harnessed this potential to leverage the unique features and attributes of thermal data to advance tracking systems.

Our implementation involved the collection of data using a FLIR ADK thermal sensor and a JAI GO-5100C RGB sensor, yielding the world's first (to the best of our knowledge) large-scale RGB-Thermal MOT dataset. We fine-tuned state-of-the-art object detection models on this dataset, both for RGB and thermal images. Subsequently, we benchmarked leading MOT models on the dataset and developed a novel box association method that incorporates thermal information. This approach considers the thermal and motion aspects through a weighted average, resulting in a comprehensive similarity matrix that combines the strengths of both modalities.

The results are compelling. Notably, ByteTrack and OCSort, two state-of-the-art MOT models, exhibited improved performance when our proposed box association method was employed. The fusion of thermal and motion-based similarity scores proved advantageous, making the tracking system more adaptable, robust to occlusions, and resistant to false positives and negatives.

Chapter 6

Conclusion and Future Work

Our work in Chapter 3, in addition to the previous work using thermal sensors, prove that thermal data is a rich source of information especially under poor lighting and weather conditions. They offer distinctive features for objects in a scene (thermal signature) that could be used for advanced scene perception. This, in addition to the rapid advancements in machine learning, and the growing industry of autonomous systems, raise the need for versatile perception systems that are robust to different lighting and weather conditions. Specifically for MOT, even though it is a core task in machine vision with many applications in real-world environments, there hasn't been real focus on utilizing thermal sensors for robust MOT.

In this thesis, we explored the potential of thermal sensors in the realm of computer vision, specifically focusing on multiple object detection and tracking. Our investigations led us to uncover the unique advantages of thermal sensors in challenging lighting and environmental conditions, where traditional RGB sensors tend to struggle. We have presented empirical evidence demonstrating the superior performance of thermal sensors in object detection and tracking tasks, particularly when dealing with scenarios like low light, glare, and adverse weather.

The contributions of this thesis are threefold. Firstly, we introduced two

new datasets, the FLIR ADAS Thermal Dataset and the enhanced City Scene RGB-Thermal MOT Dataset, both designed to facilitate research in thermal imaging and sensor fusion.

Secondly, we conducted extensive experiments and analysis, revealing the strengths and limitations of thermal sensors and RGB sensors in various scenarios. Our research emphasized the importance of sensor fusion, where combining data from different modalities can significantly enhance the overall perception capability of computer vision systems. This was evident in our findings, as we observed that the fusion of thermal and RGB data consistently outperformed single-modal approaches, particularly in challenging conditions.

Lastly, we proposed new models and algorithms to improve the performance of multiple object tracking in the thermal spectrum. Our innovative box association mechanism, which incorporates both motion similarity and thermal object identity, demonstrated enhanced tracking accuracy and robustness. We introduced a sigmoid-activated gating mechanism to improve fusion in object detection, showcasing the potential of learning the importance of pixels in each modality.

Overall, our research has laid the foundation for advancing the field of thermal computer vision and sensor fusion. The findings and methodologies presented in this thesis have practical implications for a wide range of applications, from autonomous vehicles and surveillance systems to robotics and augmented reality. As we continue to push the boundaries of computer vision, the integration of thermal sensors and sensor fusion techniques will play a pivotal role in enhancing the reliability and performance of computer vision systems in the real world.

6.1 Key Findings and Contributions

Our research revealed several significant findings and contributions:

- **Thermal Sensors for Object Detection:** Through extensive empirical studies, we demonstrated the efficacy of thermal sensors for object detection, especially under challenging lighting conditions. Training object detectors on thermal data consistently outperformed those trained on RGB data in scenarios with poor lighting and adverse weather.
- **Transfer Learning in Different Spectrums:** We investigated the impact of transfer learning when training detectors on datasets from different spectrums. Our experiments emphasized that transfer learning is most effective when applied within the same spectrum as the initial weights, highlighting the importance of spectral consistency in training.
- **Dynamic Confidence Cut-off (DCC) Mechanism:** We introduced a novel DCC mechanism to enhance the motion similarity association in tracking-by-detection based MOT systems. This mechanism, based on predicted bounding box size, improved tracking accuracy and robustness, particularly in challenging scenarios.
- **RGB-Thermal Sensor Fusion:** Extending our exploration into sensor fusion, we conducted in-depth experiments on fusing RGB and thermal data for 2D object detection. Our findings showcased the superiority of early fusion of data from these two modalities, emphasizing the robustness of thermal data under diverse lighting conditions.
- **Sigmoid-Activated Gating Mechanism:** We introduced a novel gating mechanism that uses a sigmoid activation function to learn the importance of each pixel in both RGB and thermal modalities. This mechanism noticeably improves fusion results when combined with different fusion operators.

