論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Learning-Based Pedestrian Detection on Visible and Thermal Images with Multiple Regressors Considering Misalignment
著者(和文)	WANCHAITANAWONGNapat
Author(English)	Napat Wanchaitanawong
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12852号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:田中 正行,塚越 秀行,中臺 一博,原 精一郎,川上 玲,奥冨 正敏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12852号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

TOKYO INSTITUTE OF TECHNOLOGY

Learning-Based Pedestrian Detection on Visible and Thermal Images with Multiple Regressors Considering Misalignment

Napat Wanchaitanawong

A dissertation submitted in fulfillment of the
requirements for the degree of Doctor of Engineering

in the
School of Engineering
Department if Systems and Control Engineering

September 2024

Declaration of Authorship

I, Napat WANCHAITANAWONG, declare that this dissertation titled, "Learning-Based Pedestrian Detection on Visible and Thermal Images with Multiple Regressors Considering Misalignment" and the work presented in it are my own. I confirm that:

• This work was done wholly or mainly while in candidature for a research degree at this University.

• Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

• Where I have consulted the published work of others, this is always clearly attributed.

• Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.

• I have acknowledged all main sources of help.

• Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Pedestrian detection, essential in computer vision, has advanced significantly with the integration of visible and thermal sensors, enhancing robustness in varying illumination and cluttered environments. However, the assumption of perfect alignment between modalities often reduces accuracy in real-world scenarios due to differences in sensor viewpoints, calibration errors, and temporal asynchronization. This dissertation proposes novel methods to address the misalignment problem in multi-modal pedestrian detection. In this work, the term "multi-moda" specifically refers to the combination of visible and thermal sensors. It begins with a comprehensive review of existing pedestrian detection methods, covering both traditional techniques and deep learning-based approaches. This review highlights the pressing challenges, particularly the issue of misalignment between visible and thermal modalities, guiding the development of our novel detection methods. To tackle these challenges, we develop new evaluation metrics, including the multi-modal Intersection over Union (IoU^{M}) and Multi-Modal Log-Average Miss Rate (MR^{M}), to accurately measure detection precision across different sensor modalities. The dissertation then introduces two advanced multi-modal pedestrian detection frameworks. The multi-modal Faster R-CNN framework modifies both the Region Proposal Network (RPN) and the detector to incorporate a multi-modal regressor, generating bounding box pairs that accurately locate pedestrians in both modalities despite significant misalignment. The multi-modal Single Shot MultiBox Detector (SSD) framework enhances real-time detection capabilities and improves accuracy and robustness in misaligned environments by incorporating multi-modal regressors, object-based training techniques, and shifting data augmentation. Key contributions of this work include the development of the multi-modal IoU^{M} and MR^{M} metrics, the implementation of a multi-modal regressor architecture, which generates unique bounding box pairs to accurately locate pedestrians in both modalities, and the introduction of object-based training specifically designed for paired data from different modalities. Both frameworks are validated through extensive experiments, including simulated disparity tests, demonstrating superior performance in misaligned scenarios on the KAIST Multispectral Pedestrian Dataset. Additionally, ablation studies verify the effectiveness of multi-modal regressors compared to traditional single-modal regressors and the utility of IoU^{M} as a criterion. In the final chapters, the dis-

sertation discusses the main findings and contributions, providing a detailed analysis of the experimental results. The discussion emphasizes the strengths and limitations of the proposed methods, offering insights for future research directions. This work significantly advances pedestrian detection technology in misalignment environments, paving the way for safer and more reliable applications in autonomous driving and surveillance. Future work will focus on further enhancing alignment techniques, integrating real-time processing capabilities, and exploring advanced fusion strategies to continue advancing this critical field.

Acknowledgement

First and foremost, I would like to thank my academic supervisor Professor Masutoshi Okutomi and Professor Masayuki Tanaka for choosing me as their disciple in International Graduate Program of Tokyo Institute of Technology, for giving me an invaluable opportunity to extend my knowledge under their guidance, and for supporting and supervising my study continually. Without them, this dissertation or even my study would not have become complete. I am very grateful for all of their contribution.

I also would like to thank Takashi Shibata from NTT Corporation for his support and inspiration he gave me all the time during my work. His advice and suggestion helped improving my work to a great extent.

Lastly, I would like to thank MEXT (Japanese Ministry of Education, Culture, Science and Technology) for their financial support. Their funding allowed me to smoothly focus on my study.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgement iv

1 Introduction 1
- 1.1 Background and Motivation of Overall Research 1
- 1.2 Objectives and Contributions 7
- 1.3 Dissertation Organization 9

2 Existing Methods Review 12
- 2.1 Pedestrian Detection 12
  - 2.1.1 Introduction 12
  - 2.1.2 Traditional Pedestrian Detection Methods 12
  - 2.1.3 Deep Learning-Based Pedestrian Detection Methods . . 14
  - 2.1.4 Limitations of Pedestrian Detection Methods 21
- 2.2 Multi-Modal Pedestrian Detection 21
  - 2.2.1 Introduction 21
  - 2.2.2 Naive Feature Fusion 22
  - 2.2.3 Adaptive Feature Fusion 23
  - 2.2.4 Challenges 24

3 Proposed Evaluation Metrics 26
- 3.1 Multi-Modal IoU 26
- 3.2 Multi-Modal MR 29

4 Proposed Multi-Modal Faster R-CNN Considering Misalignment 31
- 4.1 Background 31
- 4.2 Methodology 37
  - 4.2.1 Multi-Modal RPN 37
  - 4.2.2 Multi-Modal Detector 42
  - 4.2.3 Multi-modal NMS 43
  - 4.2.4 Multi-Modal Mini Batch Sampling 45

4.3 Experiment . . . 46
4.3.1 Dataset . . . 46
4.3.2 Implementation Details . . . 46
4.3.3 Evaluation Details . . . 47
4.3.4 Comparison with Existing Methods . . . 48
4.3.5 Ablation Study . . . 52
4.4 Conclusion . . . 54

5 Proposed Multi-Modal Single Shot MultiBox Detector Considering Misalignment 57
5.1 Background . . . 57
5.2 Methodology . . . 61
5.2.1 Multi-Modal Regressor For Single-Stage Network . . . 62
5.2.2 Object-Based Training . . . 64
5.2.3 Multi-Modal NMS . . . 65
5.2.4 Shifting Data Augmentation . . . 65
5.3 Experiment . . . 67
5.3.1 Dataset . . . 67
5.3.2 Implementation and Details . . . 67
5.3.3 Evaluation Details . . . 68
5.3.4 Comparison with existing methods . . . 70
5.3.5 Ablation Study . . . 73
5.4 Conclusion . . . 73

6 Discussion and Conclusion 75
6.1 Discussion . . . 75
6.2 Conclusion . . . 81
6.3 Future Work . . . 82

Bibliography 84

Chapter 1

Introduction

In this chapter, we introduce the background, motivation, and objectives of this research. Additionally, the structure of the thesis is outlined to guide the reader through the subsequent chapters.

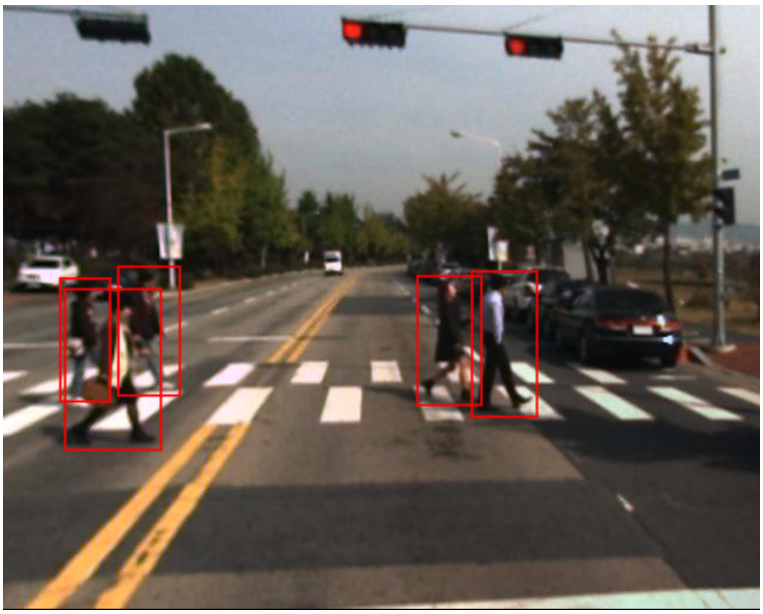


Figure 1.1: Example of pedestrian detection result on a color image. Red boxes represent predicted bounding boxes.

1.1 Background and Motivation of Overall Research

Pedestrian detection is a crucial task in computer vision, essential for applications ranging from surveillance to autonomous driving. This task involves identifying and locating pedestrians within an image or video stream. Accurate pedestrian detection is vital for ensuring the safety and efficiency of

these systems. For instance, in the context of autonomous vehicles, reliable pedestrian detection is paramount to prevent accidents and ensure the safety of both pedestrians and passengers. In surveillance, accurate pedestrian detection enhances security by enabling the monitoring of public spaces, identifying suspicious activities, and preventing potential threats. Figure 1.1 illustrates an example of pedestrian detection results on a color image, where red boxes represent predicted bounding boxes

A pedestrian is defined as a person who is walking, particularly in areas where vehicles are present. This term is often used in the context of traffic environments to emphasize individuals on foot who interact with vehicular traffic. However, pedestrian detection is not essentially the same as human detection. Pedestrian detection specifically targets individuals who are walking or moving on foot, excluding those who are riding vehicles like bicycles or motorcycles. On the other hand, human detection encompasses a broader scope, aiming to identify any person regardless of their specific activity or mode of transportation, such as walking, standing, or riding a bike. Consequently, human detection systems are designed to detect all humans, including those who are stationary or traveling on vehicles. Therefore, while pedestrian detection should not classify bikers or motorcyclists as pedestrians, human detection should include all these cases to ensure comprehensive coverage of all human activities. Moreover, the term "pedestrian detection" is commonly used because the primary application involves detecting people in traffic environments, where the focus is on ensuring the safety of individuals navigating areas shared with vehicles. This terminology is consistent with the usage in existing research and practical applications. Additionally, many pedestrian datasets distinguish between pedestrians and people traveling by vehicle. For example, the Cityscapes Dataset [2, 62] includes classes for both pedestrians and riders (cyclists and motorcyclists). Similarly, the KITTI Dataset [14] and the KAIST Multispectral Pedestrian Dataset [24] include pedestrians and cyclists as separate classes. These datasets help illustrate the importance of distinguishing between pedestrian detection and broader human detection. However, in the trend of multi-modal pedestrian detection on the KAIST dataset, only the pedestrian class is typically considered; cyclists are classified as pedestrians. This practice is also reflected in the evaluation benchmark for the KAIST dataset. In our work, we follow this convention by considering only the pedestrian class in our evaluations.

Pedestrian detection, while being a subset of the broader field of object detection, presents unique challenges. Object detection involves identifying

and locating various objects within an image, including categories such as vehicles, animals, and household items. Pedestrian detection focuses exclusively on human figures and must handle a wide range of poses, occlusions, and varying appearances. This specificity and complexity make pedestrian detection more challenging than general object detection, which deals with a broader range of object types and is often less affected by the high variability in appearance that human figures present. Additionally, pedestrian detection systems are typically optimized for the unique characteristics of human shapes and movements, often resulting in higher accuracy for this particular task compared to general object detection networks. Pedestrian detection models are trained on datasets rich in pedestrian instances, allowing them to learn more detailed and relevant features specific to pedestrian detection. Furthermore, many research studies and practical applications focus on pedestrian detection in traffic environments, contributing to the continuous improvement and refinement of these models.

Bounding boxes (BBs) are a standard and efficient method for object detection, particularly in real-time applications. Using BBs allows for simpler and faster calculations, making the detection process less computationally intensive compared to segmentation, which requires detailed pixel-wise classification. Bounding boxes provide sufficient accuracy for many applications, such as pedestrian detection, where the primary goal is to locate and track humans within a frame. While segmentation can offer more detailed information, it is often unnecessary for the task of detecting and tracking pedestrians, where the precision of bounding boxes is adequate to achieve the desired objectives. Additionally, the computational simplicity of BBs makes them more suitable for real-time applications like autonomous driving and surveillance, where quick and reliable detections are critical.

Early pedestrian detection methods relied heavily on handcrafted features and heuristics approaches. Techniques such as Histogram of Oriented Gradients (HOG) [3], Haar-like features [48], and deformable part-based models [10] were among the pioneering approaches. These methods focused on extracting specific features from images that could distinguish pedestrians from other objects. However, they often struggled to generalize across different scenarios due to variations in lighting conditions, occlusions, changes in viewpoint, and background clutter. The performance of these methods was often limited by their reliance on manually designed features, which could not capture the complex variations in pedestrian appearance and environmental conditions.

The advent of deep learning has revolutionized the field of computer vision, including pedestrian detection. Convolutional Neural Networks (CNNs) [30, 31] and other deep learning models have demonstrated remarkable success in learning complex features directly from data, leading to significant improvements in detection performance [46]. Deep learning-based methods such as Region-based CNN (R-CNN) [17], Single Shot MultiBox Detector (SSD) [37], and You Only Look Once (YOLO) [41] have set new benchmarks in object detection, including pedestrian detection. These models have shown the ability to generalize better across diverse conditions compared to traditional methods, owing to their capability to learn hierarchical feature representations.

Recent advancements in pedestrian detection include the integration of visible and thermal sensors, enabling multi-modal pedestrian detection. This approach leverages the complementary information provided by different modalities to enhance detection performance. Visible light cameras capture detailed visual information, while thermal cameras detect heat signatures, making it possible to detect pedestrians even in low-light or nighttime conditions. Multi-modal pedestrian detection systems are particularly robust in challenging conditions such as poor illumination, cluttered backgrounds, and adverse weather. As shown in Figure 1.2, the left image pair demonstrates low brightness in the visible modality, making pedestrian recognition challenging. However, with the supplementary information from the paired thermal modality, detecting pedestrians in the dark becomes feasible. By combining the strengths of both modalities, these systems can achieve higher accuracy and reliability.

Despite these advancements, multi-modal pedestrian detection faces several challenges, with one of the primary issues being the misalignment between visible and thermal modalities. Misalignment occurs due to differences in the viewpoints of the sensors, variations in resolution, and temporal discrepancies. The right image pair in Figure 1.2 illustrates the misalignment issue, where pedestrians appear in different positions in the visible and thermal images. This misalignment complicates the process of fusing data from both sensors, as the algorithm must match corresponding regions from images that do not perfectly overlap, making it difficult to combine information from both modalities accurately, which is crucial for reliable detection and localization.

Misalignment poses a significant problem because it causes discrepancies in the data captured by different sensors. These discrepancies can lead to

Figure 1.2: Examples of multi-modal pedestrian detection results. The left image pair demonstrates low brightness in the visible modality, making pedestrian recognition challenging, while the paired thermal modality facilitates detection. The right image pair illustrates misalignment, where the same pedestrians appear in different positions across modalities. Red boxes represent predicted bounding boxes, and lines between them indicate their paired relations.

detection errors, such as false positives or false negatives. When the visible and thermal images are not perfectly aligned, it becomes difficult to match the corresponding regions in both images accurately. This misalignment can result in a pedestrian detected in the visible image not aligning correctly with the thermal image, causing the system to either miss the detection or inaccurately locate the pedestrian. Furthermore, differences in sensor viewpoints can cause one sensor to capture parts of the scene that are not visible to the other sensor, adding to the complexity of aligning the images. Variations in resolution also contribute to the difficulty, as the details and quality of the images from each sensor do not match, making it challenging

to merge the information seamlessly.

The motivation for addressing this problem is driven by the critical need for accurate and reliable pedestrian detection in various applications, such as autonomous driving and surveillance. In autonomous driving, the safety of both pedestrians and passengers depends on the vehicle's ability to accurately detect and respond to the presence of pedestrians. In surveillance, effective pedestrian detection is essential for monitoring public spaces and ensuring security. Therefore, improving the alignment between sensor modalities can significantly enhance the performance of multi-modal pedestrian detection systems, leading to safer and more reliable applications.

To address these challenges, this dissertation aims to develop robust methods to mitigate the misalignment between visible and thermal modalities, thereby improving detection accuracy. While many existing methods, including ours, assume weak misalignment, our main contribution lies in the development of a network that can generate bounding box (BB) pairs from weakly misaligned data. This network maintains paired relations between the visible and thermal modalities, ensuring accurate localization of objects in both modalities despite the misalignment. The approach involves the introduction of novel techniques and algorithms designed to align the data from visible and thermal sensors more accurately. By enhancing the alignment, the proposed methods aim to reduce detection errors and improve the overall performance of pedestrian detection systems. Extensive experimental evaluations validate the effectiveness of our methods, demonstrating significant improvements in detection accuracy and robustness.

Ultimately, this work aspires to make significant contributions to the field of multi-modal pedestrian detection, paving the way for safer and more efficient systems in applications like autonomous driving and surveillance. By improving the robustness and accuracy of these systems, this research aims to enhance the safety and reliability of technologies that rely on accurate pedestrian detection.

In this dissertation, we use the term "multi-modal" to refer specifically to the combination of visible and thermal sensor modalities. While "multi-modal" can encompass a wide range of sensor types, including audio, touch, and temperature sensors, our focus is on the integration of visible and thermal images for pedestrian detection. This usage aligns with common practices in pedestrian detection, where "multi-modal" or "multispectral" [24, 25, 49, 29, 32, 61, 33, 59, 19, 64, 56, 57, 26] often refers to the fusion of visible and thermal data due to their complementary nature and significant applications

in areas such as autonomous driving and surveillance.

1.2 Objectives and Contributions

The primary objective of this dissertation is to develop and evaluate novel methods for multi-modal pedestrian detection, with a focus on mitigating the misalignment between visible and thermal modalities to improve detection accuracy. Our specific goals are:

1) Conducting a comprehensive review of existing pedestrian detection methods, covering both traditional techniques and deep learning-based approaches.

2) Identify and understand the specific challenges in multi-modal pedestrian detection, with a particular focus on the misalignment between visible and thermal modalities.

3) Develop novel multi-modal pedestrian detection methods that specifically address the issue of misalignment and improve detection performance in various conditions.

4) Introduce new evaluation metrics to better assess the performance of multi-modal detection methods, ensuring a more accurate measurement of their effectiveness.

5) evaluate the proposed methods through extensive experiments on benchmark datasets, comparing their performance against existing state-of-the-art (SOTA) approaches.

6) Analyze the experimental results thoroughly to understand the strengths and limitations of the proposed methods, and to identify potential areas for future research.

As we pursued these objectives, several significant contributions emerged from our work. Each step in our research journey led us to deeper insights and innovative solutions that collectively advance the field of multi-modal pedestrian detection. Our initial comprehensive review of pedestrian detection methods allowed us to clearly identify the pressing challenges, par-

ticularly the issue of misalignment between visible and thermal modalities. This foundational understanding was crucial as it guided the development of our novel detection methods. We designed these methods specifically to tackle the misalignment problem, incorporating sophisticated algorithms that enhance both robustness and accuracy. Recognizing the need for better evaluation tools, we introduced the multi-modal Intersection over Union (IoU^{M}) metric. This new metric provides a more precise assessment of detection performance across different modalities. Additionally, we developed the Multi-Modal Log-Average Miss Rate (MR^{M}) to evaluate how well detection methods perform under varying levels of misalignment, offering a comprehensive framework that more accurately reflects real-world conditions. We then moved forward to validate our methods rigorously through extensive experiments on benchmark datasets. These experiments included simulated disparity tests to systematically evaluate the impact of misalignment on detection accuracy. By comparing our methods with existing state-of-the-art (SOTA) approaches, we demonstrated significant improvements in detection performance, confirming the efficacy and robustness of our proposed solutions. In analyzing the experimental results, we gained valuable insights into the strengths and limitations of our methods. This thorough analysis not only validated our approaches but also highlighted potential areas for future research, ensuring continuous advancements in this critical field.

In summary, our work has led to several key contributions:

We have developed advanced multi-modal pedestrian detection techniques that effectively handle the challenges of misalignment between visible and thermal sensor modalities. These methods include a network capable of generating bounding box (BB) pairs from weakly misaligned data, maintaining paired relations between the visible and thermal modalities. This ensures accurate localization of objects in both modalities, even under significant misalignment conditions. The sophisticated algorithms integrated into our methods enhance the robustness and accuracy of detection systems, ensuring reliable performance.

2) We have introduced the multi-modal Intersection over Union (IoU^{M}) metric, a new evaluation metric that provides a more precise assessment of detection bounding boxes across different modalities. Additionally, we propose the Multi-Modal Log-Average Miss Rate (MR^{M}) to evaluate detection performance under varying levels of misalignment. This metric offers a compre-

hensive evaluation framework, highlighting the effectiveness of the proposed methods in handling misalignment issues.

3) We conducted extensive experiments on benchmark datasets to validate the effectiveness of the proposed multi-modal detection methods. These experiments included simulated disparity tests to systematically evaluate the impact of misalignment on detection accuracy. Our methods were compared against existing state-of-the-art (SOTA) approaches, demonstrating significant improvements in detection performance. A detailed analysis of the experimental results was performed to understand the strengths and limitations of our methods, providing valuable insights for future research and advancements.

By achieving these objectives, we aim to contribute to the advancement of pedestrian detection technology, fostering the development of safer and more reliable systems across various domains.

1.3 Dissertation Organization

This section outlines the structure of the dissertation, as illustrated in Figure 1.3.

Chapter 2: Existing Methods Review

This chapter comprehensively reviews existing methods in the field of pedestrian detection. It covers both traditional techniques, such as Histogram of Oriented Gradients (HOG) and Haar-like features, and modern deep learning-based approaches, including Region-Based Convolutional Neural Networks (R-CNN) and Single Shot MultiBox Detector (SSD). The chapter highlights the strengths and limitations of these methods, laying the groundwork for proposing novel solutions.

Chapter 3: Proposed Evaluation Metrics

In this chapter, new evaluation metrics for multi-modal pedestrian detection are introduced. The focus is on developing comprehensive and effective metrics to assess detection accuracy across different sensor modalities. Key metrics include multi-modal Intersection over Union (IoU^{M}) and Multi-Modal Log-Average Miss Rate (MR^{M}), providing a standardized framework

for evaluating the performance of multi-modal pedestrian detection systems.

Chapter 4: Proposed Multi-Modal Faster R-CNN Considering Misalignment

This chapter presents a novel approach to multi-modal pedestrian detection using the Faster R-CNN framework. The proposed method is specifically designed to address misalignment issues between different sensor modalities, enhancing detection accuracy in real-world scenarios. Detailed explanations of the methodology and experimental results are provided to demonstrate the effectiveness of the proposed approach.

Chapter 5: Proposed Multi-Modal Single Shot MultiBox Detector Considering Misalignment

Building upon the previous chapter, this section introduces another innovative method for multi-modal pedestrian detection using the Single Shot MultiBox Detector (SSD) framework. Similar to Chapter 4, the focus is on mitigating misalignment between sensor modalities to improve detection performance. The chapter includes methodological details and experimental findings to validate the efficacy of the proposed approach.

Chapter 6: Discussion and Conclusion

The final chapter reviews the main findings and contributions of the dissertation through a detailed discussion of experimental results and concludes with a summary of key insights. It provides a comprehensive overview of the proposed methods, their performance in addressing misalignment issues, and their implications for real-world applications. Additionally, it outlines potential directions for future work to further advance the field of multi-modal pedestrian detection.

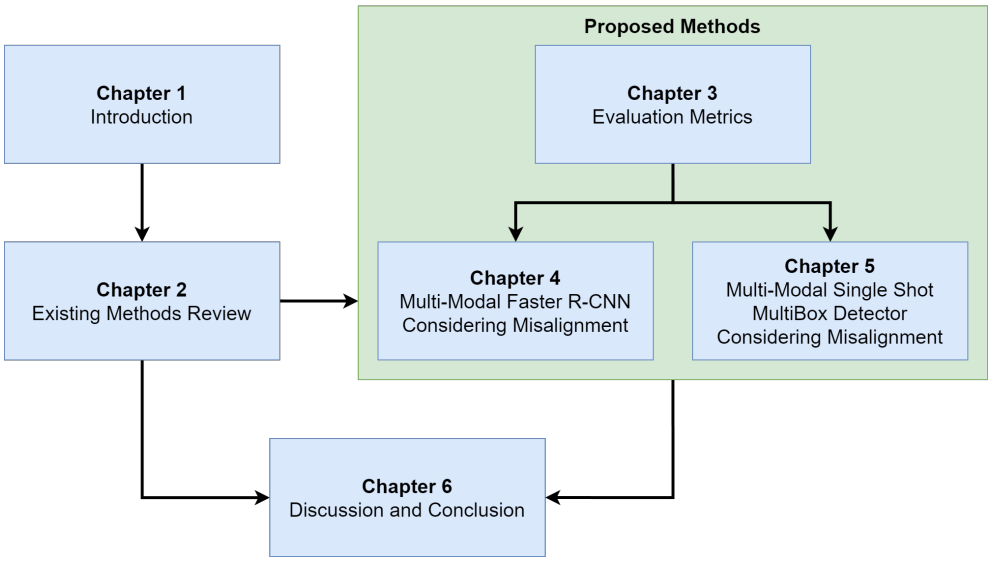


Figure 1.3: Overview of the dissertation structure highlighting the progression from the introduction and review of existing methods to the proposal and evaluation of novel approaches in multi-modal pedestrian detection. The figure illustrates the interconnection between chapters and the flow of research topics throughout the dissertation.

Chapter 2

Existing Methods Review

In this chapter, we present a comprehensive survey of pedestrian detection methods, exploring both traditional and deep learning-based techniques, as well as multi-modal approaches. We analyze their advantages, limitations, and effectiveness in various scenarios, providing a thorough understanding of the current state-of-the-art in pedestrian detection.

2.1 Pedestrian Detection

2.1.1 Introduction

Pedestrian detection is a fundamental task in computer vision with applications ranging from surveillance to autonomous driving. It involves identifying and locating pedestrians within images or video streams, which is crucial for ensuring safety and operational efficiency. Traditional methods, relying on handcrafted features and heuristic approaches, laid the groundwork for this field. Modern techniques leveraging machine learning have significantly improved accuracy and robustness. With the evolution of deep learning and the availability of large-scale datasets, pedestrian detection has seen significant advancements. This section provides an in-depth survey of pedestrian detection methods, exploring both traditional and deep learning-based approaches. We will discuss the strengths of these methods and their limitations in more complex scenarios, such as varying lighting conditions and occlusions.

2.1.2 Traditional Pedestrian Detection Methods

Traditional pedestrian detection methods have laid the groundwork for modern approaches by introducing various feature extraction and classification techniques. These methods primarily rely on handcrafted features and classical machine learning algorithms. Here, we highlight some of the most notable traditional methods.

Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) method [3], introduced by Dalal and Triggs, extracts gradient orientation information from localized image patches to represent local object appearance and shape. This method has demonstrated success in pedestrian detection tasks. However, it suffers from limitations such as sensitivity to variations in lighting conditions, occlusions, and changes in viewpoint. For instance, in crowded urban environments or under varying weather conditions, the effectiveness of HOG-based approaches diminishes.

Haar-like Features and Cascade Classifiers

Haar-like features [48], introduced by Viola and Jones, were among the early methods used for object detection, including pedestrian detection. These features capture intensity differences in rectangular regions of an image and are typically used in conjunction with cascade classifiers for efficient object detection. While they have shown effectiveness in face detection, they face challenges in pedestrian detection tasks due to similar limitations encountered by HOG-based methods. Variations in pedestrian poses, occlusions, and complex backgrounds reduce the robustness of Haar-like features in real-world scenarios.

Template Matching Methods

Template matching methods rely on predefined templates or shape models to detect pedestrians in images. These methods require accurate initialization and struggle with variations in pose, scale, and appearance [11]. For example, in situations where pedestrians exhibit diverse poses or are partially occluded, template matching methods may struggle to accurately localize pedestrians. The rigidity of these methods limits their adaptability to dynamic environments.

Deformable Part-based Models

Deformable part-based models [10], proposed by Felzenszwalb et al., address the limitations of template matching by incorporating deformable parts. These models consist of a set of parts, each with associated deformation costs, allowing flexibility in capturing variations in pedestrian appearance

and shape. However, despite their advancements, they still encounter challenges in handling complex real-world scenarios where pedestrians may exhibit diverse poses, occlusions, or variations in scale. The computational cost of deformable part-based models is also a significant concern, limiting their applicability in real-time systems.

2.1.3 Deep Learning-Based Pedestrian Detection Methods

Deep learning-based pedestrian detection methods have revolutionized the field by leveraging large datasets and complex neural network architectures to achieve superior performance compared to traditional methods. These methods automatically learn feature representations from data, eliminating the need for handcrafted features and improving detection accuracy and robustness. Below are some of the most prominent deep learning-based approaches.

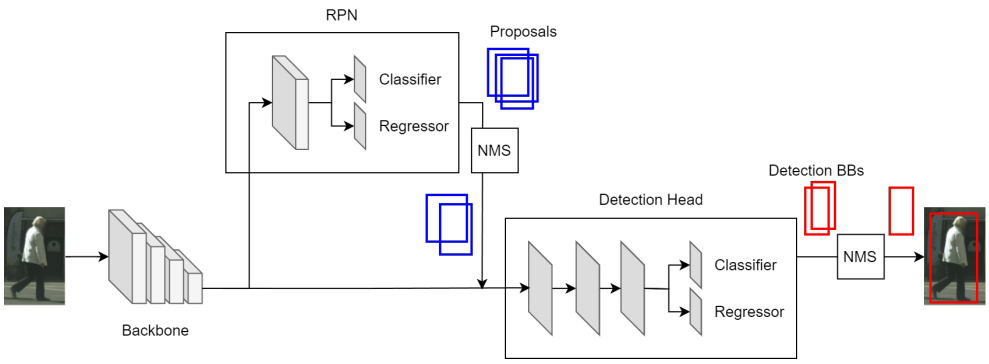


Figure 2.1: Overview of the Faster R-CNN framework. It begins with an input image, which is passed through a backbone network to extract feature maps. The Region Proposal Network (RPN) then generates proposals using anchors, predicting bounding box offsets and confidence scores. Non-Maximum Suppression (NMS) is applied to filter redundant proposals. These proposals are fed into the detection head, which includes a classifier for object classification and a regressor for further refining the bounding boxes. Another round of NMS is applied to finalize the detected objects, resulting in the final bounding boxes over the input image.

Region-Based Convolutional Neural Networks (R-CNN)

Region-Based Convolutional Neural Networks (R-CNN) [17], introduced by Girshick et al., represent a significant breakthrough in object detection by combining deep learning with region proposal techniques. R-CNN generates proposals using selective search and applies a convolutional neural network to classify and refine these proposals. The process involves three main steps: generating proposals, extracting features using a CNN, and classifying these regions. The selective search algorithm used for proposals is effective but computationally intensive, leading to slow inference speeds. While R-CNN significantly improved detection accuracy compared to traditional methods, its high computational cost limits its applicability in real-time pedestrian detection scenarios, such as in autonomous driving. Further advancements, such as Fast R-CNN [15] and Faster R-CNN [42], have been developed to address these limitations by streamlining the region proposal and classification processes.

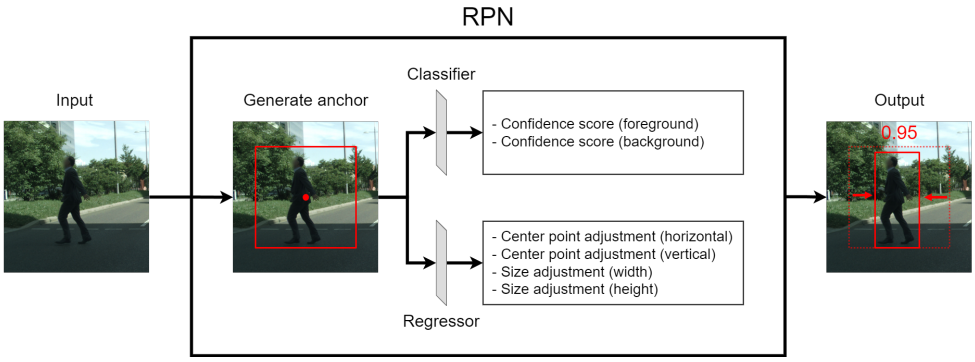


Figure 2.2: Region Proposal Network (RPN) within Faster R-CNN. The input image (represented for understanding, while the actual input is the feature map) is passed through a backbone network to extract feature maps. Although the RPN generates many anchors, only one is shown here for simplicity. The RPN uses these feature maps to generate proposals by predicting bounding box offsets and confidence scores for each anchor. The classifier in the RPN determines these confidence scores, while the regressor calculates the bounding box offsets. The bounding box offsets (shown with red arrows) adjust the predefined anchor to better fit the objects. The confidence score of foreground (displayed in red number) indicate the likelihood of the presence of an object.

Faster R-CNN is a state-of-the-art object detection framework known for its speed and accuracy. It consists of several key components that work together to achieve this task: the backbone network, the Region Proposal Network (RPN), and the detection heads. The architecture and workflow of Faster R-CNN is illustrated in Figure 2.1. Faster R-CNN begins with a backbone network, typically a pre-trained convolutional neural network such as VGG16 [45] or ResNet [22], which extracts feature maps from the input image. These feature maps contain rich information about the image's content and serve as the foundation for further processing. Next, the RPN slides over these feature maps to generate potential object regions, known as proposals. The RPN uses anchors, which are predefined bounding boxes of various sizes and aspect ratios placed at each location on the feature map. These anchors serve as reference points for predicting the locations of objects. For each anchor, the RPN predicts two things: the bounding box offsets, which are adjustments to the anchor to better fit the object, and a confidence score, which indicates the likelihood that an anchor contains an object as opposed to the background. These predictions are made using fully-connected layers within the RPN: a regressor for predicting the bounding box offsets and a classifier for predicting the confidence scores. The RPN generates proposals by applying the predicted bounding box offsets to the anchors. These proposals are the candidate regions that are likely to contain objects. The process of RPN, including its classifier and regressor, is demonstrated in Figure 2.2. Non-Maximum Suppression (NMS) is applied to filter out redundant proposals and keep the most promising ones. For each proposal, the detection heads, which also comprise a classifier and a regressor, refine the bounding box and classify the object within the region. The classifier and regressor in Faster R-CNN's detection heads work similarly to the RPN's classifier and regressor shown in Figure 2.2, but there are key differences. The detection head classifier in Faster R-CNN is designed to handle more classes than just foreground and background, providing a confidence score for each class, while the regressor further refines the bounding box coordinates to accurately enclose the object. After this step, Non-Maximum Suppression (NMS) is again applied to remove redundant bounding boxes and select the final set of detections.

In summary, Faster R-CNN is an efficient and accurate object detection framework that combines several key components: a backbone network for feature extraction, RPN for generating proposals using anchors, bounding box offsets for adjusting anchors, detection heads for classifying objects and

refining bounding boxes, and prediction scores for each class to indicate the likelihood of the detected objects. This integrated approach, with NMS applied both after RPN and detection heads, allows Faster R-CNN to quickly and accurately detect objects in images, making it a powerful tool in computer vision applications.

Faster R-CNN has several advantages, such as high accuracy, robustness, and flexibility. It achieves high accuracy in object detection due to its two-stage approach, which carefully refines proposals and classifications. The framework is robust and performs well on a variety of challenging datasets, making it a reliable choice for many applications. Additionally, Faster R-CNN can be adapted to different backbone networks and extended for various tasks beyond object detection, such as instance segmentation. However, Faster R-CNN also has some drawbacks. The two-stage process, while accurate, is computationally intensive and slower compared to single-stage detectors like SSD [37] and YOLO [41]. Due to its complexity, Faster R-CNN may not be suitable for real-time applications where high-speed processing is crucial. Moreover, the architecture is more complex to implement and requires careful tuning of hyperparameters and components to achieve optimal performance.

Regarding pedestrian detection, Faster R-CNN is particularly effective due to its high accuracy and ability to handle various challenges in object detection. Pedestrian detection requires precise localization and classification of humans in diverse environments, often with varying poses, occlusions, and lighting conditions. Faster R-CNN's two-stage approach ensures that proposals are refined before final detection, which helps in accurately identifying pedestrians amidst complex backgrounds. The backbone network extracts detailed feature maps that capture essential details of pedestrians, such as their shape and texture. The RPN generates proposals that likely contain pedestrians, even when they are partially occluded or present in varying sizes and aspect ratios. The detection heads further refine these proposals, classifying the detected regions as pedestrians and adjusting the bounding boxes to fit the actual shape of the pedestrians accurately. NMS plays a crucial role in pedestrian detection by eliminating redundant bounding boxes that may overlap significantly, ensuring that each pedestrian is represented by a single, precise bounding box. This helps in reducing false positives and improving the overall accuracy of the detection system. In pedestrian detection, the high accuracy and robustness of Faster R-CNN are significant advantages, making it a preferred choice for applications that require reliable and precise

detection, such as autonomous driving, surveillance, and crowd monitoring. However, the computational intensity of Faster R-CNN may pose challenges for real-time pedestrian detection, necessitating the use of optimized hardware or alternative methods for applications with strict latency requirements.

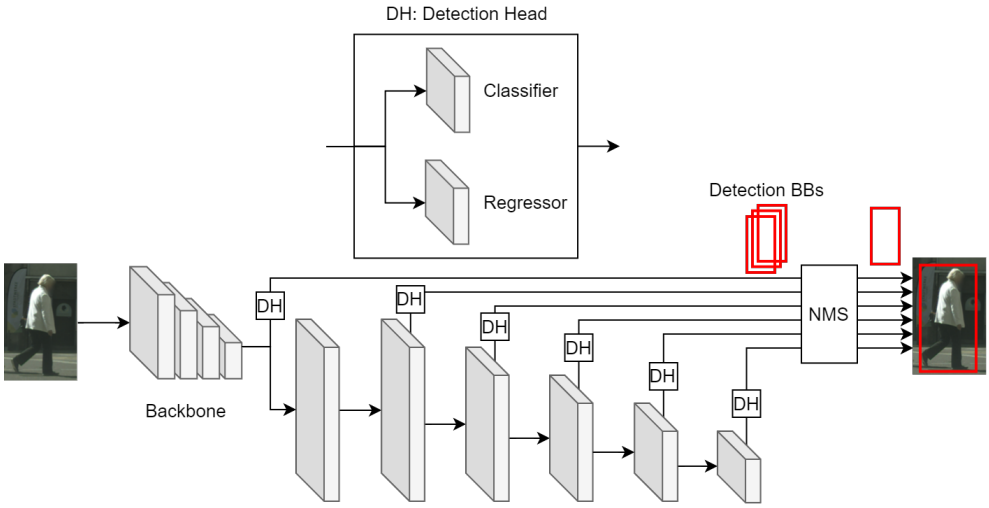


Figure 2.3: Overview of the Single Shot MultiBox Detector (SSD) [37] framework. It begins with an input image passed through a backbone network to extract feature maps at multiple scales. Detection heads (DH) are applied to these feature maps to predict bounding box offsets and class scores for each location. The detection heads utilize convolutional layers to generate these predictions, with separate classifier and regressor branches. Non-Maximum Suppression (NMS) is applied to filter redundant boxes and keep the most confident detections.

Single Shot MultiBox Detector (SSD)

Single Shot MultiBox Detector (SSD) [37], is a popular object detection framework designed for high-speed and accurate detection. Unlike Faster R-CNN, which uses a two-stage approach, SSD is a single-stage detector that directly predicts bounding boxes and class scores from feature maps at multiple scales. SSD achieves high detection accuracy and efficiency by predicting bounding boxes and class probabilities at multiple feature maps with different resolutions.

The architecture and workflow of SSD is illustrated in Figure 2.3. SSD begins with a backbone network, typically a pre-trained convolutional neural network like VGG16 [45] or ResNet [22], to extract feature maps from the input image. These feature maps provide rich hierarchical information about the image, capturing both low-level details and high-level semantic content. In SSD, detection heads are applied to multiple feature maps extracted from different layers of the backbone network. Each detection head, consisting of a classifier and a regressor, is responsible for predicting a fixed set of bounding boxes and their corresponding class scores for each location in the feature map. These predictions include bounding box offsets, which adjust the predefined anchor boxes (default boxes) to better fit the objects, and confidence scores, which indicate the likelihood of each class. The classifier and regressor in SSD's detection heads work similarly to the classifier and regressor in the detection head of Faster R-CNN, as shown in Figure 2.2, but there are key differences. In SSD, the classifier and regressor are applied at multiple scales and directly on the feature maps, allowing for simultaneous detection of objects of varying sizes, whereas in Faster R-CNN, these components refine the proposals generated by the RPN.+ SSD uses convolutional layers (typically 3x3) to generate these predictions. For each feature map cell, the convolutional layer outputs a set of bounding box coordinates and confidence scores for multiple classes. The advantage of using multiple feature maps is that it allows SSD to detect objects at different scales and aspect ratios, improving its ability to handle objects of varying sizes. The predefined anchor boxes in SSD come in various sizes and aspect ratios, covering different parts of the feature maps. These anchors act as reference points for the bounding box predictions, similar to the anchors in Faster R-CNN. After obtaining the predictions from all the detection heads, Non-Maximum Suppression (NMS) is applied to filter out redundant bounding boxes and retain the most confident detections. This process ensures that the final output consists of the most accurate and relevant object detections.

SSD is particularly effective for pedestrian detection due to its high speed and ability to handle objects at multiple scales. Pedestrian detection requires accurate localization and classification of humans in diverse environments, often with varying poses, occlusions, and lighting conditions. SSD's use of multiple feature maps allows it to detect pedestrians of different sizes and aspect ratios more effectively. The backbone network extracts detailed feature maps that capture essential details of pedestrians, such as their shape and texture. The detection heads apply convolutional layers to these feature

maps to predict bounding boxes and class scores for each location. By using predefined anchor boxes, SSD can handle the variations in pedestrian sizes and poses efficiently. NMS plays a crucial role in pedestrian detection by eliminating redundant bounding boxes that may overlap significantly, ensuring that each pedestrian is represented by a single, precise bounding box. This helps in reducing false positives and improving the overall accuracy of the detection system.

SSD offers several advantages, including high speed, simplicity, and multi-scale detection. It is faster than two-stage detectors like Faster R-CNN because it performs detection in a single stage, making it suitable for real-time applications, such as autonomous driving and video surveillance. The single-stage approach simplifies the detection pipeline, making SSD easier to implement and optimize, and end-to-end training is more straightforward compared to the two-stage approach. SSD's use of multiple feature maps allows it to detect objects at different scales, improving its ability to handle objects of varying sizes and aspect ratios. However, SSD also has some drawbacks. While it is fast, it may not achieve the same level of accuracy as two-stage detectors like Faster R-CNN, particularly for small objects. The single-stage approach can sometimes result in lower precision and higher false positives. Managing the predefined anchor boxes and their configurations can be complex, and ensuring that the anchors cover all possible object sizes and aspect ratios requires careful tuning. SSD's reliance on local feature maps may result in limited context for object detection, affecting its performance in scenarios with complex backgrounds or occlusions.

In summary, SSD is a single-stage object detector that performs detection in one go by using a backbone network to extract feature maps, applying detection heads to multiple feature maps to predict bounding boxes and class scores, and using predefined anchor boxes to guide the predictions. NMS is applied to remove redundant detections and keep the most confident ones. This approach allows SSD to achieve high-speed and accurate object detection, making it suitable for real-time applications. However, it may trade off some accuracy, especially for small objects, compared to two-stage detectors like Faster R-CNN.