- **Thermal-Based MOT Enhancement:** In the domain of multiple object tracking, we introduced a novel box association mechanism that leverages both motion similarity and thermal object identity. This innovative approach enhanced tracking accuracy and robustness by considering the distinct thermal signatures of objects.
- **RGB-Thermal MOT Dataset:** To facilitate further research in the field, we contributed a large-scale RGB-Thermal MOT dataset, annotated for pedestrian tracking. This dataset is expected to serve as a valuable resource for researchers in the areas of MOT and thermal vision perception.

6.2 Implications

The findings and contributions of this research have several important implications:

- **Sensor Fusion for Enhanced Perception:** The study highlights the importance of sensor fusion in computer vision systems. Combining data from different spectrums, such as RGB and thermal, enhances the robustness, adaptability, and reliability of autonomous systems, particularly in real-world scenarios with variable lighting and weather conditions. It also highlights the importance of carefully selecting the best fusion operator depending on the particular model used and application domain.
- **Practical Applications:** The research outcomes have direct implications for real-world applications, including autonomous vehicles, surveillance systems, and security applications. Leveraging thermal sensors in these domains can significantly improve the safety and performance of these systems.

- **Cross-modality Transfer Learning Analysis:** Our findings on transfer learning shed light on the importance of domain consistency in model training. Understanding the impact of data spectrum on transfer learning can guide the development of more effective training strategies. We show that applying cross-modality transfer learning is still effective, but not as much as intra-modality transfer learning.

6.3 Future Work

This Ph.D. thesis opens avenues for future research in several directions:

- **Advanced Fusion Methods:** Future work can explore more advanced fusion methods for combining RGB and thermal data, including deep learning-based approaches that learn optimal fusion strategies. It is also possible to explore methods like fusion-architecture search, similar to neural architecture search where the process of finding the most efficient fusion operator for a particular task is automated.
- **Thermal-Only MOT:** Investigating and developing multiple object tracking systems that operate exclusively in the thermal spectrum can be a promising direction, leveraging the unique advantages of thermal sensors. Our experiments from Chapter 5 provide promising results for thermal-only MOT models.
- **Real-World Deployments:** Evaluating the practical deployment of thermal-based computer vision systems in real-world applications, such as autonomous vehicles and surveillance networks, is crucial for translating research advancements into tangible benefits.

In conclusion, this Ph.D. thesis has contributed to advancing the field of computer vision by exploring the untapped potential of thermal sensors, sensor fusion, and enhanced perception in challenging real-world

scenarios. The findings and contributions presented here provide valuable insights for researchers, engineers, and developers striving to create more robust and reliable computer vision systems for the future.

Bibliography

- [1] Nouman Ali, Danish Ali Mazhar, Zeshan Iqbal, Rehan Ashraf, Jawad Ahmed, and Farrukh Zeeshan Khan. "Content-based image retrieval based on late fusion of binary and local descriptors". In: *arXiv preprint arXiv:1703.08492* (2017).
- [2] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. "Efficient pedestrian detection at nighttime using a thermal camera". In: *Sensors* 17.8 (2017), p. 1850.
- [3] Greet Baldewijns, Glen Debar, Gert Mertes, Tom Croonenborghs, and Bart Vanrumste. "Improving the accuracy of existing camera based fall detection algorithms through late fusion". In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2017, pp. 2667–2671.
- [4] George Barnum, Sabera Talukder, and Yisong Yue. "On the benefits of early fusion in multimodal representation learning". In: *arXiv preprint arXiv:2011.07191* (2020).
- [5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. "Tracking without bells and whistles". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 941–951.
- [6] Keni Bernardin and Rainer Stiefelhagen. "Evaluating multiple object tracking performance: the clear mot metrics". In: *EURASIP Journal on Image and Video Processing* 2008 (2008), pp. 1–10.