You Only Look Once (YOLO)

You Only Look Once (YOLO) [41], introduced by Redmon et al., is another real-time object detection framework that divides the input image into a grid

and predicts bounding boxes and class probabilities directly from the grid cells. YOLO's approach to detection is based on a single neural network that performs both object localization and classification in one forward pass, making it extremely fast. The grid-based system allows YOLO to predict multiple bounding boxes and their associated class probabilities simultaneously. While YOLO offers real-time performance with high frame rates, it may sacrifice some accuracy compared to methods like SSD and R-CNN, particularly in scenarios with small objects or complex backgrounds. Despite these challenges, YOLO's speed and simplicity have made it a popular choice for various applications requiring rapid object detection.

2.1.4 Limitations of Pedestrian Detection Methods

Despite significant advancements in single-modal pedestrian detection methods, several limitations persist. These methods often struggle with sensitivity to variations in lighting conditions, occlusions, changes in viewpoint, and scale variations. The complexity of real-world scenarios presents additional challenges, including crowded environments, diverse pedestrian poses, and varying environmental conditions. Achieving robust and accurate pedestrian detection in such settings remains a daunting task. Furthermore, single-modal systems can suffer from calibration errors and dynamic changes in the scene, complicating detection tasks. Addressing these limitations is crucial for developing pedestrian detection systems that can reliably operate in diverse and challenging real-world scenarios. Future research should enhance the robustness of detection algorithms to effectively handle varying conditions and incorporate advanced calibration techniques to mitigate these issues. Additionally, exploring the integration of complementary techniques even within a single modality can provide more comprehensive detection capabilities, improving accuracy and reliability in dynamic environments.

2.2 Multi-Modal Pedestrian Detection

2.2.1 Introduction

Multi-modal pedestrian detection leverages multiple sensor modalities to enhance detection performance and robustness. By combining data from different sensors, such as RGB cameras, thermal cameras, depth sensors (Li-

DAR), and radar, multi-modal approaches can mitigate the limitations of single-modal systems. These methods aim to improve detection accuracy and reliability under various challenging conditions, such as low lighting, occlusions, and dynamic environments. In this section, we will focus on the fusion of color (RGB) and thermal sensors, highlighting the differences between naive and adaptive fusion methods.

2.2.2 Naive Feature Fusion

KAIST Multispectral Pedestrian Detection (KAIST) dataset [24] has been widely used in the research field of multi-modal pedestrian detection. Despite non-CNN-based approaches such as Aggregate Channel Features (ACF) [8] in the early days, the CNN-based approach is mainstream in this field currently [25, 49, 20, 29, 53, 40, 32, 19, 33, 59, 57]. The main challenge in the early days was how to combine and make use of information from both modalities, as with other computer vision applications [35, 43, 44]. MSDS-RCNN [32] proposes a framework that simultaneously performs pedestrian detection and segmentation in multispectral imagery, leading to improved detection performance compared to methods that simply combine features from both modalities. Most importantly, most of the existing methods strictly assume that visible-thermal image pairs are geometrically aligned. These methods merely fuse both modalities' features in corresponding pixel positions directly.

Although many geometric calibration and image alignment methods for multi-modal cameras have been proposed [39, 27, 9], accurate and dense alignment for each pixel is still an open problem. Naive feature fusion does not account for variations in sensor characteristics, such as different resolutions and fields of view, leading to incomplete or inconsistent feature representation. Furthermore, these methods are prone to errors in scenarios with dynamic backgrounds or when pedestrians are partially occluded in one of the modalities, reducing the overall robustness of the detection system. Examples of naive fusion include simple concatenation or averaging of pixel-wise features without any consideration for the spatial and temporal context of the detected objects. As a result, their detectors suffer dramatically worse performance in poorly aligned regions.

2.2.3 Adaptive Feature Fusion

Adaptive fusion incorporates advanced algorithms to dynamically adjust the fusion process based on the quality and relevance of the information from each modality. Machine learning models, such as convolutional neural networks (CNNs), are often employed to learn the optimal fusion strategy from large datasets, enhancing the system's ability to handle various conditions and environments. Additionally, adaptive fusion can implement weighting mechanisms that prioritize features from one modality over another based on context, such as giving more weight to thermal features in low-light conditions or relying more on RGB features in well-lit environments.

AR-CNN [61] significantly addresses the misalignment issue in multimodal CNN-based pedestrian detection. The authors analyzed the position shift problem, proposed the Aligned Region CNN (AR-CNN), and provided KAIST-Paired annotation. Their method predicts the shift distance between modalities for each Region of Interest (RoI), relocates the visible region into the thermal area, and then aligns them together. This approach successfully improved performance over previous methods that did not consider misalignment, highlighting the importance of addressing this issue for better detection accuracy. However, AR-CNN does not output bounding boxes (BBs) for each modality explicitly, making their BBs inaccurate when there is misalignment.

Similarly, MBNet [64] does not output BBs for each modality explicitly, resulting in inaccuracies when misalignment occurs. Furthermore, these methods assume "weak misalignment," typically defined as a shift of up to 10 pixels. For instance, AR-CNN's experiments involved shifting visible images by up to 10 pixels and evaluating detection results in the thermal modality, while MBNet did not conduct specific misalignment experiments but operates under the assumption of weak misalignment. Therefore, weak misalignment is generally defined as no greater than 10 pixels.

To overcome these limitations, future work should explore advanced alignment techniques and robust fusion strategies capable of handling more severe misalignment conditions. Integrating real-time processing capabilities and optimizing computational efficiency will be crucial for the practical deployment of these systems. By continuing to innovate and address these challenges, adaptive feature fusion can significantly advance the field of multimodal pedestrian detection, leading to more reliable and accurate systems capable of performing well in diverse and dynamic environments.

2.2.4 Challenges

Despite the progress made in multi-modal pedestrian detection, several challenges remain:

1) Misalignment between Modalities

Integrating data from different sensors can lead to misalignment issues, where pedestrians appear at different positions or scales across modalities. This misalignment arises due to variations in sensor placement, differing fields of view, and inconsistencies in sensor calibration. Such discrepancies negatively impact detection accuracy, as the alignment errors can cause incorrect feature mapping and poor fusion of information. Addressing misalignment requires sophisticated calibration techniques and advanced alignment algorithms that can dynamically adjust to ensure accurate data fusion.

2) Modality Imbalance

Variations in sensor characteristics, such as resolution, sensitivity, and range, can lead to modality imbalance. One modality may dominate over others, leading to an unequal contribution of information from different sensors. For instance, thermal cameras might be more effective in low-light conditions, while RGB cameras perform better in daylight. Additionally, different lighting conditions can affect the reliability of each modality differently, with thermal imaging being less affected by changes in ambient light compared to RGB cameras. To ensure fair representation and effective integration of information, adaptive weighting mechanisms and robust fusion strategies are necessary to balance the influence of each modality based on the context and environment.

3) Domain Adaptation

Multi-modal pedestrian detection models often struggle to generalize well across diverse real-world scenarios. Differences in environmental conditions, sensor configurations, and data distributions between training and deployment environments can degrade model performance. Domain adaptation techniques are essential to bridge this gap, allowing models trained on one dataset to perform effectively in various settings. This includes methods such as transfer learning, domain adversarial training, and data augmentation techniques that simulate different environmental conditions to enhance model robustness.

Researchers are actively working to address these challenges to advance the state-of-the-art in multi-modal pedestrian detection and develop more effective and robust pedestrian detection systems for real-world applications. Overcoming these obstacles will enhance the performance and reliability of detection systems, paving the way for their deployment in various real-world applications, from autonomous vehicles to surveillance systems. In conclusion, while significant steps have been made in the field of multi-modal pedestrian detection, the journey towards creating fully reliable and robust systems is ongoing. Continued research and innovation are essential to surmount these challenges, ultimately leading to safer and more efficient pedestrian detection solutions in our ever-evolving technological landscape.

Chapter 3

Proposed Evaluation Metrics

This chapter explains our proposed evaluation metrics that we use in our training and performance testing for our multi-modal detectors in detail. First, Multi-modal IoU (IoU^{M}) is introduced. Second, Multi-modal MR (MR^{M}) is introduced.

3.1 Multi-Modal IoU

Intersection over Union (IoU) is a fundamental evaluation metric in object detection tasks. It is widely used because it provides a clear and straightforward measure of how well the predicted bounding boxes match the ground truth bounding boxes. IoU helps in quantifying the accuracy of object localization, which is crucial for applications where precise detection is necessary. By evaluating the overlap between the predicted and actual bounding boxes, IoU allows researchers and practitioners to assess the performance of detection algorithms. The IoU is defined as:

$$IoU = \frac{GT \cap DT}{GT \cup DT}, \tag{3.1}$$

where GT and DT denote ground truth and detection bounding boxes, respectively. $GT \cap DT$ represents the area of intersection of ground truth and detection bounding boxes, while $GT \cup DT$ represents the area of union of ground truth and detection bounding boxes. A higher IoU indicates a better overlap between the predicted and ground truth bounding boxes, signifying more accurate detection. IoU is crucial for assessing the precision of object localization and is often used as a threshold to determine whether a detection is considered a true positive or false positive. In traditional pedestrian detection tasks, the IoU threshold is usually set at 0.5, meaning that a predicted bounding box must overlap with at least 50% of the ground truth bounding box to be considered a correct detection. This metric is widely used due to its simplicity and effectiveness in providing a clear measure of localization accuracy.

In the context of multi-modal object detection, specifically for pedestrian detection using visible and thermal modalities, it is essential to ensure accurate localization across both modalities. Traditional IoU does not account for potential misalignment between different modalities. When there is a misalignment between modalities, the coordinates of each object in both modalities are not the same. If we are only concerned about the precision of one modality, another modality will have poor precision, which can lead to inaccuracies in evaluating the performance of multi-modal detection systems. To address this limitation, we propose the Multi-Modal Intersection-over-Union ($\mathrm{IoU^M}$), which extends the traditional IoU metric to consider the overlap of bounding boxes in both visible and thermal modalities. This metric provides a more comprehensive evaluation of detection performance in multi-modal contexts by ensuring that detections are accurately localized in both modalities simultaneously. In order to measure the ability to handle both modalities, especially when the level of misalignment is high, we introduce a new evaluation metric, which we call "multi-modal IoU ($\mathrm{IoU^M}$)" defined as:

$$IoU^{\mathrm{M}} = \frac{(GT^V \cap DT^V) + (GT^T \cap DT^T)}{(GT^V \cup DT^V) + (GT^T \cup DT^T)}, \tag{3.2}$$

where GT^V and GT^T denote paired ground truth bounding boxes referring to the same object in visible and thermal modalities, respectively. Similarly, DT^V and DT^T denote paired detection bounding boxes referring to the same object in visible and thermal modality, respectively.

By incorporating the IoUM metric, we can effectively evaluate the precision of our multi-modal pedestrian detection system, ensuring that it performs accurately in both visible and thermal modalities, even in the presence of misalignment. The reason we use this formula instead of just average both modalities' IoU is mainly because of misalignment, which can make the position and size of objects vary between modalities, resulting in inequality between visible IoU ($\mathrm{IoU^V}$) and thermal IoU ($\mathrm{IoU^T}$). To take misalignment into account, we measure the ratio of both modalities' intersections to both modalities' unions. $\mathrm{IoU^M}$ can be used to determine the precision of detection bounding boxes in both modalities. Moreover, in order to thoroughly evaluate each modality, we define visible IoU ($\mathrm{IoU^V}$) as IoU in visible modality and thermal IoU ($\mathrm{IoU^T}$) as IoU in thermal modality. This comprehensive evaluation is crucial for applications such as autonomous driving and surveillance, where reliable detection of pedestrians across different environmental

conditions is imperative.

Algorithm 1: Greedy matching strategy to compute true positives, false positives, and false negatives

Input: the set of detection results B_d;
the set of detection scores S_d;
the set of ground truths B_g;
number of detection results N_d;
number of ground truth N_g;
matching threshold T;
Output: the set of true positives B_{tp};
the set of false positives B_{fp};
the set of false negatives B_{fn};
initialization;
sort the detection results B_d; in descending order according to their corresponding detection confidence scores S_d;
for $i < N_d$ **do**
    for $i < N_g$ **do**
        compute the IoU O_j between the bounding-box B_d^i and the ground-truth B_g^i;
    end
    compute maximum IoU $O_m = maxO_j$ and corresponding index $j_m = argmaxO_j$;
    if $O_m > T$ **then**
        add the corresponding B_d^{jm} to the set B_{tp};
        remove B_d^{jm} and B_g^{jm} from B_d and B_g;
    else
        add the corresponding B_d^{jm} to the set B_{fp};
        remove B_d^{jm} from B_d;
    end
end
Add B_g to B_fn;
return B_{tp}, B_{fp}, B_{fn};

3.2 Multi-Modal MR

Following the traditional evaluation of object detection, we categorize detection bounding boxes and ground truth bounding boxes into true positives, false positives, and false negatives in order to evaluate detection results. The traditional way to do that is greedy matching algorithm as shown in Algorithm 1. The traditional way to do that is the greedy matching algorithm. In short, the algorithm sort all the detection bounding boxes based on prediction scores from high to low. We then iterate through all unmatched ground truth bounding boxes for each detection bounding box and then match the pair with the highest IoU above IoU threshold (usually 0.5). We continue the iteration until we reach the maximum number of false positives. Matched detection bounding boxes will become true positives, unmatched detection bounding boxes will become false positives, and unmatched ground truth bounding boxes will become false negatives. In pedestrian detection, we value false negatives the most since miss detection could be crucial in real-life applications. The lower the false negatives, the better. One of the evaluation metrics we use is miss rate, defined as:

$$Miss\ Rate = \frac{Number\ of\ false\ negatives}{Number\ of\ all\ objects}. \tag{3.3}$$

In multi-modal pedestrian detection, performance is traditionally measured by log-average miss rate (MR) suggested by Dollar et al [7]. MR is defined by geometrical mean of miss rates at specific false positives per image (FPPI) evenly divided in log space, which can be formulated as:

$$Log-Average\ Miss\ Rate\ (MR)\ =\ \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}\ =\ exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln a_i\right], \tag{3.4}$$

where a_1, a_2, ..., a_n are miss rates at n different FPPI evenly spaced in log space. Miss rate is the proportion of false negative results to total objects, and FPPI is the proportion of false positive results to total images. Traditionally, we use 9 miss rates at evenly spaced FPPI over $[10^{-2}, 10^{0}]$ in log space (10^{-2}, $10^{-1.75}$, $10^{-1.5}$, ..., 10^{0}) to calculate MR, at which we call MR^{-2}. The lower the MR, the better.

The original KAIST dataset only had a single common annotation for each object in both modalities, despite misalignment between them. Therefore, MR was evaluated based on IoU between detection bounding boxes

and ground truth bounding boxes representing locations of objects for both modalities. Their annotation also has many errors, such as imprecise localization, misclassification, and misaligned regions[32]. Aware of the issue, many researchers relabeled KAIST annotation to solve the above errors. Liu et al.[25] provided improved annotation for the testing, which has become the standard annotation for performance evaluation. Li et al. [32] provided sanitized annotation for the training and demonstrated the effects caused by different kinds of annotation errors. Zhang et al. [61, 60] provided revolutionary KAIST-paired annotation, which carefully localizes pedestrians in both modalities and builds their relationships. They also evaluated the detection performance by MR^V and MR^T, which denote MR evaluating by visible annotation and thermal annotation, respectively. However, those evaluations were performed separately, and their detection results have no relationship between visible and thermal bounding boxes, which makes In order to evaluate the precision of detection results in both modalities pairwise, we change the criteria of the greedy matching algorithm from IoU to IoU^M, which represents MR based on IoU^M, "multi-modal MR (MR^M)". To use this metric, the detection results must be pairs of bounding boxes; each pair locates the same object in both modalities, which could have different coordinates due to misalignment. Not only is MR^M able to measure the precision of bounding boxes in both modalities simultaneously, but it also measures the ability to correctly match objects between modalities with misalignment since the detection bounding box pair can mismatch with other nearby objects, which can potentially become false negative, resulting in lower MR^M. We experiment using MR^M as an evaluation metric to demonstrate its effectiveness in measuring the detection performance against misalignment.

Chapter 4

Proposed Multi-Modal Faster R-CNN Considering Misalignment

In this chapter, we explain the principle of our proposed Multi-Modal Faster R-CNN Considering Misalignment in detail.

4.1 Background

Pedestrian detection is a crucial task in computer vision with significant applications in autonomous driving [55] and video surveillance systems [12]. Accurate pedestrian detection is essential for ensuring the safety and efficiency of these systems. For instance, in the context of autonomous vehicles, reliable pedestrian detection is paramount to prevent accidents and ensure the safety of both pedestrians and passengers. In surveillance, accurate pedestrian detection enhances security by enabling the monitoring of public spaces, identifying suspicious activities, and preventing potential threats.

The first era of pedestrian detection methods typically relies on handcrafted features extracted from visible images (e.g., RGB images). These traditional methods include techniques such as Histogram of Oriented Gradients (HOG) [3], Haar-like features [48], and deformable part-based models (DPM) [10]. HOG features capture edge and gradient information, while Haar-like features detect object shapes by computing differences in intensity between rectangular regions. DPMs model pedestrians using a collection of part detectors that account for variations in human pose and appearance. While these methods were groundbreaking at their inception, they struggle with generalizing across different scenarios due to variations in lighting conditions, occlusions, changes in viewpoint, and background clutter. The performance of these traditional methods is limited by their reliance on manually designed features, which cannot capture the complex variations in pedestrian appearance and environmental conditions.

The advent of deep learning marked the second era of pedestrian detection, significantly improving performance by leveraging convolutional neural

networks (CNNs) [30]. These deep learning-based methods initially also relied solely on visible images. They utilize CNNs to automatically learn feature representations from large datasets, resulting in superior performance compared to traditional handcrafted feature-based approaches. Notable examples include Region-based Convolutional Neural Networks (R-CNN) [17], Single Shot MultiBox Detector (SSD) [37], and You Only Look Once (YOLO) [41] frameworks. R-CNN extracts region proposals and classifies each proposal using a CNN. SSD directly predicts bounding boxes and class scores from feature maps at multiple scales, achieving high accuracy and speed. YOLO frames object detection as a single regression problem, predicting bounding boxes and class probabilities directly from full images in one evaluation. However, when using only visible images, these single-modal deep learning methods still face challenges under adverse conditions such as low lighting, occlusions, and cluttered backgrounds [54, 58, 38, 34, 63].

To address these limitations, various approaches have been proposed to combine multiple modalities (e.g., visible and far-infrared) [24, 18] and utilize the highly apparent regions of these modalities together. The integration of visible and thermal sensors enables multi-modal pedestrian detection, leveraging the complementary information provided by different modalities to enhance detection performance. Visible light cameras capture detailed visual information, while thermal cameras detect heat signatures, making it possible to detect pedestrians even in low-light or nighttime conditions. Multi-modal pedestrian detection systems are particularly robust in challenging conditions such as poor illumination, cluttered backgrounds, and adverse weather.

Despite these advancements, existing multi-modal methods often assume perfect alignment between the visible and thermal images, which is rarely the case in real-world scenarios. Misalignment can occur due to differences in sensor viewpoints, calibration errors, and temporal discrepancies [47, 1]. Figure 4.1 shows examples of misaligned annotations between visible and thermal images, highlighting the difficulties in achieving perfect alignment. This misalignment complicates the process of fusing data from both sensors, as the algorithm must match corresponding regions from images that do not perfectly overlap, making it difficult to combine information from both modalities accurately. Such discrepancies can lead to detection errors, such as false positives or false negatives.

In recent years, several methods have been proposed to integrate visible and thermal data for pedestrian detection. These methods typically employ a two-stream Faster R-CNN framework [25, 32, 40, 33], where features from

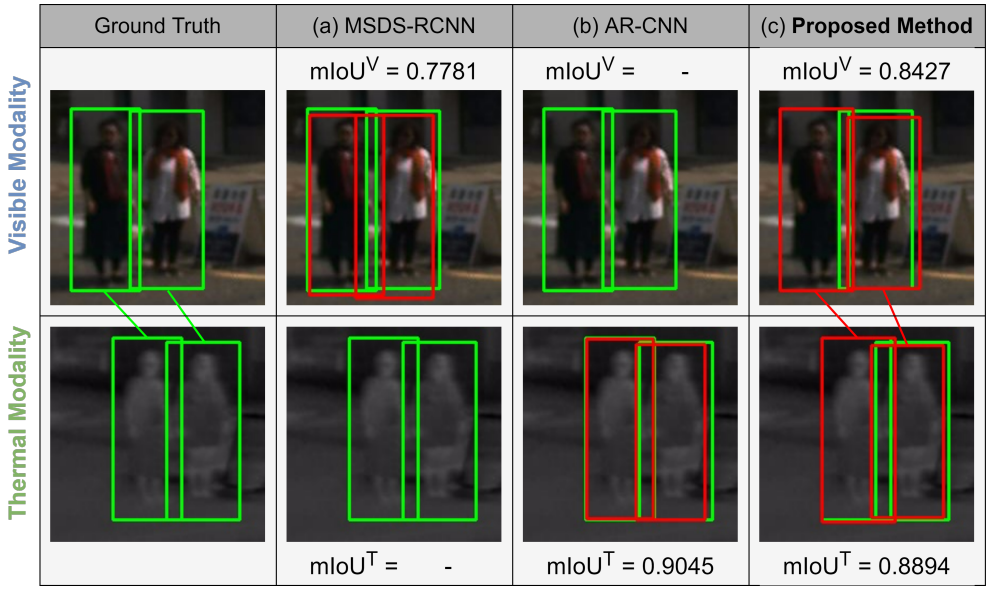


Figure 4.1: Visualization examples of ground truth annotations by Zhang et al. [61] (boxes in green), detection results (boxes in red), and overlap area between them, measured by mean visible IoU (mIoU^{V}) and mean thermal IoU (mIoU^{T}) of MSDS-RCNN [32], AR-CNN [61], and the proposed method. Image patches are cropped from visible-thermal image pairs in the same position from KAIST Multispectral Pedestrian Detection dataset [24] with large misalignment. (a) MSDS-RCNN [32], (b) AR-CNN [61], (c) proposed method.

both modalities are combined to improve detection accuracy. For example, MSDS-RCNN [32] combines detection and semantic segmentation tasks to optimize the model; however, without any consideration of misalignment, it is very sensitive to misalignment and can only precisely locate pedestrians in the visible modality, as shown in 4.1 (a). The fundamental assumption for this two-stream approach is that alignment between the two modalities is perfect.

To address the misalignment issue, some recent works have incorporated alignment modules within the Faster R-CNN framework. For example, the Aligned Region CNN (AR-CNN) [61] predicts shift distances between modalities for each region of interest (RoI) and relocates the visible region into the

thermal area. This approach successfully improves performance over previous methods that did not consider misalignment, highlighting the importance of addressing this issue for better detection accuracy. Another method, MBNet [64], focuses on mitigating modality imbalance and aligning features between the two modalities adaptively. However, these methods still output only one set of bounding box coordinates, neglecting the fact that objects can appear in different positions in the two modalities due to misalignment, as shown in Figure 4.1 (b). Despite predicting the shift distances of objects between modalities, AR-CNN can only precisely locate pedestrians in the thermal modality. Similarly, MBNet does not output bounding boxes for each modality explicitly, resulting in inaccuracies when misalignment occurs.

Existing methods fail to account for the fact that objects can appear in different positions in the two modalities, which can lead to significant errors in real-world applications where perfect alignment is nearly impossible. This misalignment issue results in reduced detection accuracy and increased false positives or false negatives, undermining the reliability of pedestrian detection systems. Given these limitations, it is evident that a more robust solution is required to handle misalignment effectively and improve detection accuracy. Therefore, we propose a multi-modal Faster R-CNN that is robust against misalignment. We use several novel strategies for the proposed multi-modal detection, including 1) multi-modal regressor, 2) multi-modal mini-batch sampling, and 3) multi-modal non-maximum suppression (NMS). The proposed method detects each object as a pair of bounding boxes with different coordinates in each modality, maintaining paired relations between the visible and thermal modalities. This allows for accurate localization of objects in both modalities, even in the presence of significant misalignment, as shown in Figure 4.1 (c). Note that despite the differences in the position of detection bounding boxes between modalities, all bounding boxes have paired relations between modalities; each pair indicates the same object in both modalities. Consequently, the proposed method can accurately pinpoint all objects in both modalities and match them regardless of displacement caused by misalignment.

Figure 4.2 shows the different Faster R-CNN-based approaches to multi-modal pedestrian detections. As shown in Figure 4.2 (a), a typical two-stream Faster R-CNN fuses features from both modalities directly without handling the disparity between each object. This approach assumes that the features from visible and thermal modalities are perfectly aligned, which is often not the case in real-world scenarios. As a result, the fused features may not

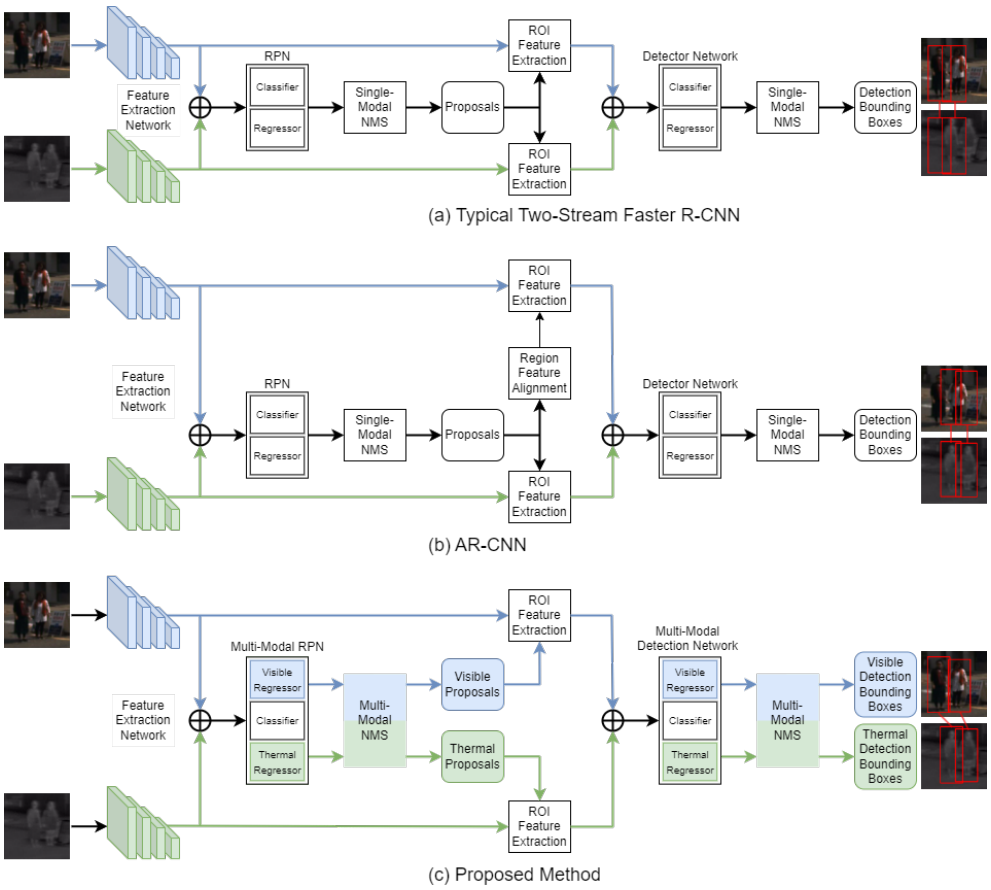


Figure 4.2: Comparison of multi-modal pedestrian detection frameworks based on Faster R-CNN. Blue and green blocks/paths represent properties of visible and thermal modalities, respectively. $\oplus$ denotes channel-wise concatenation. (a) Typical two-stream Faster R-CNN, (b) AR-CNN [61], (c) proposed method.

accurately represent the objects in both modalities, leading to suboptimal detection performance. These methods can only output detection bounding boxes for either the visible or thermal modality, but not both, which limits their effectiveness in multi-modal settings. This limitation is particularly problematic in scenarios with significant misalignment between the modalities, where the fused features fail to capture the true positions of objects.

Explicitly addressing the misalignment problem, AR-CNN integrates Re-

gion Feature Alignment to align each visible region with its counterpart thermal region before the detection network. This approach improves alignment by predicting shift distances between modalities for each region of interest (RoI) and relocating the visible region into the thermal area. As shown in Figure 4.2 (b), AR-CNN successfully aligns the features, which helps to improve detection accuracy compared to methods that do not consider misalignment. However, their method still has limitations. AR-CNN only outputs detection bounding boxes according to the position of objects in the thermal modality alone. This means that it does not fully utilize the information from both modalities, as it relies primarily on the thermal images for final detection. Consequently, any inaccuracies in the thermal modality can directly affect the overall detection performance.

Our proposed method, on the contrary, is designed to handle misalignment more effectively. Installed with a multi-modal regressor for both the RPN and detector and newly introduced multi-modal NMS, our method can output pairs of bounding boxes, which accurately locate objects in both modalities. As demonstrated in Figure 4.2 (c), the multi-modal regressor allows the detection network to predict bounding box coordinates separately for each modality, maintaining paired relations between the visible and thermal detections. This ensures that each object is accurately localized in both modalities, even in the presence of significant misalignment. The multi-modal NMS further refines the detections by considering the paired bounding boxes and suppressing redundant detections across both modalities. By addressing the misalignment issue and utilizing information from both visible and thermal images, our proposed method significantly enhances detection accuracy and robustness in challenging scenarios.

In summary, this chapter addresses the critical issue of misalignment in multi-modal pedestrian detection by proposing a robust Faster R-CNN framework. By introducing a multi-modal regressor, multi-modal mini-batch sampling, and multi-modal NMS, our approach ensures accurate localization and matching of objects in both visible and thermal modalities. This method significantly enhances detection performance in real-world scenarios, where misalignment between sensor modalities is a common challenge. Ultimately, this work contributes to the advancement of pedestrian detection technology, fostering the development of safer and more reliable systems across various domains, including autonomous driving and surveillance.

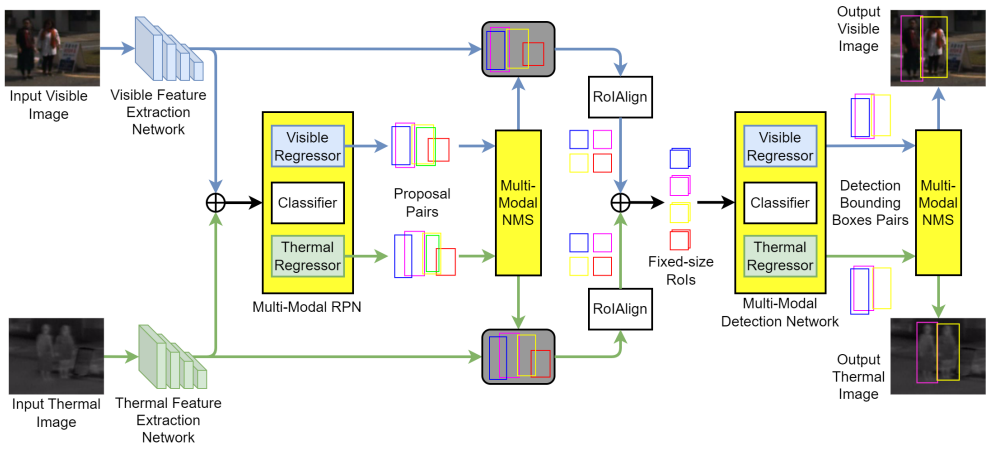


Figure 4.3: The overall architecture of our network. We extend Faster R-CNN into a two-stream network to take visible-thermal image pairs as input, then return pairs of detection bounding boxes as output for both modalities. Yellow blocks represent notable changes introduced in our method: RPN with visible and thermal regressors, detection heads with visible and thermal regressors, and detection outputs consisting of pairs of bounding boxes. Blue and green blocks/paths represent properties of visible and thermal modalities, respectively. RoIs and bounding boxes with the same color represent their paired relations. $\oplus$ denotes channel-wise concatenation.

4.2 Methodology

We adopt Faster R-CNN [42] architecture and extend it into two-stream network for multi-modal imaging, which consists of multi-modal RPN, multi-modal NMS, and multi-modal detector. Moreover, our multi-modal mini-batch sampling strategy are introduced. Overview of our network structure is shown in Figure 4.3.

4.2.1 Multi-Modal RPN

The proposed multi-modal RPN has a regressor for each modality, enabling proposals from each modality to adjust their sizes and positions independently. This is different from a single-modal regressor, which applies the same adjustments to both visible and thermal modalities, leading to inaccu-

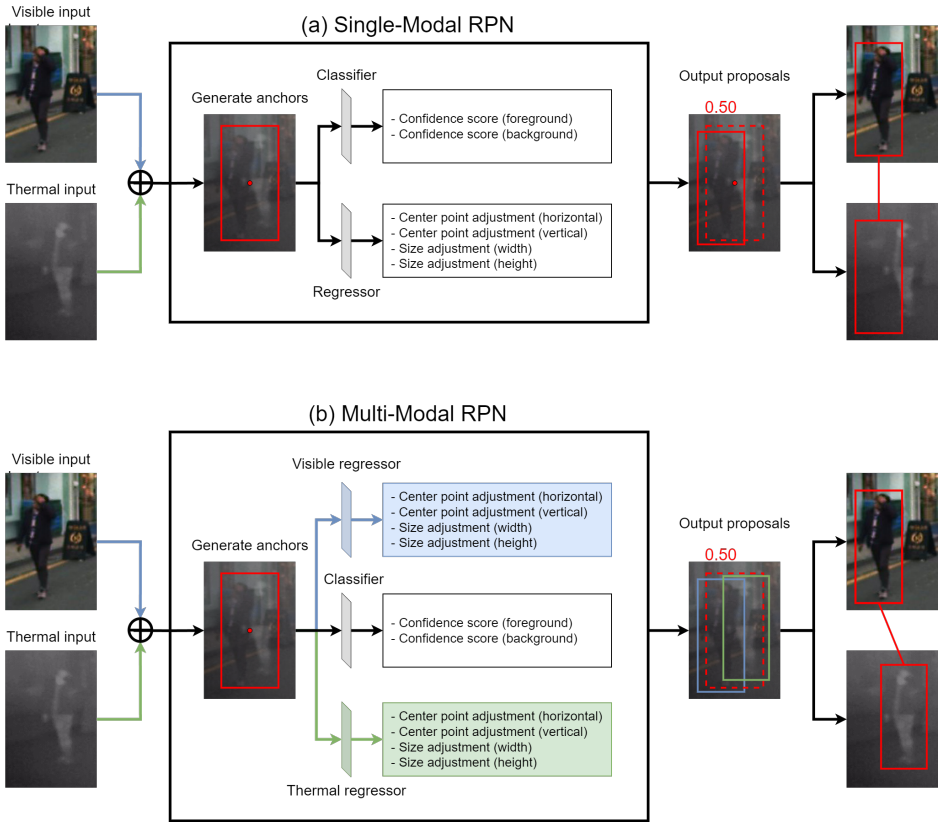


Figure 4.4: Comparison of single-modal and multi-modal regressor approaches. The input image pair is represented for understanding, while the actual input is the feature maps. Although the RPN generates many anchors, only one is shown here for simplicity. (a) Traditional single-modal regressor approach, where misalignment results in inaccurate bounding boxes. (b) The proposed multi-modal regressor, which performs bounding box adjustments in both modalities separately, leading to accurate bounding box alignment in both visible and thermal images.

racies when there is significant misalignment. The multi-modal regressor is designed to handle the differences between visible and thermal modalities, ensuring precise localization of pedestrians even when there is a significant misalignment between the modalities. It performs adjustments on both the horizontal and vertical center points as well as the width and height of the

bounding boxes. This approach ensures that both visible and thermal modalities contribute to the learning process, allowing our model to precisely locate pedestrians in both modalities. In contrast, the single-modal regressor returns the exact same set of proposals to both pipelines, which can result in inaccuracies when the modalities are not perfectly aligned. Our multi-modal regressor, however, returns proposal pairs with different positions for visible and thermal bounding boxes, providing more accurate regions of interest (RoI) for the detector in cases of significant misalignment. To further illustrate the differences and improvements, Figure 4.4 compares the single-modal and multi-modal regressor approaches. The top section shows the detection process without the multi-modal regressor, where misalignment issues result in inaccurate bounding boxes. The bottom section demonstrates the improvements achieved with the multi-modal regressor, where the bounding boxes are correctly adjusted to match the pedestrian's position in both modalities.

The proposed multi-modal RPN generates proposal pairs as its output, via classifier predicting each proposal pair a confidence score. To keep paired relations of proposals, we use multi-modal NMS (Section 4.2.3) to filter the best 300 out of many redundant proposal pairs. All remaining proposals will be applied with RoIAlign [21] operation to extract their feature maps into the exact size of 7×7 before returning to channel-wise concatenate with their corresponding pairs, resulting in well-aligned RoI for the detector. While single-modal regressor returns the exact same set of proposals to both pipelines, multi-modal regressor returns proposal pairs with different positions for visible and thermal bounding boxes, which gives more accurate RoI for detector in case of significant misalignment. We employ the loss function of RPN from Faster R-CNN [42] and add one more regression loss to optimize the precision of both modals, which is defined as:

$$L(\{p_i\}, \{t_i^V\}, \{t_i^T\}, \{p_i^*\}, \{t_i^{V*}\}, \{t_i^{T*}\}) \; = \\ \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \frac{\lambda}{N_{reg}} \sum_i p_i^* \left[L_{reg}^V(t_i^V, t_i^{V*}) + L_{reg}^T(t_i^T, t_i^{T*}) \right], \tag{4.1}$$

where i is the index of the anchor, p_i is the predicted probability of anchor i being an object. p_i^* is ground truth label of anchor i, which equals 1 if anchor i is positive (overlaps with an object above the high threshold) and 0 if anchor i is negative (overlaps with an object below the low threshold). L_{cls} is a cross-entropy over object and not object classes, which is defined as:

$$L_{cls}(p, p^*) \quad = \quad -\Big(p^*\ log(p) + (1-p^*)\ log(1-p)\Big) \tag{4.2}$$

t_i^V, t_i^T are vectors representing parameterized coordinates of predicted bounding box pairs as $t = (t_x, t_y, t_w, t_h)$ that associate with anchor i in visible and thermal modalities, respectively, and t_i^{V*}, t_i^{T*} are that of the ground truth bounding box pairs. Regression losses L_{reg}^V, L_{reg}^T are smooth L_1 loss for visible and thermal modalities, respectively, which are only activated for positive anchors ($p_i^* = 1$), defined as:

$$L_{reg}(t, t^*) \quad = \quad \sum_{j \in \{x,y,w,h\}} smooth_{L1}(t_j - t_j^*), \tag{4.3}$$

in which

$$smooth_{L1}(x) \quad = \quad \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{4.4}$$

N_{cls} is mini-batch size and N_{reg} is number of anchor locations. Following Girshick et al. [16], the parameterized coordinates $t = (t_x, t_y, t_w, t_h)$ for regression are scale-invariant translation and log-space shift relative, defined as:

$$\begin{aligned} t_x &= (x_p - x_a)/w_a, \quad & t_x^* &= (x^* - x_a)/w_a, \\ t_y &= (y_p - y_a)/h_a, \quad & t_y^* &= (y^* - y_a)/h_a, \\ t_w &= log(w_p/w_a), \quad & t_w^* &= log(w^*/w_a), \\ t_h &= log(h_p/h_a), \quad & t_h^* &= log(h^*/h_a), \end{aligned} \tag{4.5}$$

where x, y denote the bounding box's center coordinates and w, h denote its width and height. Variables p, a, and * denote coordinates of prediction, anchor, and ground truth bounding boxes, respectively. We set $\lambda = 1$ for all experiments.

In multi-modal pedestrian detection, specifically when using visible and thermal images, misalignment between the two modalities poses significant challenges. One might consider using separate anchors for different modalities, which involves matching all possible combinations of anchors between the modalities. However, this approach has several drawbacks that make it less effective.

Firstly, using separate anchors for different modalities and matching all combinations significantly increases computational complexity. Each modality would require its own set of anchors. If we consider N anchors for the visible modality and M anchors for the thermal modality, this approach would require evaluating $N \times M$ combinations. For instance, if $N = 1000$ and $M = 1000$, the system would need to evaluate 1,000,000 anchor combinations. This combinatorial explosion can lead to a dramatic increase in the number of anchors to be processed, resulting in slower detection speeds and higher computational costs. Such an approach is impractical for real-time applications like autonomous driving or surveillance, where rapid and efficient processing is crucial.

Secondly, matching all combinations of separate anchors does not inherently solve the misalignment problem. Misalignment occurs because the positions and sizes of objects in the visible and thermal images do not match perfectly. Even with separate anchors, the system still requires a robust mechanism to align the detected objects accurately across both modalities. Without addressing this fundamental issue, separate anchors with all combinations would provide only a partial and inefficient solution, potentially leading to increased false positives and misdetections.

Thirdly, using separate anchors for different modalities complicates the training process. The system must learn to generate and regress anchors for each modality independently while also learning to match the combinations of anchors between the modalities. This complexity can make the training process more challenging and less stable, potentially leading to suboptimal performance. It also requires extensive annotated data for both modalities to train effectively, which may not always be available.

Instead, our approach involves using a multi-modal regressor that can handle the bounding box adjustments independently for each modality while maintaining their paired relations. This method ensures accurate localization of objects in both modalities and effectively addresses the misalignment issue without the drawbacks associated with separate anchors and matching all combinations. By leveraging a unified framework, we achieve better computational efficiency, simpler training, and improved detection accuracy in multi-modal pedestrian detection.

The multi-modal regressor offers several advantages, including improved accuracy in bounding box predictions due to independent adjustments for visible and thermal modalities, which leads to better handling of misalignment issues. This approach also ensures that both modalities contribute

effectively to the learning process, enhancing the overall robustness of the detection system. However, there are also some drawbacks to using a multi-modal regressor. In scenarios where locating pedestrians in both modalities is not necessary, the additional computational cost may be considered a waste. Knowing the position of pedestrians in just one modality might be sufficient for certain applications. Additionally, in crowded scenes, the model might incorrectly match nearby individuals as the same person across modalities, leading to mismatches and potential inaccuracies.

4.2.2 Multi-Modal Detector

The proposed multi-modal detector is designed to leverage the complementary information from both visible and thermal modalities to enhance pedestrian detection accuracy. This detector includes a regressor for each modality, allowing it to independently adjust the positions and sizes of bounding boxes in the visible and thermal images. This independent adjustment is crucial for handling the misalignment between the modalities, ensuring that the detection is precise in both. However, both modalities share a single classifier, which predicts a confidence score for each pair of bounding boxes, determining the likelihood of them containing a pedestrian.

Similar to the multi-modal RPN, the multi-modal detector employs multi-modal Non-Maximum Suppression (NMS) to eliminate vague and overlapping bounding box pairs. This step is vital in refining the detection results by filtering out redundant boxes and keeping the most accurate ones. Multi-modal NMS considers the information from both modalities, which helps in reducing false positives and enhancing the overall detection performance. In the end, the detection results consist of pairs of bounding boxes for both modalities. These bounding boxes have different sizes and positions in the visible and thermal images, accurately representing the pedestrians in each modality while maintaining their paired relations. This approach ensures that the final bounding boxes are precise for both modalities, addressing the challenges of misalignment and varying appearances in different imaging conditions.

The multi-modal detector operates similarly to the multi-modal RPN in terms of its structure and function, as depicted in Figure 4.4 (B). Both systems use separate regressors for each modality and employ multi-modal NMS to refine the outputs. However, there are key differences between the two. The RPN is primarily focused on generating initial region proposals

that highlight potential pedestrian locations. These proposals serve as rough guesses that need further refinement. The multi-modal detector takes these initial proposals and enhances them by adjusting the bounding box coordinates more accurately and assigning confidence scores through a shared classifier. This additional processing step ensures that the final detections are precise and reliable.