- [7] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Uptcroft. “Simple online and realtime tracking”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468.
- [8] Anil Bhattacharyya. “On a measure of divergence between two statistical populations defined by their probability distribution”. In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–110.
- [9] Guillem Brasó and Laura Leal-Taixé. “Learning a neural solver for multiple object tracking”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6247–6257.
- [10] Garrick Brazil, Xi Yin, and Xiaoming Liu. “Illuminating Pedestrians via Simultaneous Detection and Segmentation”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, June 2017. DOI: [10.1109/iccv.2017.530](https://doi.org/10.1109/iccv.2017.530).
- [11] Daniel Skomedal Breland, Aveen Dayal, Ajit Jha, Phaneendra K Yalavarthy, Om Jee Pandey, and Linga Reddy Cenkeramaddi. “Robust hand gestures recognition using a deep CNN and thermal images”. In: *IEEE Sensors Journal* 21.23 (2021), pp. 26602–26614.
- [12] Zhaowei Cai and Nuno Vasconcelos. “Cascade R-CNN: Delving into High Quality Object Detection”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. DOI: [10.1109/cvpr.2018.00644](https://doi.org/10.1109/cvpr.2018.00644).
- [13] Ej Miguel Francisco Caliwag, Angela Caliwag, Bong-Ki Baek, Yongrae Jo, Hae Chung, and Wansu Lim. “Distance estimation in thermal cameras using multi-task cascaded convolutional neural network”. In: *IEEE Sensors Journal* 21.17 (2021), pp. 18519–18525.

- [14] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. "Observation-centric sort: Rethinking sort for robust multi-object tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 9686–9696.
- [15] Injoong Chang, Ji-Yeul Bae, Namkyu Lee, Hwkyuen Kwak, and Hyung Hee Cho. "Thermal signature characteristics of clothed human considering thermoregulation effects". In: *Journal of the Computational Structural Engineering Institute of Korea* 32.2 (2019), pp. 109–116.
- [16] Chen Chen, Bin Liu, Shaohua Wan, Peng Qiao, and Qingqi Pei. "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system". In: *IEEE Transactions on Intelligent Transportation Systems* 22.3 (2020), pp. 1840–1852.
- [17] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. "MMDetection: Open MMLab Detection Toolbox and Benchmark". In: *arXiv preprint arXiv:1906.07155* (2019).
- [18] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. "Multi-view 3D Object Detection Network for Autonomous Driving". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Nov. 2016. DOI: [10.1109/cvpr.2017.691](https://doi.org/10.1109/cvpr.2017.691).
- [19] Peng Chu and Haibin Ling. "Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6172–6181.

- [20] MMTracking Contributors. *MMTracking: OpenMMLab video perception toolbox and benchmark*. <https://github.com/open-mmlab/mtracking>. 2020.
- [21] Eduardo R Corral-Soto and Liu Bingbing. “Understanding strengths and weaknesses of complementary sensor modalities in early fusion for object detection”. In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2020, pp. 1785–1792.
- [22] Christyan Cruz Ulloa, Guillermo Prieto Sánchez, Antonio Barrientos, and Jaime Del Cerro. “Autonomous thermal vision robotic system for victims recognition in search and rescue missions”. In: *Sensors* 21.21 (2021), p. 7346.
- [23] Xuerui Dai, Xue Yuan, and Xueye Wei. “TIRNet: Object detection in thermal infrared images for autonomous driving”. In: *Applied Intelligence* 51.3 (2021), pp. 1244–1261.
- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [25] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. “CenterNet: Keypoint Triplets for Object Detection”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. DOI: [10.1109/iccv.2019.00667](https://doi.org/10.1109/iccv.2019.00667). arXiv: [1904.08189](https://arxiv.org/abs/1904.08189).
- [26] Wassim El Ahmar, Yahya Massoud, Dhanvin Kolhatkar, Hamzah Al-Ghamdi, Mohammad Alja’afreh, Riad Hammoud, and Robert Laganieri. “Enhanced Thermal-RGB Fusion for Robust Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 365–374.

- [27] Wassim A El Ahmar, Dhanvin Kolhatkar, Farzan Erlik Nowruzi, Hamzah AlGhamdi, Jonathan Hou, and Robert Laganieri. "Multiple Object Detection and Tracking in the Thermal Spectrum". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 277–285.
- [28] Farzan Erlik Nowruzi, Wassim A El Ahmar, Robert Laganieri, and Amir H Ghods. "In-vehicle occupancy detection with convolutional networks on thermal images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.
- [29] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. "Tood: Task-aligned one-stage object detection". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society. 2021, pp. 3490–3499.
- [30] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.
- [31] Xu Gao and Tingting Jiang. "Osmo: Online specific models for occlusion in multiple object tracking under surveillance scene". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 201–210.
- [32] Cinmayii A Garillos-Manliguez and John Y Chiang. "Multimodal Deep Learning via Late Fusion for Non-Destructive Papaya Fruit Maturity Classification". In: *2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. IEEE. 2021, pp. 1–6.

- [33] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. “Yolox: Exceeding yolo series in 2021”. In: *arXiv preprint arXiv:2107.08430* (2021).
- [34] Ross Girshick. “Fast R-CNN”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015. DOI: [10.1109/iccv.2015.169](https://doi.org/10.1109/iccv.2015.169). eprint: [1504.08083](https://arxiv.org/abs/1504.08083).
- [35] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Nov. 2014. DOI: [10.1109/cvpr.2014.81](https://doi.org/10.1109/cvpr.2014.81). eprint: [1311.2524](https://arxiv.org/abs/1311.2524).
- [36] Adam Glowacz. “Fault diagnosis of electric impact drills using thermal imaging”. In: *Measurement* 171 (2021), p. 108815.
- [37] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. “Pedestrian Detection: Domain Generalization, CNNs, Transformers and Beyond”. In: *ECCV2022*. Jan. 2022.
- [38] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. DOI: [10.1109/iccv.2017.322](https://doi.org/10.1109/iccv.2017.322). eprint: [1703.06870](https://arxiv.org/abs/1703.06870).
- [39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Dec. 2015). DOI: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90). arXiv: [1512.03385](https://arxiv.org/abs/1512.03385).
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

- [41] Peiyun Hu and Deva Ramanan. "Finding Tiny Faces". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Dec. 2016. DOI: [10.1109/cvpr.2017.166](https://doi.org/10.1109/cvpr.2017.166).
- [42] Deng-Yuan Huang and Chia-Hung Wang. "Optimal multi-level thresholding using a two-stage Otsu optimization approach". In: *Pattern Recognition Letters* 30.3 (2009), pp. 275–284.
- [43] Nianchang Huang, Qiang Jiao, Qiang Zhang, and Jungong Han. "Middle-level feature fusion for lightweight RGB-D salient object detection". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 6621–6634.
- [44] Ino. *Video analytics dataset - technologies*. URL: <https://www.ino.ca/en/technologies/video-analytics-dataset/>.
- [45] Rudolph Emil Kalman. "A new approach to linear filtering and prediction problems". In: (1960).
- [46] Jin Kyu Kang, Toan Minh Hoang, and Kang Ryoung Park. "Person re-identification between visible and thermal camera images based on deep residual CNN using single input". In: *IEEE Access* 7 (2019), pp. 57972–57984.
- [47] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Dimitris Dres, and Dimitris Drakoulis. "Man overboard event detection from RGB and thermal imagery: Possibilities and limitations". In: *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. 2020, pp. 1–6.
- [48] Aleksandr Kim, Aljoša Ošep, and Laura Leal-Taixé. "Eagermot: 3d multi-object tracking via sensor fusion". In: *2021 IEEE International conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 11315–11321.