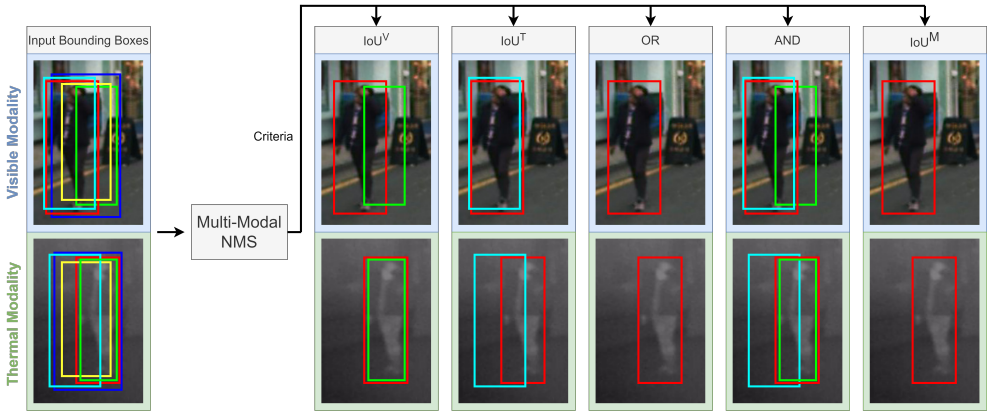
To optimize the performance of the multi-modal detector, we adopt the loss function from Fast R-CNN [39] and introduce an additional regression loss This combined loss function is defined as:

$$L(\{p_i\}, \{t_i^V\}, \{t_i^T\}, \{p_i^*\}, \{t_i^{V*}\}, \{t_i^{T*}\}) = \sum_i L_{cls}(p_i, p_i^*) + \lambda \sum_i p_i^* \left[L_{reg}^V(t_i^V, t_i^{V*}) + L_{reg}^T(t_i^T, t_i^{T*}) \right], \tag{4.6}$$

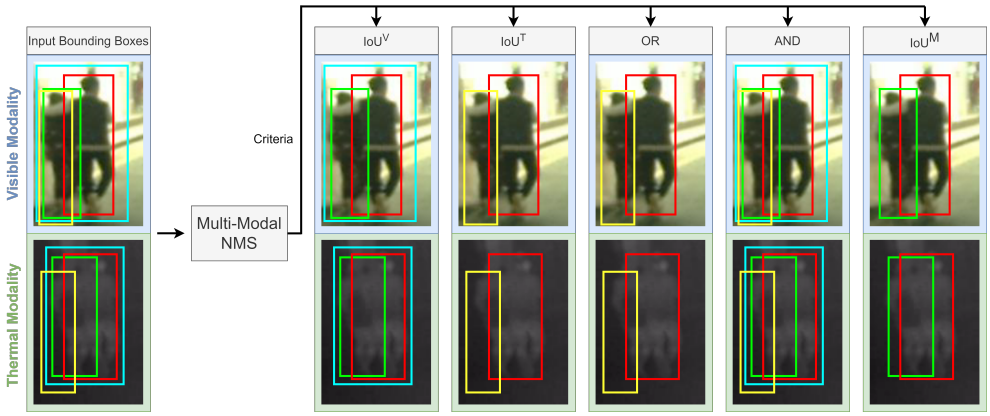
where i is the index of a bounding box pair, L_{cls} is a cross-entropy for class probability p_i and true class p_i^* (Eq. 4.2) of bounding box pair i, since we only consider pedestrian class, there are only two classes, pedestrian and non-pedestrian, in which p_i^* equals 1 and 0, respectively. Regression losses L_{reg}^V, L_{reg}^T are smooth L_1 loss (Eq. 4.3 and 4.4) over predicted regression offsets t_i^V, t_i^T and regression targets t_i^{V*}, t_i^{T*} of bounding box pair i for visible and thermal modalities, respectively, which are also parameterized as Eq. 4.5. We set $\lambda = 1$ for all experiments.

4.2.3 Multi-modal NMS

Non-maximum Suppression (NMS) is a technique for selecting one entity out of many overlapping entities, usually using IoU as a suppression criterion, i.e., when the IoU between bounding boxes exceeds the threshold, the bounding box with a lower prediction score is suppressed. Since CNN-based methods generate many dense bounding boxes mostly detecting the same objects, the detection results are cluttered with unnecessary bounding boxes. Therefore, we use NMS to remove those lower-quality bounding boxes, keeping only the best bounding boxes to locate the objects. However, since we need to keep paired relations of bounding boxes between visible and thermal modalities in this procedure, we must select and suppress bounding boxes in a pairwise approach, or paired relations would be lost in the suppression process. For this purpose, we attempt various criteria that can select and suppress bounding boxes in a pairwise manner for our NMS.



(a) Single pedestrian



(b) Two pedestrians with partial occlusion

Figure 4.5: Examples of multi-modal NMS processes with different criteria on the scenes where large misalignment is present. Bounding boxes with the same color reflect paired relations between them. Left side shows bounding box pairs prior to multi-modal NMS. Right side shows results of multi-modal NMS with different criteria. (a) Single pedestrian, (b) two pedestrians with partial occlusion.

As a naive extension, we can use either IoU^{V} or IoU^{T} as multi-modal NMS criteria, i.e., making one modality a dictator. In AR-CNN [61], IoU^{T} is used for NMS criteria, neglecting information of visible modality. Accordingly, we can use the logical operation of IoU^{V} and IoU^{T} as criteria. For OR operation, if either IoU^{V} or IoU^{T} exceeds the threshold, the bounding box

pair with a lower score will be suppressed. For AND operation, if both IoU^V and IoU^T exceed the threshold, the bounding box pair with a lower score will be suppressed. Lastly, we use the proposed IoU^M as criteria, where the proportion of intersection and union from both modalities is considered. Examples of proposed multi-modal NMS results with different criteria are demonstrated in Figure 4.5. Several detection bounding box pairs around objects are shown before and after the multi-modal NMS process.

From Figure 4.5 (a), IoU^V and IoU^T clearly show weakness, considering only one modality, these criteria can not get rid of all ambiguous bounding boxes around the same object, which is also the same for logical operator AND. On the contrary, logical operator OR and IoU^M can suppress needless bounding boxes correctly. In case of multiple pedestrians with partial occlusion and misalignment, which is not uncommon in the real situation, correct bounding boxes could be removed if not handled properly. As shown in Figure 4.5 (b), IoU^V and logical operator AND, while preserving correct bounding boxes, fail to remove poor bounding boxes around the objects. IoU^t and logical operator OR remove most poor bounding boxes, including correct green bounding boxes. As a result, less precise yellow bounding boxes remain. Meanwhile, IoU^M can suppress and keep all correct bounding boxes precisely. IoU^M is our best candidate for NMS criteria since it considers the overlaps between bounding boxes in both modalities, unlike other criteria. The experiment to demonstrate performance comparison between all NMS criteria is also conducted (Section 4.3.5).

4.2.4 Multi-Modal Mini Batch Sampling

We follow sampling strategies from [42] and [15]. But since our approach has one regressor for each modality exclusively and we need to keep paired relations for all proposals and RoIs, we select training samples as anchor pairs and RoI pairs. For this purpose, we use IoU^M as selection criteria instead of IoU. For RPN, we assign positive and negative labels to anchor pairs that have IoU^M overlap higher than 0.63 with any ground truth bounding box pair and lower than 0.3 for all ground truth bounding box pairs respectively, for a total of 256 anchor pairs, whereas positive labels can take up to 128 anchor pairs. For detector, we assign positive and negative labels to RoI pairs that have IoU^M overlap with any ground truth bounding box pair higher than 0.5 and lower than 0.5 but higher than 0.1 respectively, for a total of 128 RoI pairs, whereas positive labels can take up to 32 RoI pairs.

4.3 Experiment

First, we describe the dataset we used in our experiments. Second, the details of our implementation are clarified. Third, we indicate evaluation details of our experiments, which include the explanation of simulated disparity experiment. Fourth, we illustrate and discuss the results of our experiments compared with existing methods, divided into performance comparison and qualitative comparison. Finally, we conducted ablation experiments to verify the effectiveness of our network's components, multi-modal regressor, and multi-modal NMS.

4.3.1 Dataset

KAIST dataset [24] was used in our experiments. It is one of the widely used multi-modal pedestrian datasets, with more than 90,000 frames recorded both day and night to consider changes in light conditions. It was initially assumed to be geometrically aligned. However, the annotations have many errors [32], such as imprecise localization, misclassification, and misaligned regions. Many researchers constructed their improved version of annotations to use instead of the original. Improved annotations provided by Liu et al.[25] has officially been used as standard annotations for performance benchmark. Zhang et al. [61] carefully analyzed the misalignment problem of KAIST dataset and were the first to provide paired annotations for KAIST dataset, locating objects for each modality individually and building their paired relations. Since we focus on the misalignment problem, we adopted their annotations to use in our work for training and testing.

4.3.2 Implementation Details

We adopt VGG-16 [45] pre-trained on ImageNet [4, 28] as our two-stream backbone networks as in AR-CNN [61]. We train the network for three epochs with a learning rate of 0.005 and one additional epoch with a learning rate of 0.0005 by Stochastic Gradient Descent (SGD) optimizer with 0.9 momentum and 0.0005 weight decay. We select 8,892 images from the training set containing informative pedestrians for the training. Image resolution is fixed to 640×512. All images are horizontally flipped and append to original training data for data augmentation. For multi-modal RPN's mini-batch sampling, we set IoU^{M} of high and low thresholds at 0.63 and 0.3, respectively. For

multi-modal detector's mini-batch sampling, we set IoU^{M} of high threshold and low thresholds at 0.5 and 0.1, respectively. For the first NMS following RPN, we set IoU^{M} threshold at 0.7 in the training to generate proposals with more variation in precision, which can benefit the training of the detector, and we set IoU^{M} threshold at 0.55 in the testing to generate bounding boxes with higher precision. For the second NMS following detector, we set IoU^{M} threshold at 0.53.

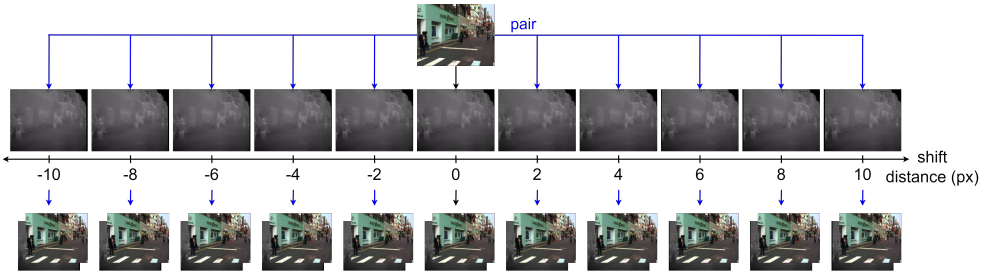


Figure 4.6: An example of simulated disparity experiment. The top image is the original visible image, while the middle and bottom rows illustrate the thermal images shifted in both directions and their corresponding results after pairing with the visible images.

4.3.3 Evaluation Details

To thoroughly evaluate the effectiveness of our method against misalignment, we introduce "simulated disparity experiment" between modalities as an experiment set up by shifting thermal images by 2, 4, 6, 8, and 10 pixels horizontally in both directions to imitate the misalignment, which mainly occurs in the horizontal direction. There is no change to visible images. However, In case of any pedestrian goes over the image border as a result of shifting, that pedestrian will be ignored from the evaluation. Subsequently, we will have 11 subsets of different misalignments as test data for each horizontal shift. Figure 4.6 illustrates the simulated disparity experiment setup. This simulated disparity setting allows us to systematically evaluate the impact of varying degrees of misalignment on our multi-modal pedestrian detection method.

Moreover, mean and standard deviation (SD) of MR^M over all subsets are calculated. For performance comparison, detection performance was mea-

sured by MR^M over the range of [10^{-2}, 10^0] FPPI with IoU^M threshold of 0.5 (MR^M_{50}) and 0.75 (MR^M_{75}), respectively, for all simulated disparity distances. Additionally, MR curves are plotted by using the mean and the worst miss rate of all simulated disparity distances over the range of [10^{-2}, 10^0] FPPI with IoU^M threshold of 0.5 (MR^M_{50}) and 0.75 (MR^M_{75}), respectively. For qualitative comparison, detection performance was measured by mean multi-modal Intersection over Union ($mIoU^M$) between all ground truth bounding boxes and detection bounding boxes with the highest $mIoU^M$ overlap in each scene. For ablation study, detection performance was measured by MR^M over the range of [10^{-2}, 10^0] FPPI with IoU^M threshold of 0.5 (MR^M_{50}) for all simulated disparity distances. All experiments were performed under reasonable con-Figureuration [24], i.e., only pedestrians taller than 55 pixels under partial or no occlusion are considered. Only 2,252 frames sampled from the test set with 20-frame skips were used in the performance test as traditional.

4.3.4 Comparison with Existing Methods

We selected three existing methods for our experiments, MSDS-RCNN [32] is representative of methods without misalignment consideration, AR-CNN [61] and MBNet [64] are methods that consider misalignment, trained by KAIST-paired annotations provided by Zhang et al. [61]. For methods that do not generate paired detection bounding boxes, we substituted their detection bounding boxes with those from one modality. Our method is currently the only one that generates paired bounding box outputs for both visible and thermal modalities, ensuring more accurate and reliable detection despite misalignment.

Performance Comparison

As shown in Table 4.1, MSDS-RCNN, the only method not considering misalignment, performs much poorer than other methods, especially when the simulated disparity is significant. For the proposed method, performance is mediocre when there is no simulated disparity, and at a shift distance of -2, performing worse than MBNet [64] by about 8%. Note that AR-CNN [61] and MBNet [64] show better performance at shift distance of -2 than without simulated disparity. This demonstrates that the dataset has some misalignment from the beginning, and simulated disparity could align objects in certain circumstances. Our method, however, shows a noticeable performance

improvement, achieving the best MR_{50}^{M} when disparities are larger than 4 pixels, demonstrating the effectiveness of our method against misalignment. Our method also achieves the best performance at all disparities when IoU^{M} threshold is 0.75, as shown in Table 4.2. This indicates that our method performs the best when the requirement for bounding boxes' precision is strict, showcasing the superior precision of the proposed method's detection bounding boxes. Additionally, we have the lowest mean and standard deviation (SD) for both MR_{50}^{M} and MR_{75}^{M}, demonstrating our best overall performance across all misalignment situations and our robustness against misalignment.

The MR plots by mean miss rate (solid lines) and worst miss rate (dashed lines) over all simulated disparity distances are shown in Figure 4.7. The MR plot of our proposed method's mean miss rate is at the bottom, with the MR plot of the worst case being comparable to MR plots of other state-of-the-art (SOTA) methods in Figure 4.7 (a). Our method achieved the best MR_{50}^{M}, calculated by the geometric mean of miss rate over $[10^{-2}, 10^{0}]$ FPPI, for both mean miss rate and worst miss rate, outperforming AR-CNN [61] by 13% and 27%, respectively. From Figure 4.7 (b), our proposed method clearly outshines other SOTA methods, with plots at the bottom for both mean and worst-case scenarios, demonstrating superior robustness to misalignment and bounding box precision in both modalities. Similarly, our method achieved the best MR_{75}^{M} for both mean miss rate and worst miss rate, surpassing MBNet [64] by 25% and 32%, respectively.

All other methods generate a single bounding box for each object, representing the same position in both modalities. This approach can lead to inaccuracies when there is misalignment because the object may appear in different locations in each modality. Relying on a single bounding box position for both modalities reduces precision under misalignment conditions. In contrast, our method uniquely generates paired detection bounding boxes for both visible and thermal modalities. This capability ensures more accurate and reliable detection, even when there is misalignment, distinguishing our approach from others. As our method is the only one that generates paired bounding boxes, direct comparisons with other methods might seem somewhat unfair. However, these comparisons highlight the superior effectiveness of our approach in handling misalignment. The evaluation metric, IoU^{M}, favors methods that accurately align bounding boxes across both modalities, further demonstrating the advantage and performance of our method. By introducing this innovative way to evaluate bounding box accuracy for multimodal detection with misalignment, we hope to inspire other researchers to

adopt and refine this approach, enhancing the precision and reliability of detection systems.

Table 4.1: Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity in the horizontal direction, by MR^M_{50}, including their mean and standard deviation over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Note that other methods generate a single bounding box for each object, representing the same position in both modalities, while our method generates paired bounding boxes. Bold values indicate the best performance.

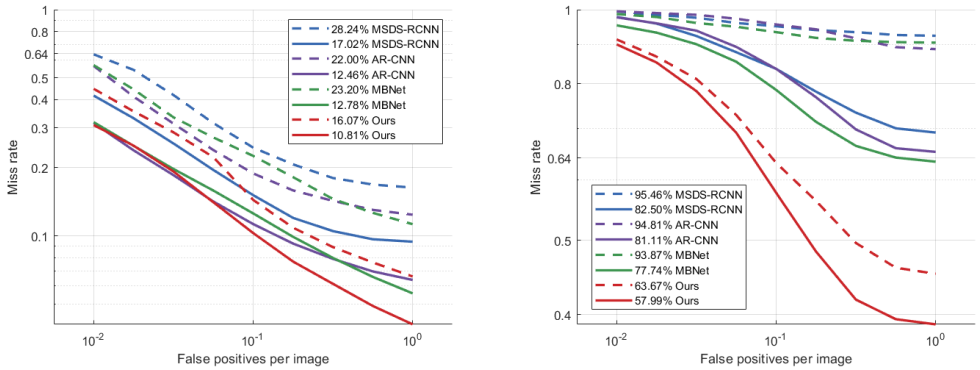
Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN[32]	27.06	18.76	15.93	12.74	12.58	11.09	11.72	13.25	15.06	21.38	27.48	17.00	5.94
AR-CNN[61]	21.61	14.65	10.43	8.67	8.22	8.79	8.68	10.10	11.02	14.65	19.84	12.42	4.69
MBNet[64]	23.14	15.31	11.02	8.92	**7.70**	**7.76**	8.64	9.88	11.17	14.87	21.70	12.74	5.43
Ours	**15.46**	**11.60**	**10.21**	**8.51**	8.43	8.28	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**	**2.77**

Table 4.2: Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity in the horizontal direction, by MR^M_{75}, including their mean and standard deviation over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Note that other methods generate a single bounding box for each object, representing the same position in both modalities, while our method generates paired bounding boxes. Bold values indicate the best performance.

Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN[32]	93.05	89.46	83.55	76.78	71.70	70.10	70.97	77.29	84.71	91.35	95.46	82.22	9.35
AR-CNN[61]	94.25	91.64	87.64	78.70	69.57	61.77	65.13	71.56	81.81	90.29	94.79	80.65	12.06
MBNet[64]	90.89	87.50	80.33	70.81	63.63	58.82	63.15	71.33	80.75	89.27	93.87	77.30	12.39
Ours [51]	**63.30**	**59.94**	**56.67**	**55.87**	**55.45**	**55.07**	**55.39**	**56.35**	**57.01**	**59.72**	**62.58**	**57.94**	**2.96**

Qualitative Comparison

We provide visualization of detection results from several state-of-the-art methods on four scenes from KAIST [24] test set with a varied amount of



(a) (b)

Figure 4.7: Comparison of state-of-the-art methods' performance on KAIST dataset by mean miss rate (solid lines) and worst miss rate (dashed lines) of eleven different simulated disparities to FPPI curves. Numbers in legend show geometric mean of mean miss rate and geometric mean of worst miss rate over FPPI in the range of $[10^{-2}, 10^{0}]$ of each method. (a) IoU^{M} threshold of 0.5, (b) IoU^{M} threshold of 0.75.

misalignment to measure each method's quality in terms of detection precision and reliability. For methods that only have detection bounding boxes localizing objects in one modality, we replicated detection bounding boxes in the other modality with the same position and showed them as dashed line bounding boxes, i.e., we replicated thermal bounding boxes for MSDS-RCNN [32] and MBNet [64], visible bounding boxes for AR-CNN [61].

Scene N1 is a scene at nighttime with no misalignment between modalities and without simulated disparity. Scene D1 is a scene at daytime with significant distortion, causing each pedestrian to have different disparities, especially the leftmost pedestrian, without simulated disparity. Scene N2 is a scene at nighttime with slightly weak misalignment between modalities, and we add simulated disparity by shifting the thermal image by 10 pixels to the left direction. Scene D2 is a scene at daytime with huge misalignment from the beginning. We then add simulated disparity by shifting the thermal image by 10 pixels to the right direction for larger misalignment. We evaluate the mean IoU^{M} of all detection bounding boxes with the highest IoU^{M} overlap with each ground truth bounding box with at least 0.01 prediction score for

each scene in order to measure the quality of bounding boxes of each method. As illustrated in Figure 4.8, in Scene N1, our method could not demonstrate its strength since there is no misalignment. Moreover, the multi-modal regressor also causes detection bounding boxes in visible modality to regress without clear information instead of staying at the same place as thermal bounding boxes, degrading the precision even further. In Scene D1, we can notice the more precise detection bounding boxes of the proposed method, especially at the leftmost pedestrian. Our method was able to adjust the detection bounding boxes in thermal modality closer to the pedestrian, despite the extreme misalignment. Still, the adjustment is not so great, which might be caused by the unusual pedestrian in dark color in thermal modality instead of bright color. In Scene N2 and D2, when misalignment is large, our method clearly shows its advantage by adjusting bounding boxes' positions for both modalities, resulting in the highest mIoU^{M}.