- [49] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. "Multispectral transfer network: Unsupervised depth estimation for all-day vision". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [50] Albert Ko and Henry YK Lau. "Robot assisted emergency search and rescue system with a wireless sensor network". In: *International Journal of Advanced Science and Technology* 3.4 (2009), pp. 69–78.
- [51] Albert WY Ko and Henry YK Lau. "Intelligent robot-assisted humanitarian search and rescue system". In: *International Journal of Advanced Robotic Systems* 6.2 (2009), p. 12.
- [52] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. "Thermal object detection in difficult weather conditions using YOLO". In: *IEEE access* 8 (2020), pp. 125459–125476.
- [53] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. "Thermal object detection in difficult weather conditions using YOLO". In: *IEEE access* 8 (2020), pp. 125459–125476.
- [54] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. "Joint 3D Proposal Generation and Object Detection from View Aggregation". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Dec. 2018. DOI: [10.1109/iros.2018.8594049](https://doi.org/10.1109/iros.2018.8594049). eprint: [1712.02294](https://arxiv.org/abs/1712.02294).
- [55] Wahyu Kurniawan, Sutrisno Ibrahim, and Meiyanto Sulisty. "People detection and tracking methods for intelligent surveillance system". In: *AIP Conference Proceedings*. Vol. 2217. 1. AIP Publishing. 2020.
- [56] Alicja Kwaśniewska, Jacek Rumiński, and Paul Rad. "Deep features class activation map for thermal face detection and tracking". In: *2017 10Th international conference on human system interactions (HSI)*. IEEE. 2017, pp. 41–47.

- [57] Redouan Lahmyed, Mohamed El Ansari, and Ayoub Ellahyani. "A new thermal infrared and visible spectrum images-based pedestrian detection system". In: *Multimedia Tools and Applications* 78.12 (2019), pp. 15861–15885.
- [58] Ichraf Lahouli, Rob Haelterman, Zied Chtourou, Geert De Cubber, and Rabah Attia. "Pedestrian Detection and Tracking in Thermal Images from Aerial MPEG Videos." In: *VISIGRAPP (5: VISAPP)*. 2018, pp. 487–495.
- [59] Hei Law and Jia Deng. "CornerNet: Detecting Objects as Paired Keypoints". In: *International Journal of Computer Vision* 128.3 (2018), pp. 642–656. DOI: [10.1007/s11263-019-01204-1](https://doi.org/10.1007/s11263-019-01204-1). arXiv: [1808.01244](https://arxiv.org/abs/1808.01244).
- [60] Ji Hoon Lee, Jong-Suk Choi, Eun Som Jeon, Yeong Gon Kim, Toan Thanh Le, Kwang Yong Shin, Hyeon Chang Lee, and Kang Ryoung Park. "Robust pedestrian detection by combining visible and thermal infrared cameras". In: *Sensors* 15.5 (2015), pp. 10580–10615.
- [61] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. "LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 392–404. DOI: [10.1109/tip.2021.3130533](https://doi.org/10.1109/tip.2021.3130533).
- [62] Jingcheng Li, Lina Yao, Binghao Li, and Claude Sammut. "Distilled Mid-Fusion Transformer Networks for Multi-Modal Human Activity Recognition". In: *arXiv preprint arXiv:2305.03810* (2023).
- [63] Chao Liang, Zhipeng Zhang, Yi Lu, Xue Zhou, Bing Li, Xiyong Ye, and Jianxiao Zou. "Rethinking the competition between detection and reid in multi-object tracking". In: *arXiv preprint arXiv:2010.12138* (2020).
- [64] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. "Deep Continuous Fusion for Multi-Sensor 3D Object Detection". In: *ECCV 2018*. Cham: Springer International Publishing, 2018, 663–678.

- [65] Daniele Liciotti, Marina Paolanti, Emanuele Frontoni, and Primo Zingaretti. "People detection and tracking from an RGB-D camera in top-view configuration: review of challenges and applications". In: *New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers* 19. Springer. 2017, pp. 207–218.
- [66] Teck-Yian Lim, Amin Ansari, Bence Major, Daniel Fontijne, Michael Hamilton, Radhika Gowaikar, and Sundar Subramanian. "Radar and camera early fusion for vehicle detection in advanced driver assistance systems". In: *Machine learning for autonomous driving workshop at the 33rd conference on neural information processing systems*. Vol. 2. 7. 2019.
- [67] Jia-Wei Lin, Ming-Hung Lu, and Yuan-Hsiang Lin. "A thermal camera based continuous body temperature measurement system". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.
- [68] Kai Lin, Yihui Li, Jinchuan Sun, Dongsheng Zhou, and Qiang Zhang. "Multi-sensor fusion for body sensor network in medical human–robot interaction scenario". In: *Information Fusion* 57 (2020), pp. 15–26.
- [69] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [70] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

- [71] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. “Focal Loss for Dense Object Detection”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. DOI: [10.1109/iccv.2017.324](#). arXiv: [1708.02002](#).
- [72] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. “PTB-TIR: A Thermal Infrared Pedestrian Tracking Benchmark”. In: *IEEE Transactions on Multimedia* 22.3 (2020), pp. 666–675. DOI: [10.1109/tmm.2019.2932615](#).
- [73] Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, and Feng Zheng. “LSOTB-TIR: A Large-Scale High-Diversity Thermal Infrared Object Tracking Benchmark”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020. DOI: [10.1145/3394171.3413922](#).
- [74] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. “SSD: Single Shot Multi-Box Detector”. In: *Computer Vision – ECCV 2016* (Dec. 2015), pp. 21–37. DOI: [10.1007/978-3-319-46448-0_2](#).
- [75] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. “Late fusion incomplete multi-view clustering”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.10 (2018), pp. 2410–2423.
- [76] Ye Liu, Xiao-Yuan Jing, Jianhui Nie, Hao Gao, Jun Liu, and Guo-Ping Jiang. “Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos”. In: *IEEE Transactions on Multimedia* 21.3 (2018), pp. 664–677.
- [77] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. “Bevfusion: Multi-task multi-sensor

- fusion with unified bird's-eye view representation". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 2774–2781.
- [78] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jan. 2022). DOI: [10.1109/cvpr52688.2022.01167](https://doi.org/10.1109/cvpr52688.2022.01167). eprint: [2201.03545](https://arxiv.org/abs/2201.03545).
- [79] Yahya Massoud. "Sensor Fusion for 3D Object Detection for Autonomous Vehicles". MA thesis. Université d'Ottawa/University of Ottawa, 2021.
- [80] Gaetano Messina and Giuseppe Modica. "Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook". In: *Remote Sensing* 12.9 (2020), p. 1491.
- [81] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv:1603.00831 [cs]* (Mar. 2016). arXiv: 1603.00831. URL: <http://arxiv.org/abs/1603.00831>.
- [82] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. "Real-time hand tracking under occlusion from an egocentric rgb-d sensor". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1154–1163.
- [83] Valentina Muşat, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. "Multi-weather city: Adverse weather stacking for autonomous driving". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2906–2915.
- [84] Mohammad Amin Nabian and Hadi Meidani. "Deep learning for accelerated seismic reliability analysis of transportation networks". In: *Computer-Aided Civil and Infrastructure Engineering* 33.6 (2018), pp. 443–458.

- [85] Francisco Soares Neves, Rafael Marques Claro, and Andry Maykol Pinto. "End-to-end detection of a landing platform for offshore UAVs based on a multimodal early fusion approach". In: *Sensors* 23.5 (2023), p. 2434.
- [86] Hoang Nguyen, Le-Minh Kieu, Tao Wen, and Chen Cai. "Deep learning methods in transportation domain: a review". In: *IET Intelligent Transport Systems* 12.9 (2018), pp. 998–1004.
- [87] Erkan Oğuz, Ayhan Küçükmanisa, Ramazan Duvar, and Oğuzhan Urhan. "A deep learning based fast lane detection approach". In: *Chaos, Solitons & Fractals* 155 (2022), p. 111722.
- [88] Timo Ojala, Matti Pietikäinen, and David Harwood. "A comparative study of texture measures with classification based on featured distributions". In: *Pattern recognition* 29.1 (1996), pp. 51–59.
- [89] Joshua Olusola, Jumana Abu-Khalaf, Sidike Paheding, and Nathir Rawashdeh. "LiDAR and RGB camera based feature extraction and evaluation for adverse weather driving". In: *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Vol. 12540. 2023, p. 125400I.
- [90] Khalid Omer and Meredith Kupinski. "Mid-fusion of road scene polarization images on pretrained RGB neural networks". In: *JOSA A* 38.4 (2021), pp. 515–525.
- [91] Gwanyong Park, Minhyung Lee, Hyangin Jang, and Changmin Kim. "Thermal anomaly detection in walls via CNN-based segmentation". In: *Automation in Construction* 125 (2021), p. 103627.
- [92] Morteza Mousa Pasandi, Tianran Liu, Yahya Massoud, and Robert Laganière. "Sensor Fusion Operators for Multimodal 2D Object Detection". In: *ISVC 2022*. Vol. 13598. LNCS. Springer, 2022, 184–195.