4.3.5 Ablation Study

We conduct ablation experiments to investigate our network's components and analyze each component's effectiveness. First, we compare the performance of a typical two-stream Faster R-CNN with our proposed models composed of either only multi-modal RPN or multi-modal detector and both of them. Second, we compare the performance of our proposed models trained and tested by different NMS criteria for both RPN and detector. Lastly, we compare the performance of our proposed models trained by different mini-batch sampling criteria for both RPN and detector.

Regressor comparison

As illustrated in Table 4.3, we can see that multi-modal regressor RPN does not significantly improve from a single-regressor, showing that the detector could not fully utilize RPN's output. Unquestionably, the performance improves drastically when a multi-modal regressor detector is implemented, especially with large misalignment, showing the benefit of locating objects in each modality individually. Finally, our network with both components has the best performance, indicating the effectiveness of both multi-modal RPN and multi-modal detector combined.

Table 4.3: Comparison of the proposed multi-modal Faster R-CNN consisting of different components on KAIST dataset, with simulated disparity in the horizontal direction, by MR_{50}^{M}, including their mean and standard deviation over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

RPN's	Detector's	Thermal images' horizontal shift distance (px)											Mean
regressor	regressor	-10	-8	-6	-4	-2	0	2	4	6	8	10	
Single	Single	24.41	16.42	13.08	11.03	10.19	10.07	11.03	11.66	13.69	15.81	21.15	14.41
Multi	Single	24.33	17.04	12.92	10.57	9.93	9.99	10.59	11.48	12.49	15.09	20.46	14.08
Single	Multi	18.54	13.25	**10.20**	9.17	8.57	8.96	9.21	9.90	11.44	13.19	17.18	11.78
Multi	Multi	**15.46**	**11.60**	10.21	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**

NMS comparison

As illustrated in Table 4.4, IoU^{M} outperforms other criteria in this experiment. IoU^{T} also performs surprisingly well as the first runner-up. The reason might be that most of the pedestrians in KAIST dataset can be recognized by thermal modalities alone, and the misalignment is not significant in most test sets. Nevertheless, this experiment shows that IoU^{M} can be utilized as NMS criteria for multi-modal pedestrian detection and perform the best compared to other criteria.

Table 4.4: Comparison of the proposed multi-modal Faster R-CNN consisting of different NMS criteria on KAIST dataset, with simulated disparity in the horizontal direction, by MR_{50}^{M}, including their mean and standard deviation over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

NMS	Thermal images' horizontal shift distance (px)											Mean
criteria	-10	-8	-6	-4	-2	0	2	4	6	8	10	
IoU^{V}	16.17	13.18	11.67	10.71	11.10	10.82	11.42	12.10	13.38	14.56	18.76	13.08
IoU^{T}	15.86	12.24	10.33	9.08	8.82	9.00	9.08	10.17	11.41	13.30	17.60	11.54
OR	16.95	12.64	11.37	9.92	10.24	9.85	10.06	10.72	12.09	13.97	16.53	12.21
AND	20.51	15.18	11.84	9.98	9.50	9.38	9.43	10.29	11.51	15.05	18.98	12.88
IoU^{M}	**15.46**	**11.60**	**10.21**	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**

Mini-batch sampling comparison

From Table 4.5, IoU^{M} has the best performance when shift distance is less than or equal to 8 and also achieves the lowest mean MR^{M}. However, IoU^{M} is inferior to OR when the shift distance is 10. The reason could be that OR is better at learning extreme cases, such as very large misalignment. Still, it has worse performance at no or weak misalignment as a trade-off. Overall, IoU^{M} has the best performance. It is worth considering how to make a multi-modal network perform well at any level of misalignment.

Table 4.5: Comparison of the proposed multi-modal Faster R-CNN trained by different mini-batch sampling criteria on KAIST dataset, with simulated disparity in the horizontal direction, by MR_{50}^{M}, including their mean and standard deviation over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

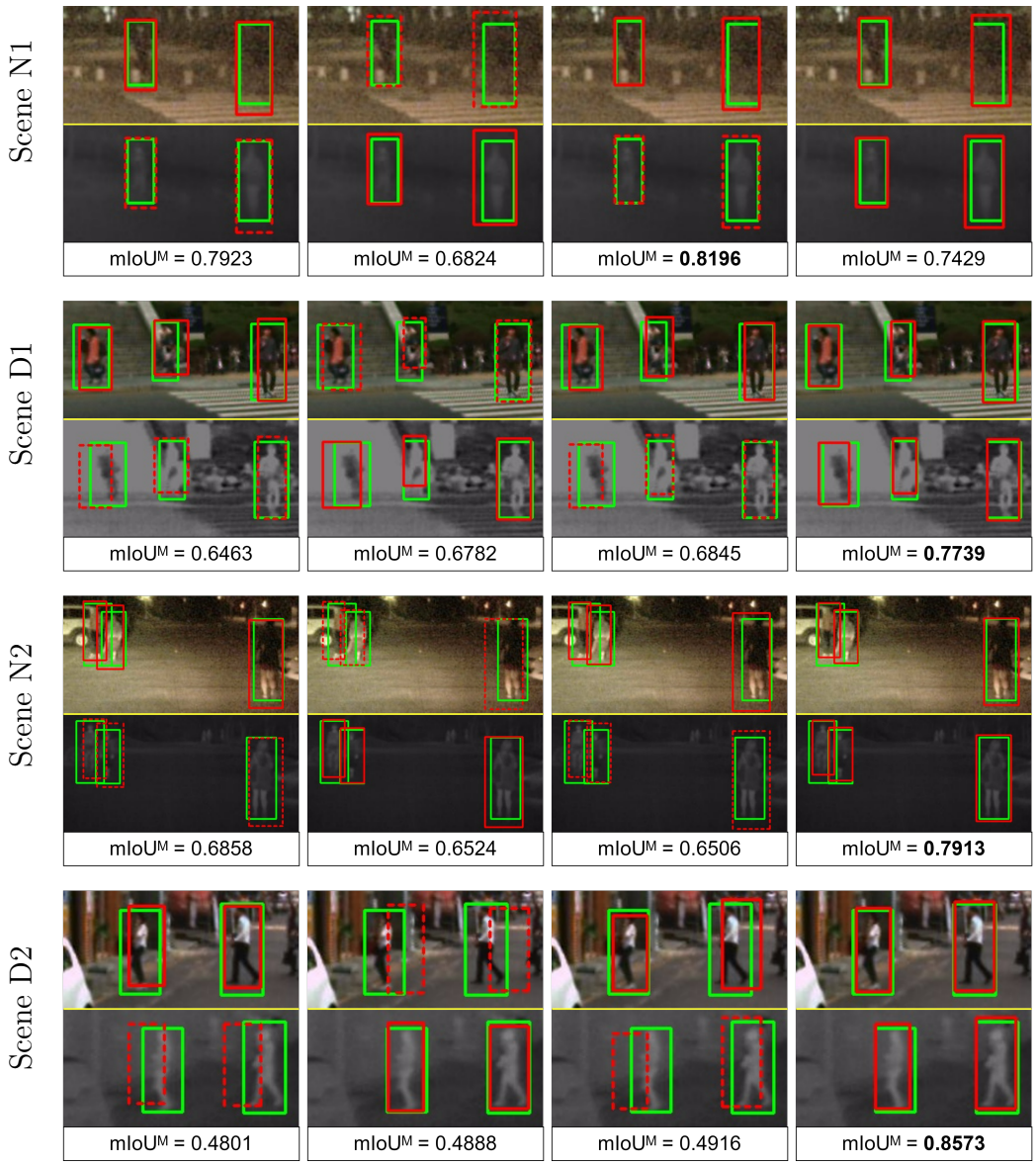
Sampling criteria	Thermal images' horizontal shift distance (px)											Mean
	-10	-8	-6	-4	-2	0	2	4	6	8	10	
IoU^{V}	20.77	16.82	14.42	13.38	13.28	13.25	13.27	14.73	15.36	17.34	21.31	15.81
IoU^{T}	14.75	11.89	10.90	9.62	9.65	9.76	9.53	10.16	10.85	13.40	16.30	11.53
OR	**14.54**	11.66	10.70	10.03	9.07	9.23	10.07	10.10	11.40	12.91	**15.26**	11.36
AND	17.03	13.07	10.92	9.97	9.39	9.46	9.62	10.76	11.74	14.38	18.52	12.26
IoU^{M}	15.46	**11.60**	**10.21**	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	15.87	**10.80**

4.4 Conclusion

In this chapter, we have analyzed the current misalignment problem of existing multi-modal pedestrian detection methods. We proposed a novel multi-modal detection method based on multi-modal regressor and IoU^{M}, consisting of multi-modal NMS, multi-modal RPN, and a multi-modal detector. Employing the proposed IoU^{M} metric during the training process, we ensured that our method could independently localize pedestrians in each modality and maintain their paired relations, demonstrating robustness to large misalignment.

Our experiments, utilizing the Multi-Modal Log-Average Miss Rate (MR^{M}) for evaluation, demonstrated that our proposed method achieves the best

performance compared to state-of-the-art methods when the misalignment is large or the precision requirement of bounding boxes is high. This robustness to misalignment and superior precision of detection bounding boxes in both modalities highlights the effectiveness of our approach. However, the performance of our method in scenarios where there is no misalignment still requires improvement. Even though we achieve the best performance when misalignment is significant, it is not yet reliable enough for critical real-life applications such as autonomous driving, where there is no room for error.



(a) MSDS[32] (b) AR-CNN[61] (c) MBNet[64] (d) Ours

Figure 4.8: Qualitative comparison examples of detection results on KAIST dataset of (a) MSDS-RCNN [32], (b) AR-CNN [61], (c) MBNet [64], and (d) ours. Scene N1 and D1 are original test images without simulated disparity. Scene N2 and D2 are test images with simulated disparity from shifting 10 pixels to the left and right direction, respectively. Green bounding boxes represent ground truth by Zhang et al. [61], and red bounding boxes represent detection results. Dashed line bounding boxes denote substituted bounding boxes for methods that do not have paired bounding boxes.

Chapter 5

Proposed Multi-Modal Single Shot MultiBox Detector Considering Misalignment

In this chapter, we explain the principle of our proposed Single Shot MultiBox Detector Considering Misalignment in detail.

5.1 Background

Pedestrian detection is one of the important research topics in computer vision [6, 5], with major applications in areas such as video surveillance systems and autonomous driving. The one-stage detection network [13, 36], including SSD [37], is one of the standard architectures for many of these vision applications. However, relying solely on the visible modality has limitations, particularly in adverse lighting conditions or cluttered backgrounds. To surpass these limitations, multiple modalities, such as visible and thermal modalities, have been used together for pedestrian detection.

Recent studies have introduced various approaches to combine information from different modalities. One of the main challenges of multi-modal pedestrian detection is the misalignment problem. Most existing methods assume that the alignment of image pairs is nearly perfect. These methods suffer performance degradation when there is misalignment. More recent methods [61] directly addressed this issue. They proposed an alignment module to adaptively align features between two modalities, which improved robustness against misalignment. However, their performance is still lacklustre when the degree of misalignment is large. Furthermore, their method is only applicable to two-stage detection networks.

In Chapter 4, we proposed a multi-modal Faster R-CNN model, which achieved state-of-the-art performance. However, there are several limitations. Despite its accuracy improvements over other state-of-the-art methods in misaligned environments, Faster R-CNN's two-stage nature complicates the

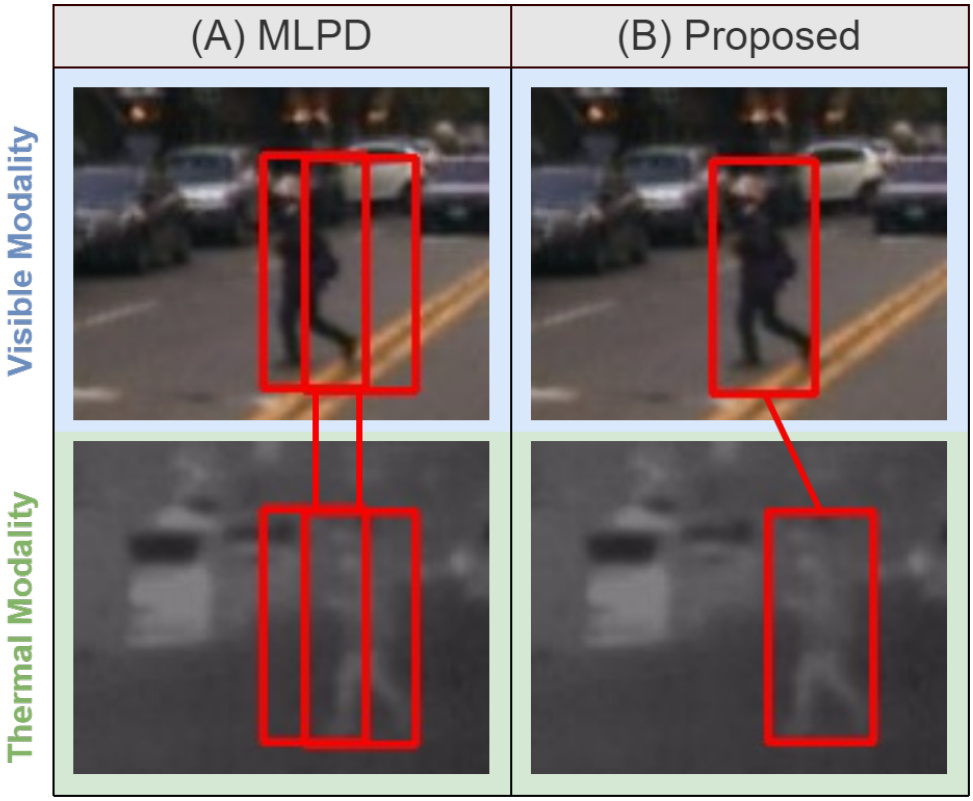
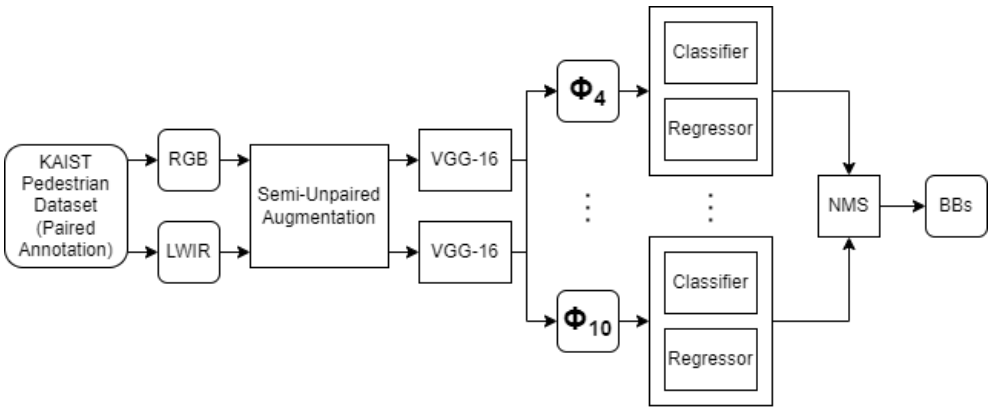


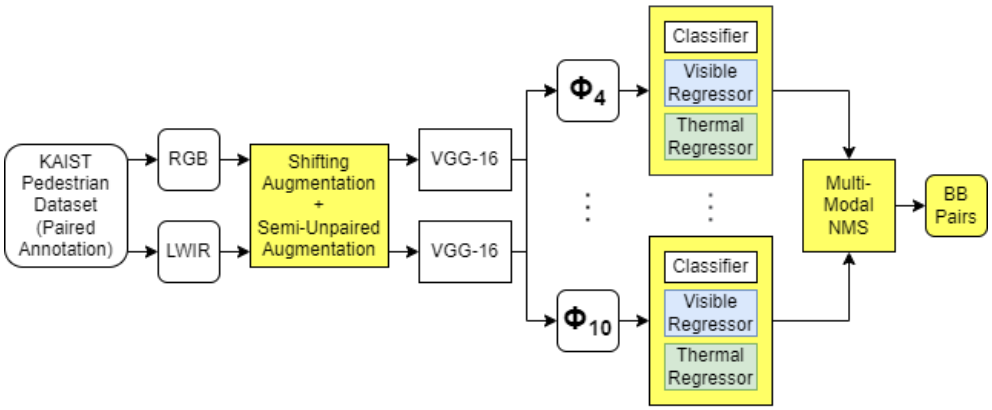
Figure 5.1: Visualization examples of detection results on KAIST dataset by (A) MLPD [26] and (B) our proposed method. Red boxes represent predicted bounding boxes and the lines between them indicate their pair relation.

fine-tuning and optimization of our multi-modal regressor. This complexity makes it challenging to maintain good performance in both no-misalignment and large-misalignment conditions, resulting in subpar performance when there is no misalignment. Additionally, the computational intensity of Faster R-CNN makes it less suitable for real-time applications.

Building on the approach taken by the Multi-Label Pedestrian Detector in the Multispectral Domain (MLPD) [26], which has shown the ability to handle unpaired visible and thermal images, we attempt to adopt MLPD's unpaired data handling techniques to improve our solution to the misalignment problem. This could be a promising approach to address large misalignment. However, as shown in Figure 5.1 (A), even with MLPD, significant misalignment issues still result in performance degradation. Specifically,



(a) MLPD



(b) Proposed Method

Figure 5.2: Comparison of multi-modal pedestrian detection frameworks based on SSD [37]. Φ_i indicates the fused feature map of feature layer i. Yellow blocks represent notable changes introduced in our method. (a) MLPD [26], (b) proposed method.

MLPD, which is trained to handle unpaired cases, perceives the same pedestrian as two different pedestrians when there is misalignment, resulting in two bounding box pairs for one person.

In this chapter, we propose a novel Multi-Modal Single Shot MultiBox

Detector that considers misalignment. Our model leverages the efficiency of SSD and incorporates innovative techniques to enhance the alignment and integration of multi-modal data. By addressing the limitations of both Faster R-CNN and MLPD, and inheriting the multi-modal regressor from our previous work to a single-stage network, we aim to achieve superior performance in multi-modal pedestrian detection, making our model robust against large misalignments while maintaining high performance in no-misalignment cases. As illustrated in Figure 5.1 (B), the pedestrian is perceived as a single object, and the proposed method correctly locates them precisely in both modalities despite misalignment.

To address the limitations identified in the MLPD framework, we propose several key enhancements. We incorporate a multi-modal regressor designed to handle the differences between modalities, improving the accuracy of bounding box predictions. Our method uses object-based training, which ensures that both visible and thermal modalities contribute to the learning process, allowing our model to precisely locate pedestrians in both modalities, even in the presence of significant misalignment. Shifting data augmentation is introduced to improve the model's ability to handle misalignment by simulating various degrees of misalignment during training. Additionally, we employ a multi-modal non-maximum suppression (NMS) technique that combines the outputs from both modalities to refine the final detections, reducing false positives and enhancing overall detection accuracy. Finally, our approach generates paired bounding boxes for each modality, acknowledging that objects may appear in different positions in visible and thermal images, thus providing more accurate localization. A comparative overview of the traditional MLPD approach and our proposed methodology is illustrated in Figure 5.2.

We will discuss the methodology of our proposed model, detailing the integration of a multi-modal regressor for single-stage networks, object-based training strategies, and the implementation of multi-modal non-maximum suppression (NMS) and shifting data augmentation. An extensive evaluation of our model's performance will be presented, comparing it with existing methods.

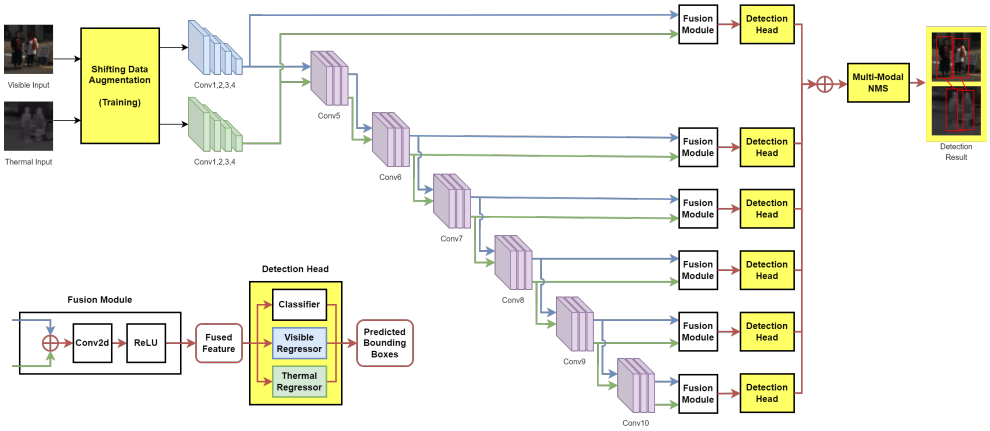


Figure 5.3: The overall architecture of our network. The framework is based on SSD [37] customized by MLPD [26]. Yellow blocks represent notable changes introduced in our method: shifting data augmentation in the training phase, detection heads with visible and thermal regressors, multi-modal NMS, and detection outputs consisting of pairs of bounding boxes. Blue, green, and red blocks/paths represent properties of visible modality, thermal modality, and fused modalities, respectively. $\oplus$ denotes channel-wise concatenation.

5.2 Methodology

We present one-stage multi-modal pedestrian detection framework inspired by SSD [37] and MLPD [26], as shown in Figure 5.3. The visible and thermal inputs initially follow distinct branches, proceeding through shared convolutional layers. Note that shifting data augmentation is only applied in the training phase, which is an addition to semi-unpaired augmentation of MLPD [26]. They are then unified within a fusion module before being input into the detection head, where we implement the proposed multi-modal regressor. The final output of the network is a set of bounding box pairs locating objects in both modalities

5.2.1 Multi-Modal Regressor For Single-Stage Network

We propose a refinement to the conventional single-stage detection head, introducing a multi-modal regressor approach for multi-modal detection. This approach is inspired by the multi-modal regressor described in Chapter 4, which successfully addresses the misalignment issue by predicting bounding box coordinates separately for each modality. The overall architecture of our network is based on the Single Shot MultiBox Detector (SSD) [37] framework, customized to incorporate multi-modal capabilities as shown in Figure 5.3. The visible and thermal inputs initially follow distinct branches, proceeding through shared convolutional layers. They are then unified within a fusion module before being input into the detection head, where we implement the proposed multi-modal regressor. Each detection head includes a regressor for each modality, which independently adjusts the positions and sizes of bounding boxes. However, a single shared classifier predicts the confidence score for each pair of bounding boxes. This design ensures that the bounding boxes are precisely aligned in each modality while maintaining the paired relations. The final output of the network is a set of bounding box pairs locating objects in both modalities.