- [93] Janez Perš, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec, and Stanislav Kovačič. “Histograms of optical flow for efficient representation of body motion”. In: *Pattern Recognition Letters* 31.11 (2010), pp. 1369–1376.
- [94] Tomáš Petříček, Vojtěch Šalanský, Karel Zimmermann, and Tomáš Svoboda. “Simultaneous exploration and segmentation for search and rescue”. In: *Journal of Field Robotics* 36.4 (2019), pp. 696–709.
- [95] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. DOI: [10.1109/cvpr.2016.91](https://doi.org/10.1109/cvpr.2016.91). eprint: [1506.02640](https://arxiv.org/abs/1506.02640).
- [96] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (June 2015), pp. 1137–1149. DOI: [10.1109/tpami.2016.2577031](https://doi.org/10.1109/tpami.2016.2577031). eprint: [1506.01497](https://arxiv.org/abs/1506.01497).
- [97] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. “Performance measures and a data set for multi-target, multi-camera tracking”. In: *European conference on computer vision*. Springer, 2016, pp. 17–35.
- [98] Jesús Joel Rivas, Felipe Orihuela-Espina, and Luis Enrique Sucar. “Recognition of affective states in virtual rehabilitation using late fusion with Semi-Naive Bayesian classifier”. In: *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 2019, pp. 308–313.
- [99] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

- [100] Francesco Secci and Andrea Ceccarelli. “On failures of RGB cameras and their effects in autonomous driving applications”. In: *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*. IEEE. 2020, pp. 13–24.
- [101] A Selvarani and GR Suresh. “Infrared thermal imaging for diabetes detection and measurement”. In: *Journal of medical systems* 43 (2019), pp. 1–11.
- [102] Haidong Shao, Min Xia, Guangjie Han, Yu Zhang, and Jiafu Wan. “Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images”. In: *IEEE Transactions on Industrial Informatics* 17.5 (2020), pp. 3488–3496.
- [103] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. “Safety-enhanced autonomous driving using interpretable sensor fusion transformer”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 726–737.
- [104] Mario Silvagni, Andrea Tonoli, Enrico Zenerino, and Marcello Chiaberge. “Multipurpose UAV for search and rescue operations in mountain avalanche events”. In: *Geomatics, Natural Hazards and Risk* 8.1 (2017), pp. 18–33.
- [105] Severin Stalder, Helmut Grabner, and Luc Van Gool. “Cascaded confidence filtering for improved tracking-by-detection”. In: *European Conference on Computer Vision*. Springer. 2010, pp. 369–382.
- [106] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. “Computer vision in the infrared spectrum: challenges and approaches”. In: *Synthesis Lectures on Computer Vision* 10.2 (2021), pp. 1–138.