In our SSD approach, there are no separate region proposals as in Faster R-CNN. Instead, the SSD framework predicts the bounding boxes and their associated confidence scores in a single step. As depicted in Figure 4.4 (b), the multi-modal SSD operates similarly to the multi-modal RPN in terms of structure and function. Both systems use separate regressors for each modality and employ multi-modal NMS to refine the outputs. However, the key difference lies in the detection process. The multi-modal RPN is part of a two-stage detector that first generates region proposals, whereas the SSD approach directly predicts the final bounding boxes and confidence scores in a single step. This streamlined process reduces computational complexity and allows for real-time applications, making it faster and less complicated to train than the two-stage Faster R-CNN approach.

In contrast to SSD, our modified loss function accounts for the multi-modal regressor setup, incorporating distinct regression losses for each modality-specific regressor. The overall loss function is expressed as follows:

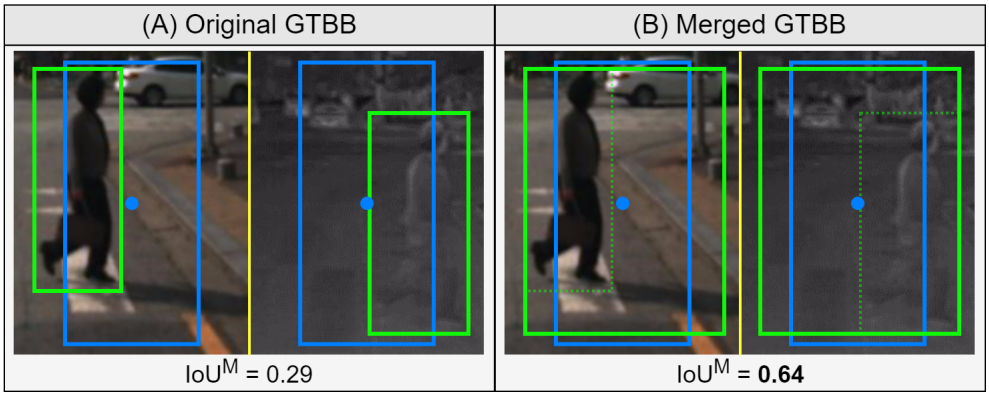


Figure 5.4: Visualization examples of (A) a ground truth bounding box (GTBB) pair from KAIST-paired annotation by [61] and (B) a merged ground truth bounding box. Green boxes represent ground truth bounding boxes. Blue boxes represent anchor boxes. Images are artificially shifted for better understanding.

$$
\begin{aligned}
L = & \sum_i L_{cls}\left(BB_i, \widehat{BB}_i\right) \\
& + \sum_i \left[w_i^v L_{reg}\left(BB_i^v, \widehat{BB}_i^v\right) + w_i^t L_{reg}\left(BB_i^t, \widehat{BB}_i^t\right)\right],
\end{aligned}
\tag{5.1}
$$

where i denotes the index of the anchor box, a predefined bounding box positioned at various points throughout the images. These anchor boxes serve the purpose of identifying objects within specific, designated regions. L_{cls} denotes classification loss, which is binary cross entropy (Equation 4.2) with sigmoid activation function. L_{reg} denotes regression loss, employing smooth L1 loss (Equation 4.3 and 4.4). BB_i^v, BB_i^t denote the visible and thermal ground truth bounding boxes of anchor box i, respectively. $\widehat{BB}_i^v$, $\widehat{BB}_i^t$ denote the predicted visible and thermal bounding boxes of anchor box i, respectively. w_i^v, w_i^t denote the visible and thermal mask, determined by multi-label of the object, adopted from MLPD [26]. In essence, w_i^v, w_i^t are set to 1 when the corresponding object is perceivable in the visible or thermal modality, respectively; otherwise, they are set to 0.

5.2.2 Object-Based Training

As discussed in the context of the multi-modal regressor, visible and thermal ground truth bounding boxes is crucial for training the multi-modal regressor. We adopted the KAIST-paired annotation developed by Zhang et al. [61]. While various methods have employed this annotation in different ways, they often did not leverage its full potential. For instance, MBNet [64] merges visible and thermal bounding boxes of each pedestrian into a unified bounding box by averaging. MLPD [26] considers the same pedestrian in both visible and thermal modalities as two distinct objects, a methodology we will now label as 'bounding box-based (BB-based) training'. In contrast, our approach considers each pedestrian as a single object with two individual coordinates for visible and thermal modalities. For unpaired objects visible exclusively in one modality, whether solely in the visible or thermal domain, they are categorized as either visible-only or thermal-only objects, respectively. Subsequently, these objects are utilized to exclusively train either the visible or thermal regressor. This training approach is referred to as 'object-based training.' This distinction allows our method to precisely locate pedestrians in both modalities, even in the presence of significant misalignment.

In the sampling process, positive samples are chosen based on the overlap between each anchor box and the ground truth bounding box. To account for potential misalignment, We unify visible and thermal bounding boxes into a single bounding box by utilizing the farthest points in both horizontal and vertical directions from the vertices of the original bounding boxes. This consolidation can improve overlap computation with the anchor box, thereby reducing the likelihood of overlooking potential samples with significant misalignment. However, during regressor training, we maintain the original ground truth bounding boxes as targets for the proposed multi-modal regressors. The visualized example of ground truth bounding boxes is depicted in Figure 5.4, where the original ground truth bounding box (Figure 5.4 (A)) serves as the target for our regressors' training: the visible regressor is trained with the visible ground truth bounding box, and the thermal regressor is trained with the thermal ground truth bounding box. In the sampling process, we utilize the merged ground truth bounding box (Figure 5.4 (B)) to calculate the overlap with the anchor box. This approach enhances the overlap measurement, especially when misalignment is significant. Here, IoU^M increases from 0.29 to 0.64. This increase could be pivotal, potentially changing the sample's classification from negative to positive.

5.2.3 Multi-Modal NMS

Our Non-Maximum Suppression (NMS) utilizes IoU^M as a suppression criterion to preserve pair relations between bounding boxes in the visible and thermal modalities. The process starts by categorizing each bounding box pair into three groups: visible-thermal, visible-only, or thermal-only objects, based on the prediction scores of both modalities. Pairs with prediction scores below the specified threshold for both modalities below the specified threshold are classified as background and discarded. If only one modality's prediction score surpasses the threshold, the pair is designated as a modality-specific object. Otherwise, it is identified as a visible-thermal object. Next, the bounding box pairs are sorted in descending order based on the average prediction scores of the visible and thermal modalities. The overlap calculation between pairs considers IoU^M, IoU of the visible modality (IoU^V), and IoU of the thermal modality (IoU^T). For visible-only or thermal-only objects, only the bounding box in the corresponding modality is considered, with the other modality treated as non-existent. When any of the overlap thresholds is exceeded, the bounding box pair with the lower score is suppressed.

By categorizing objects and applying NMS accordingly, this strategy ensures that each object type is accurately processed based on its unique characteristics. The object-based training approach enhances the multi-modal detection process, allowing us to eliminate redundant bounding boxes more effectively while maintaining the paired relations. As a result, the overall detection accuracy and robustness are significantly improved, especially in challenging conditions where objects may not be consistently visible in both modalities.

5.2.4 Shifting Data Augmentation

Furthermore, we incorporate shifting data augmentation to expose our network to misalignment scenarios, addressing a gap in the original dataset. This augmentation method involves randomly translating training images horizontally in one of the two modalities, with pixel shifts ranging from -10 to 10. This process is facilitated by a multinomial distribution, with probabilities derived from a normal distribution, to randomly determine the shift distance. As shown in Figure 5.5, we illustrate the shifting data augmentation process. For each training image, a shift distance in the range

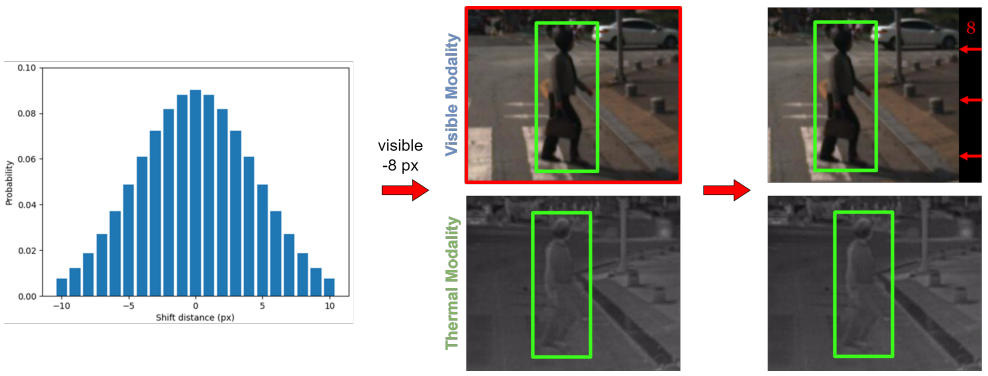


Figure 5.5: shifting data augmentation (DA) process used in our multi-modal pedestrian detection method. For each training image, a shift distance in the range of [-10, 10] pixels is randomly picked based on a multinomial distribution (normal distribution with a standard deviation of 4). The histogram on the left shows the probability distribution of the shift distances. In this example, the random result is a shift of -8 pixels for the visible modality image. The resulting shifted images for both the visible and thermal modalities are shown on the right, with green bounding boxes indicating the ground truth pedestrians.

of [-10, 10] pixels is randomly picked based on a multinomial distribution (normal distribution with a standard deviation of 4). The histogram on the left shows the probability distribution of the shift distances. In this example, the random result is a shift of -8 pixels for the visible modality image. The resulting shifted images for both the visible and thermal modalities are shown on the right, with green bounding boxes indicating the ground truth pedestrians. This augmentation strategy increases the amount of training data with varying degrees of misalignment, enhancing the robustness of our model in effectively handling misalignment scenarios.

Initially, the entire network is trained without shifting data augmentation. Subsequently, upon achieving a well-performing model on the validation dataset, we proceed to freeze all layers of the network except the regressors and re-train the previous checkpoint with shifting data augmentation. This step further enhances localization performance, particularly when dealing with misalignment data. This augmentation strategy contributes to the robustness of our model in handling misalignment challenges. Subsequently, other semi-unpaired augmentations adopted from MLPD [26] are still ap-

plied.

5.3 Experiment

5.3.1 Dataset

The KAIST dataset [24] stands out as one of the extensively utilized multimodal pedestrian datasets, featuring over 90,000 frames recorded during both day and night to account for varying light conditions. Initially presumed to be geometrically aligned, the dataset's annotations revealed numerous errors, including imprecise localization, misclassification, and misaligned regions, as reported by prior studies [32, 61]. To address these issues, several researchers [25] have created improved versions of annotations as alternatives to the original dataset. Among these alternatives, the annotations provided by Liu et al. [25] have been officially designated as the standard for performance benchmarking. Acknowledging the misalignment issues within the KAIST dataset, Zhang et al. [61] conducted a careful analysis and became the pioneers in offering paired annotations for the dataset. Their approach involved locating objects for each modality individually and establishing pair relations. Their annotations have become crucial for subsequent researchers aiming to address misalignment problems in their work.

5.3.2 Implementation and Details

We adopted an SSD modified into MLPD [26]. The architecture utilized VGG16 pre-trained on ImageNet with batch normalization for Conv1 to Conv5, and the remaining convolutional layers (Conv6 onwards) were initialized with values drawn from a normal distribution (std=0.01). The model underwent training with Stochastic Gradient Descent (SGD), using an initial learning rate, momentum, and weight decay of 0.0001, 0.9, and 0.0005, respectively. The mini-batch size was set to 6, and the input image size was resized to 512 (H) x 640 (W) We integrated MLPD's semi-unpaired data augmentation, maintaining the same parameters, and introduced our shifting data augmentation to bolster the training process against misalignment. The standard deviation of the normal distribution for the shifting data augmentation was set to 4. The prediction score threshold of NMS is set to 0.1. The overlap threshold IoU^M, IoU^V, and IoU^T of NMS are set to 0.425, 0.75,

and 0.75, respectively. First, we train the whole network without shifting data augmentation for 30 epochs. Then, we continue the training from last checkpoint only on multi-modal regressor with shifting data augmentation for another 30 epochs.

5.3.3 Evaluation Details

We conducted our experiments using the KAIST Dataset [24]. Given our specific focus on addressing the misalignment problem, we adopted the annotations provided by Zhang et al. [61] for both training and testing. Recognizing that the test data did not include sufficient scenes with significant misalignment, we conducted a "simulated disparity experiment" to replicate misalignment at various degrees, which we used in Chapter 4. In this setup, we horizontally shift the thermal images of the test data by 2, 4, 6, 8, and 10 pixels in both directions, while the visible images remained unchanged. This process results in 11 subsets of test data with different degrees of misalignment. This experiment provides a clear understanding of the influences of misalignment at different levels.

We evaluated the performance of our methods against all available state-of-the-art methods with accessible source code. For methods producing a single bounding box for each object, we substituted visible and thermal bounding boxes with that single bounding box. The detection performance was quantified using the Multi-Modal Log-Average Miss Rate (MR^M) over the range of $[10^{-2}, 10^0]$ False Positive Per Image (FPPI) with an IoU^M threshold of 0.5 (MR^M_{50}). This evaluation metric allowed us to assess the methods consistently and comprehensively. Furthermore, for a fair comparison, since we are the only method that generates paired detection bounding boxes for both visible and thermal modalities, while other methods locate the pedestrian in only one of the two modalities, we introduced visible MR ($\mathrm{MR^V}$) representing MR based on $\mathrm{IoU^V}$, and thermal MR ($\mathrm{MR^T}$) representing MR based on $\mathrm{IoU^T}$ in our experiments. These metrics allowed us to evaluate the precision of detection results in both modalities separately, providing a comprehensive assessment of the performance of our multi-modal detection system.

Table 5.1: Comparison with state-of-the-art methods on KAIST dataset, on simulated disparity experiment by MR^M_{50}, mean, and standard deviation across shifted distances. Note that other methods generate a single bounding box for each object, representing the same position in both modalities, while our method generates paired bounding boxes. Bold values indicate the best performance.

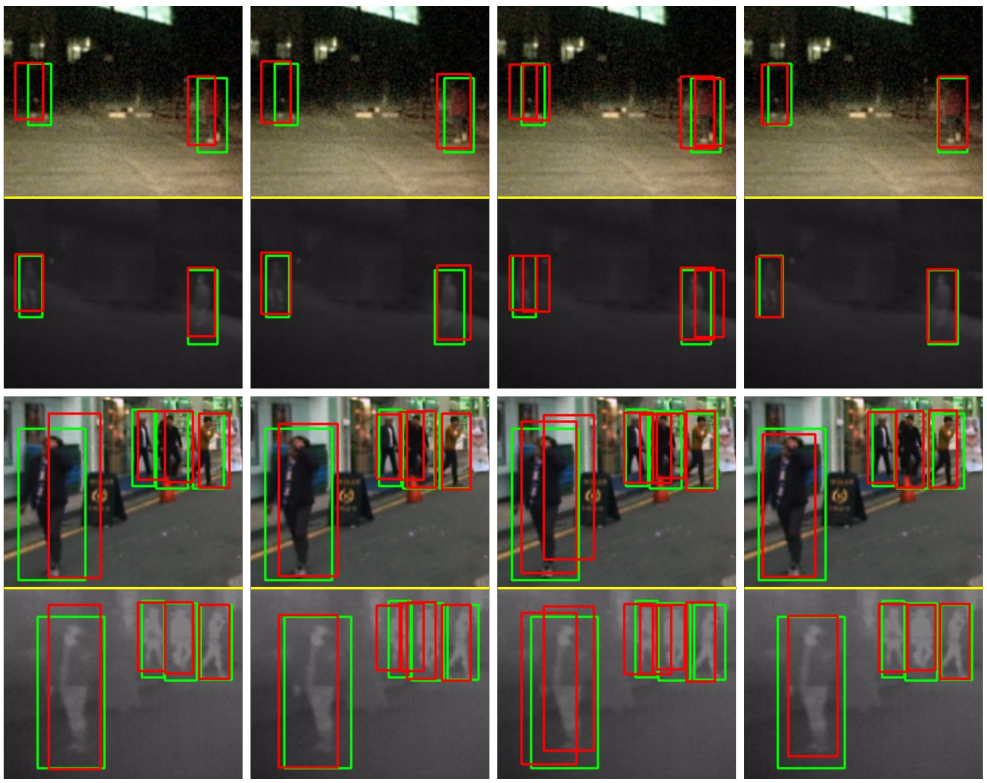
Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN [32]	27.06	18.76	15.93	12.74	12.58	11.09	11.72	13.25	15.06	21.38	27.48	17.00	5.94
AR-CNN [61]	21.61	14.65	10.43	8.67	8.22	8.79	8.68	10.10	11.02	14.65	19.84	12.42	4.69
MBNet [64]	23.14	15.31	11.02	8.92	7.70	7.76	8.64	9.88	11.17	14.87	21.70	12.74	5.43
MLPD [26]	21.33	13.07	9.57	**7.10**	7.07	6.97	7.89	9.49	10.59	15.27	21.86	11.84	5.48
Our Faster R-CNN [51]	15.46	11.60	10.21	8.51	8.43	8.28	8.50	9.14	10.31	12.51	15.87	10.80	**2.77**
Ours [52]	**14.82**	**10.21**	**8.20**	7.27	**6.76**	**6.84**	**7.21**	**8.34**	**9.35**	**11.93**	**15.22**	**9.65**	3.08

Table 5.2: Comparison with state-of-the-art methods on KAIST dataset, on simulated disparity experiment by MR^V_{50}, mean, and standard deviation across shifted distances. Bold values indicate the best performance.

Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN [32]	30.43	21.90	16.91	12.98	12.24	11.28	12.36	14.30	17.92	24.73	33.16	18.93	7.65
AR-CNN [61]	70.83	57.92	37.70	19.14	11.40	9.12	13.41	17.54	27.80	43.26	61.39	33.59	22.05
MBNet [64]	38.56	26.42	16.25	10.27	8.82	7.89	8.68	10.38	13.81	19.19	29.05	17.21	10.11
MLPD [26]	32.17	23.01	15.20	10.22	8.01	7.41	**7.84**	9.91	15.05	23.56	32.08	16.77	9.47
Our Faster R-CNN [51]	**19.33**	**16.43**	**12.86**	10.83	10.10	9.32	9.63	10.09	**11.87**	**13.69**	**17.99**	**12.92**	**3.53**
Ours [52]	23.40	17.53	12.98	**9.58**	**7.80**	**7.37**	7.93	**9.09**	12.88	16.78	23.29	13.51	5.98

Table 5.3: Comparison with state-of-the-art methods on KAIST dataset, on simulated disparity experiment by MR^T_{50}, mean, and standard deviation across shifted distances. Bold values indicate the best performance.

Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN [32]	38.89	29.52	22.02	15.92	14.35	12.51	13.92	16.28	20.17	27.71	34.88	22.38	9.09
AR-CNN [61]	**11.23**	**9.64**	**8.82**	**8.09**	8.04	9.05	8.16	**8.22**	**8.97**	**10.13**	**10.23**	**9.14**	**1.05**
MBNet [64]	26.06	18.82	13.14	10.04	8.99	8.12	9.89	12.35	16.05	22.84	28.60	15.90	7.21
MLPD [26]	23.06	16.03	12.78	9.22	**7.63**	7.94	9.13	10.43	13.99	18.37	24.81	13.94	6.00
Our Faster R-CNN [51]	12.92	11.31	10.41	8.83	8.45	8.55	8.46	9.40	9.85	11.41	12.87	10.22	1.69
Ours [52]	14.10	12.08	11.59	9.02	8.00	**7.62**	**7.68**	10.04	10.19	12.33	15.91	10.78	2.71



(a) AR-CNN [61] (b) MBNet [64] (c) MLPD [26] (d) ours

Figure 5.6: Qualitative comparison examples of detection results on KAIST dataset of (a) AR-CNN [61], (b) MBNet [64], (c) MLPD [26], and (d) ours. Green boxes represent ground truth bounding boxes. Red boxes represent predicted bounding boxes. Image pairs are cropped in the same position to make the contrast between methods more apparent. Prediction score threshold is set to 0.1. Thermal images of scene 1 and 2 are shifted to the left and right direction by 10 pixels, respectively.

5.3.4 Comparison with existing methods

Performance comparison.

Table 5.1 shows the performance comparison between various state-of-the-art methods, including our past work [50] on MR^M. The proposed method emerges as the top-performing solution, consistently outperforming state-

of-the-art approaches across various simulated disparity distances on the KAIST dataset. Specifically, at smaller misalignment distances, our model showcases performance comparable to the MLPD baseline, indicating that the introduced modifications maintain competitive accuracy under standard conditions. However, the strength of the proposed method becomes clear at larger misalignment distances (e.g., -10 pixels), where it significantly surpasses MLPD and other methods. This performance at larger misalignment distances highlights the effectiveness of the proposed model in addressing challenges associated with substantial misalignment. Additionally, the proposed method consistently outperforms our previous work across all shift distances, demonstrating enhanced robustness and performance in handling misalignment challenges. This improvement is particularly evident at larger misalignment distances, emphasizing the effectiveness of the novel components and strategies incorporated into the proposed model. The mean and standard deviation values further support the reliability and stability of the proposed method across diverse misalignment scenarios.

Tables 5.2 and 5.3 present the comparison of MR^V and MR^T across different methods, respectively. These metrics consider the accuracy of bounding boxes in one modality only, which can lead to good performance in one modality but poor accuracy or missed detections in the other. Our multi-modal Faster R-CNN [52] achieves the best performance in MR^V overall and outperforms our multi-modal SSD on both MR^V and MR^T at larger shift distances, while our multi-modal SSD excels under no misalignment conditions.

The reason our multi-modal SSD excels in MR^M but lags in MR^V and MR^T is that MR^M accounts for the precision of bounding boxes in both modalities simultaneously and ensures correct pedestrian matching across modalities. In contrast, MR^V and MR^T do not consider cross-modality matching, allowing our multi-modal Faster R-CNN to locate pedestrians in one modality more accurately, but perform worse in matching pedestrians across modalities. Interestingly, AR-CNN, which relocates the visible region into the thermal area, shows the best performance in MR^T but performs poorly in MR^V. While our multi-modal SSD may not lead in a single modality, it still achieves top performance among available state-of-the-art methods.

The experimental results indicate that when considering the precision of bounding box pairs and the correct matching of pedestrians across modalities, our proposed MR^M is the most reliable benchmark. Overall, our multi-modal SSD demonstrates superior performance, highlighting its effectiveness

in multi-modal pedestrian detection with weak misalignment when the precision of bounding boxes in both modalities are considered.

Qualitative comparison.

Figure 5.6 illustrates comparison examples of detection results on the KAIST dataset, showcasing the performance of our method against other state-of-the-art approaches. i) First Scene: In a scene where pedestrians are separate but challenging to recognize due to dark lighting and substantial misalignment, our method stands out, producing precise bounding boxes for all pedestrians, whereas alternative methods either struggle to locate pedestrians accurately or generate multiple bounding boxes for a single pedestrian, leading to false positives. ii) Second Scene: In a more crowded scene where pedestrians are numerous and clearly distinct, but serious misalignment is present, our method showcases its ability to generate accurate bounding boxes for all pedestrians. In contrast, other methods encounter challenges in precise localization. For instance, MLPD resorts to creating two bounding box pairs for a single pedestrian, attempting to cover them in both modalities. This example serves as a clear visual representation of the advantages offered by our proposed method. It highlights the effectiveness of our approach in handling misalignment and achieving accurate multi-modal detections in both visible and thermal modalities.

Table 5.4: Performance Evaluation of Varied Components and Training Strategies in the Proposed Network on the KAIST dataset with simulated disparity experiment by MR_{50}^{M}, mean, and standard deviation across shifted distances.

Type of regressor	Training strategy	Shifting data augmentation	Thermal images' horizontal shift distance (px) -10	-8	-6	-4	-2	0	2	4	6	8	10	Mean	SD
Single	BB-based	-	22.91	15.01	9.74	7.87	7.04	7.17	8.43	9.66	11.04	15.19	21.12	12.29	5.57
Single	Object-based	-	20.36	13.27	8.78	7.72	6.93	7.21	7.69	9.18	10.32	13.61	20.41	11.41	4.98
Single	Object-based	✓	20.21	13.65	9.42	7.94	7.28	7.20	7.85	8.61	11.07	13.36	19.60	11.47	4.74
Multi	BB-based	-	19.74	12.92	9.84	8.24	6.90	6.94	7.95	9.50	10.51	14.04	20.34	11.54	4.77
Multi	Object-based	-	17.10	11.59	8.67	7.99	7.35	6.99	7.37	**8.34**	9.45	12.53	16.03	10.31	3.56
Multi	Object-based	✓	**14.82**	**10.21**	**8.20**	**7.27**	**6.76**	**6.84**	**7.21**	**8.34**	**9.35**	**11.93**	**15.22**	**9.65**	**3.08**

5.3.5 Ablation Study

Table 5.4 provides an insightful ablation study, exploring the impact of different components and training strategies on our proposed network's performance. We examined variations in the type of regressor, training strategy (BB-based or Object-based), and the inclusion of shifting data augmentation. The results indicate that the performances of single-regressor networks are almost the same. They could not utilize from the object-based training and misalignment data. Furthermore, the integration of multi-modal regressors, combined with a BB-based training strategy, does not lead to any performance improvement. This is because BB-based training does not consider the relationship between objects in different modalities. In contrast, combining multi-modal regressors with an object-based training strategy ensures precise pedestrian localization, facilitating accurate pedestrian matching even under varying degrees of misalignment, ultimately leading to improved performance. Additionally, the incorporation of shifting data augmentation allows the model to learn from data exhibiting diverse misalignment, providing valuable insights not present in the original training data and contributing to the best performance.

5.4 Conclusion

In this chapter, we proposed a one-stage multi-modal pedestrian detection network leveraging a multi-modal regressor and object-based training to address the misalignment challenges prevalent in existing methods. By integrating the proposed IoU^{M} metric into our training process, we ensured accurate localization and robust performance across different modalities.

The simulated disparity experiments on the KAIST dataset and the proposed MR^{M} demonstrated the superiority of our proposed method, allowing us to clearly see the differences in performance compared to state-of-the-art methods. The combination of a multi-modal regressor, object-based training, and shifting data augmentation collectively contributes to enhanced performance, showcasing the model's robustness in scenarios with varying degrees of misalignment.

These findings affirm that our approach effectively addresses the misalignment problem and significantly improves detection performance compared to existing state-of-the-art methods. Our method presents a lightweight and

high-performance solution for multi-modal pedestrian detection with misalignment, applicable to video surveillance systems and autonomous driving. Future work will aim to further optimize computational efficiency and explore the integration of additional sensor modalities to enhance the robustness and accuracy of multi-modal pedestrian detection systems.

Chapter 6

Discussion and Conclusion

In this chapter, we collect the results from our experimental evaluations and discuss the implications of our findings. We then conclude by summarizing the key contributions of our research and outlining potential directions for future studies.

6.1 Discussion

In this section, we analyze the results from our experimental evaluations and explore the significance of our findings in the context of multi-modal pedestrian detection. The discussion is structured to address the following key areas: performance analysis, effectiveness of proposed evaluation metrics, practical applications, and limitations.

Performance Analysis

Key factors of our proposed method include the incorporation of multi-modal regressor mechanisms, object-based training, shifting image augmentation, and multi-modal non-maximum suppression (NMS). Our experimental evaluations in the ablation study revealed that the proposed methods, which include the multi-modal Faster R-CNN and SSD frameworks, significantly outperform traditional single-modal detection methods.

A major challenge in multi-modal pedestrian detection is the misalignment between visible and thermal images. Our methods address this issue by integrating position regressors into both the Faster R-CNN and SSD frameworks. These multi-modal regressor mechanisms enable independent localization of pedestrians in each modality while maintaining their paired relationships. The robustness of our approach to varying levels of misalignment was confirmed through experimental results, which showed superior performance metrics even in scenarios with artificially induced misalignment. By leveraging complementary information from visible and thermal sensors, our models achieve much higher detection accuracy and robustness, particularly

under challenging conditions such as severe misalignment. This was demonstrated by our simulated disparity experiment, where the results consistently showed improvement in our proposed evaluation metrics as the severity of misalignment increased.

We benchmarked our methods against several state-of-the-art multi-modal pedestrian detection approaches, including MSDS-RCNN, AR-CNN, MBNet, and MLPD. The results showed that our proposed frameworks consistently outperformed these methods, particularly in terms of robustness to misalignment and overall detection accuracy across all artificial misalignment levels. These comparisons underscore the advancements our methods bring to the field of multi-modal pedestrian detection, highlighting their potential for improving detection accuracy and robustness in real-world applications.

However, there are some drawbacks to the multi-modal regressor approach. In scenarios where locating pedestrians in both modalities is unnecessary, the additional computational cost may be considered a waste. Knowing the position of pedestrians in just one modality might be sufficient for certain applications. Additionally, in crowded scenes, the model might incorrectly match nearby individuals as the same person across modalities, leading to mismatches and potential inaccuracies. These drawbacks highlight the trade-offs involved in using a multi-modal regressor and suggest areas for future research to enhance the overall efficiency and accuracy of multi-modal pedestrian detection systems.

Effectiveness of Proposed Evaluation Metrics

We introduced the multi-modal Intersection over Union (IoU^{M}) and the multi-modal Log-Average Miss Rate (MR^{M}) as new evaluation metrics tailored for multi-modal pedestrian detection. These metrics provide a comprehensive assessment of detection performance by accounting for the precision of bounding boxes across both modalities. The use of IoU^{M} and MR^{M} in our evaluations demonstrated their effectiveness in capturing the detailed aspects of multi-modal detection performance. These metrics are particularly valuable as they clearly assess the accuracy of both visible and thermal modalities in the evaluation, offering a complete and accurate measure of the detection system's performance. Experiments utilizing these metrics clearly showed the differences and improvements brought by our methods, particularly in scenarios with significant misalignment, compared to state-of-the-art approaches. By ensuring that both modalities are accurately evaluated,

IoU^{M} and MR^{M} provide a better performance evaluation framework, leading to more reliable and robust detection systems.

Practical Applications and Implications

The practical implications of our research are significant, particularly for applications in autonomous driving and surveillance systems. In autonomous driving, accurate and reliable pedestrian detection is crucial for ensuring safety. The enhanced robustness to environmental challenges and misalignment provided by our methods can contribute to safer autonomous vehicle navigation. Our methods are particularly advantageous when dealing with impractical camera systems that may not be perfectly aligned or calibrated, a common issue in real-world deployments. Furthermore, our models excel in different lighting conditions, effectively leveraging visible and thermal sensors to maintain high detection accuracy in both daylight and low-light scenarios. In surveillance, our methods can improve the accuracy of pedestrian monitoring in diverse conditions, enhancing security and operational efficiency. By improving detection performance across different lighting environments, our methods ensure reliable monitoring and security under a wide range of operational conditions. However, in scenarios where locating pedestrians in both modalities is unnecessary, the additional computational cost may be considered a waste. Knowing the position of pedestrians in just one modality might be sufficient for certain applications, making the multi-modal regressor an over-engineered solution in such cases.

Potential Additional Modalities for Multi-Modal Detection

In multi-modal pedestrian detection, there is potential to enhance detection performance and robustness by incorporating additional sensor modalities beyond the visible and thermal cameras currently used. While we already use infrared (thermal) imaging, integrating more than two modalities simultaneously could provide even greater benefits, assuming adequate data is available. Potential additional modalities include:

1) LiDAR (Light Detection and Ranging):

LiDAR sensors provide precise 3D information about the environment by measuring the time it takes for a laser pulse to reflect off objects and return to the sensor. Integrating LiDAR data can help in accurately determining the

distance and size of detected objects, which is particularly useful in scenarios with complex backgrounds or where precise spatial information is crucial.

2) Radar:

Radar sensors can detect objects and measure their velocity, providing additional information about the movement of pedestrians. This modality is less affected by weather conditions such as rain, fog, or snow, making it a valuable complement to visible and thermal cameras, which may struggle in such environments.

3) Depth Cameras:

Depth cameras, such as stereo cameras or Time-of-Flight (ToF) cameras, capture depth information by measuring the distance to objects in the scene. This helps in distinguishing pedestrians from the background and improving detection accuracy in crowded environments.

Incorporating these additional modalities can provide a more comprehensive understanding of the environment, improve detection accuracy, and enhance robustness against various challenging conditions. Developing advanced fusion techniques to effectively combine information from multiple modalities would be crucial for this integration. This approach requires extensive annotated data for all modalities involved to ensure effective training and performance.

Potential Approach for Using Video/Temporal Information

In real-time applications, we can use temporal information from video streams to enhance the performance of multi-modal pedestrian detection systems. Objects in a scene exhibit continuous motion, and leveraging temporal information helps in understanding these motion patterns over time. Since objects do not appear or disappear instantaneously but move smoothly from one position to another, using data from previous frames can significantly improve detection accuracy. Integrating temporal information from video streams into multi-modal pedestrian detection systems offers several advantages:

1) Improved Tracking:

Temporal information helps in tracking pedestrians across multiple frames,

reducing the likelihood of false positives and missed detections. By maintaining a continuous track of detected pedestrians, the system can more accurately determine their trajectories and predict their future positions.

2) Enhanced Robustness:

Using temporal data allows the system to leverage motion patterns and continuity, which helps in distinguishing pedestrians from other moving objects. Temporal consistency ensures that objects moving coherently are correctly identified as pedestrians, even in cases of partial occlusion or abrupt changes in appearance.

3) Reduction of Misalignment Effects:

Temporal information can aid in compensating for misalignment issues by providing a history of object positions. This historical data can be used to correct the alignment between modalities over time, ensuring that detections remain accurate despite minor misalignments.

4) Contextual Awareness:

Video data enables the system to understand the context of the scene better. For example, pedestrians tend to move in specific patterns, such as walking along sidewalks or crossing streets. By analyzing these patterns over time, the system can improve its detection accuracy and reduce false positives from non-pedestrian objects.

In real-time applications, it is important to note that future frames (e.g., $t + 1$) cannot be used because they are not available during the current processing frame (t). Instead, the system must rely on previous frames to extract temporal information. This constraint necessitates the use of algorithms that can efficiently utilize past data to enhance current detections. Incorporating temporal information involves using algorithms such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) [23] networks, which are designed to handle sequential data. Additionally, optical flow techniques can capture motion information between consecutive frames, providing valuable data for improving detection performance.

Potential Extension to Additional Classes and Activities

Currently, our pedestrian detection system primarily focuses on detecting pedestrians as a single class. However, there is potential to extend the system to include additional classes and predict activities, which can provide more detailed and useful information. Potential extensions include:

1) Specific Classes:

The system can be extended to detect other classes such as cyclists, children, and adults. This would involve training the model with annotated data for these specific classes to enable the detection of diverse types of road users.

2) Predicting Activities:

Beyond just detecting the presence of pedestrians or other classes, the system can be enhanced to predict activities. For example, it could determine whether a pedestrian is walking, running, or standing still, or if a cyclist is riding slowly or fast. This additional context can be valuable for applications like autonomous driving, where understanding the behavior of road users is crucial for making informed decisions.

By incorporating these extensions, the detection system can provide a more comprehensive analysis of the scene, leading to better decision-making and enhanced safety in applications such as autonomous driving and surveillance.

Limitations

While our methods show substantial improvements, there are still areas that require further research. One limitation is that our current approach assumes all pedestrians can be detected in both modalities, as we remove unpaired cases from the evaluation datasets since we focus on the misalignment problem the most. Nevertheless, addressing the detection of unpaired cases is important for improving the applicability of our methods in real-world scenarios. Additionally, we assume that the misalignment is not too extreme, expecting it to be weak (less than 10 pixels), as severe misalignment could result in one pedestrian being considered as two unpaired pedestrians. Our regressor is limited by the predefined bounding boxes (anchors) in both modalities, making it challenging to adjust if the object is not contained within these anchors in both modalities. Another limitation is the computational com-

plexity of our methods, which could impact their real-time implementation and practical deployment. Although computational cost is essential for real-life applications, we have not yet optimized our methods for this aspect. Optimizing the computational efficiency of our models without sacrificing performance is a crucial area for future work. Lastly, in crowded scenes, the model might incorrectly match nearby individuals as the same person across modalities, leading to mismatches and potential inaccuracies. This is particularly problematic when dealing with large groups of people, as the likelihood of such mismatches increases, potentially degrading the overall performance of the detection system. By acknowledging these limitations, we can better understand the trade-offs involved in using a multi-modal regressor and identify areas for future improvement to enhance the overall efficiency and accuracy of multi-modal pedestrian detection systems.

6.2 Conclusion

In this dissertation, we have addressed the critical challenges in multi-modal pedestrian detection, with a particular emphasis on overcoming misalignment between visible and thermal modalities. Our work has led to several key contributions and advancements in the field.

We introduced advanced multi-modal pedestrian detection techniques that effectively handle the challenges posed by misalignment. These methods incorporate sophisticated algorithms designed to enhance the robustness and accuracy of detection systems. Notably, our approaches leverage multi-modal regressor mechanisms, object-based training, shifting data augmentation, and multi-modal non-maximum suppression (NMS). These techniques, along with the implementation of multi-modal Faster R-CNN and Single Shot MultiBox Detector (SSD) frameworks, contribute significantly to improved performance under misaligned conditions.

A key innovation in our work is the generation of paired detection bounding boxes for both visible and thermal modalities. This capability ensures that the same object is accurately localized in both modalities, even in the presence of misalignment. By maintaining paired relations between the bounding boxes in each modality, our method addresses the limitations of traditional approaches that use a single bounding box position for both modalities, which can lead to inaccuracies under misalignment.

We also introduced the multi-modal Intersection over Union (IoU^{M}) met-

ric and the Multi-Modal Log-Average Miss Rate (MR^M). The IoU^M provides a more precise assessment of detection bounding boxes across different modalities, addressing the limitations of traditional single-modal metrics. It was integral to both the training of our models and the non-maximum suppression process, ensuring more accurate and reliable detections. MR^M offers a comprehensive evaluation of detection performance under varying levels of misalignment, providing a realistic measure of the effectiveness of multimodal detection systems.

The concept of simulated disparity experiment was another key contribution, allowing us to systematically assess the robustness of our methods under various levels of misalignment. By artificially introducing various levels of misalignment into the dataset, we were able to rigorously test and validate the effectiveness of our proposed methods. This approach provided deeper insights into how misalignment affects detection performance and highlighted the robustness of our techniques.

Our experimental evaluations were extensive and thorough. We conducted tests on benchmark datasets, systematically evaluating the impact of misalignment on detection accuracy. These experiments demonstrated that our methods significantly outperform existing state-of-the-art approaches, particularly in scenarios with severe misalignment.

In summary, while this dissertation has made significant strides in multimodal pedestrian detection, numerous opportunities for further research and development remain. By continuing to innovate and collaborate across disciplines, we can develop more reliable and effective pedestrian detection systems with wide-ranging applications in safety, security, and transportation.

6.3 Future Work

In this section, we outline several directions for future research to further advance the field of multi-modal pedestrian detection.

1) Real-Time Implementation

Further research is needed to optimize the proposed methods for real-time processing, which is crucial for applications such as autonomous driving where both timely and reliable detection are essential. Achieving a zero-mistake rate in real-time implementation will be critical to ensure safety and reliability.

2) Incorporate More Advanced Techniques

Developing more sophisticated alignment algorithms that can handle severe misalignment conditions and dynamically adjust to varying environmental factors, along with incorporating advanced network architectures, will significantly contribute to improved alignment and detection accuracy.

3) Evaluating on More Diverse Datasets

Evaluating the proposed methods on more diverse datasets will help verify their generalizability and effectiveness across different scenarios, ensuring that they perform well in a variety of real-world conditions.

4) Extending to Other Object Detection Tasks

The principles and methods developed in this research can be extended to other object detection tasks beyond pedestrian detection, potentially benefiting a wide range of applications in computer vision.

5) Integration with Other Modalities

Exploring the integration of additional sensor modalities, such as LiDAR or radar, could further enhance detection performance and robustness.

Bibliography

[1] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Navneet. Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[5] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Proc. CVPR*. IEEE, 2009.

[6] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI*, 2011.

[7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(4):743–761, 2012.

[8] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(8):1532–1545, 2014.

[9] Jing Dong, Byron Boots, Frank Dellaert, Ranveer Chandra, and Sudipta Sinha. Learning to align images using weak geometric supervision. In *International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2018.

[10] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based

models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[11] D. M. Gavrila. Pedestrian detection from a moving vehicle. In David Vernon, editor, *Computer Vision — ECCV 2000*, pages 37–49, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[12] Ujwalla Gawande, Kamal Hajari, and Yogesh Golhar. *Pedestrian Detection and Tracking in Video Surveillance System: Issues, Comprehensive Review, and Challenges*. 01 2020.

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[16] 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*, pages 580–587, 2014.

[17] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[18] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio Manuel López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6), 2016.

[19] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

[20] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 621–626, 2016.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[22] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[24] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015.

[25] Shu Wang Jingjing Liu, Shaoting Zhang and Dimitris Metaxas. Multispectral deep neural networks for pedestrian detection. In *British Machine Vision Conference (BMVC)*, pages 73.1–73.13, 2016.

[26] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. In *IEEE RA-L*, 2021.

[27] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2103–2112, 2015.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Everest Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012.

[29] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, 2017.

[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[32] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, 2018.

[33] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[34] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996, 2018.

[35] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing (TIP)*, 22(7):2864–2875, 2013.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, 2017.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proc. ECCV,*. Springer, 2016.

[38] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6043, 2017.

[39] Yuka Ogino, Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Coaxial visible and fir camera system with accurate geometric calibration. In *Thermosense: Thermal Infrared Applications XXXIX*, volume 10214, page 1021415. International Society for Optics and Photonics, 2017.

[40] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 80:143–155, 2018.

[41] 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)*, pages 779–788, 2016.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[43] Thapanapong Rukkanchanunt, Masayuki Tanaka, and Masatoshi Okutomi. Full thermal panorama from a long wavelength infrared and visible camera system. *Journal of Electronic Imaging*, 28(3):1 – 10, 2019.

[44] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Misalignment-robust joint filter for cross-modal image pairs. In *Proceed-*

ings of the IEEE International Conference on Computer Vision (ICCV), pages 3295–3304, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[46] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro. Deep convolutional neural networks for pedestrian detection. *Signal Processing: Image Communication*, 47:482–489, 2016.

[47] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[49] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

[50] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi. Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal iou. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–6. IEEE, 2021.

[51] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi. Multi-modal pedestrian detection with misalignment based on modal-wise regression and multi-modal IoU. *Journal of Electronic Imaging*, 2023.

[52] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi. Multi-modal pedestrian detection via dual-regressor and object-based training for one-stage object detection network. *Electronic Imaging*, 36(17):111–1–111–1, 2024.

[53] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4236–4244, 2017.

[54] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Convolutional channel features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[55] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

[56] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2020.

[57] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 72–80, 2021.

[58] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 443–457, 2016.

[59] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019.

[60] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2021.

[61] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[62] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017.

[63] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[64] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 787–803, 2020.