- [107] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. "Computer vision in the infrared spectrum: challenges and approaches". In: *Synthesis Lectures on Computer Vision* 10.2 (2021), pp. 1–138.
- [108] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications". In: *Computer Vision and Image Understanding* 116.2 (2012), pp. 210–221.
- [109] Ioannis Vernikos, Dimitrios Koutrintzes, Eirini Mathe, Evaggelos Spyrou, and Phivos Mylonas. "Early fusion of visual representations of skeletal data for human activity recognition". In: *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*. 2022, pp. 1–5.
- [110] Nataša Vlahović and Zeljko Djurovic. "Robust tracking of moving objects using thermal camera and speeded up robust features descriptor". In: *International Journal of Adaptive Control and Signal Processing* 35.4 (2021), pp. 549–566.
- [111] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "Scaled-YOLOv4: Scaling Cross Stage Partial Network". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). DOI: [10.1109/cvpr46437.2021.01283](https://doi.org/10.1109/cvpr46437.2021.01283).
- [112] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *arXiv preprint arXiv:2207.02696*. 2022.
- [113] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. "Cross-modality paired-images generation for RGB-infrared person re-identification". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 12144–12151.

- [114] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 3623–3632.
- [115] Jiawei Wang, Ruixiang Chen, and Zhaocheng He. "Traffic speed prediction for urban transportation network: A path based deep learning approach". In: *Transportation Research Part C: Emerging Technologies* 100 (2019), pp. 372–385.
- [116] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. "Deep Parametric Continuous Convolutional Neural Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. DOI: [10.1109/cvpr.2018.00274](https://doi.org/10.1109/cvpr.2018.00274). arXiv: [2101.06742](https://arxiv.org/abs/2101.06742).
- [117] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. "Repulsion Loss: Detecting Pedestrians in a Crowd". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Nov. 2017. DOI: [10.1109/cvpr.2018.00811](https://doi.org/10.1109/cvpr.2018.00811).
- [118] Samuel Morton Williams, Sara Bariselli, Cristiano Palego, Richard Holland, and Paul Cross. "A comparison of machine-learning assisted optical and thermal camera systems for beehive activity counting". In: *Smart Agricultural Technology* 2 (2022), p. 100038.
- [119] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

- [120] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Mar. 2017. DOI: [10.1109/icip.2017.8296962](https://doi.org/10.1109/icip.2017.8296962). eprint: [1703.07402](https://arxiv.org/abs/1703.07402).
- [121] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. "Mid-fusion: Octree-based object-level multi-instance dynamic slam". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 5231–5237.
- [122] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. "Spatial-temporal relation networks for multi-object tracking". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3988–3998.
- [123] Shuguan Yang, Wei Ma, Xidong Pi, and Sean Qian. "A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 248–265.
- [124] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. "Bi-directional center-constrained top-ranking for visible thermal person re-identification". In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 407–419.
- [125] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. "Sensor and sensor fusion technology in autonomous vehicles: A review". In: *Sensors* 21.6 (2021), p. 2140.
- [126] Kwangjin Yoon, Jeonghwan Gwak, Young-Min Song, Young-Chul Yoon, and Moon-Gu Jeon. "Oneshotda: Online multi-object tracker with one-shot-learning-based data association". In: *IEEE Access* 8 (2020), pp. 38060–38072.

- [127] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. “Varifocalnet: An iou-aware dense object detector”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8514–8523.
- [128] Jun Zhang, Yiyao Liu, Mingxing Wen, Yufeng Yue, Haoyuan Zhang, and Danwei Wang. “L 2 V 2 T 2 Calib: Automatic and Unified Extrinsic Calibration Toolbox for Different 3D LiDAR, Visual Camera and Thermal Camera”. In: *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2023, pp. 1–7.
- [129] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. “Bytetrack: Multi-object tracking by associating every detection box”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer. 2022, pp. 1–21.
- [130] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”. In: *arXiv preprint arXiv:2110.06864* (2021).
- [131] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. “Fairmot: On the fairness of detection and re-identification in multiple object tracking”. In: *International Journal of Computer Vision* 129.11 (2021), pp. 3069–3087.
- [132] LIN Zong-ze, TANG Chao-sheng, ZENG Hao, CHENG Qing, TIAN Ben-gang, and SHI Bin. “Soil evaporation based on infrared thermal imaging technology”. In: *Chinese Journal of Geotechnical Engineering* 43.4 (2021), pp. 743–750.
- [133] Juan Zuluaga-Gomez, Zeina Al Masry, Khaled Benaggoune, Safa Meraghi, and Nourredine Zerhouni. “A CNN-based methodology for

breast cancer diagnosis using thermal images". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 9.2 (2021), pp. 131–145.