國立成功大學

工程科學系

博士論文

多模態感測器融合物件偵測技術於自動駕駛之研究

Research on Multi-modal Sensor Fusion for Object Detection in Autonomous Driving

研究生： 鍾昕燁 Student: Sin-Ye Jhong

指導教授： 賴槿峰 Advisor: Chin-Feng Lai

共同指導教授： 夏至賢 Advisor: Chih-Hsien Hsia

中華民國 113 年 11 月

國立成功大學

National Cheng Kung University

博士論文合格證明書

Certificate of Approval for Doctoral Dissertation

題目： 多模態感測器融合物件偵測技術於自動駕駛之研究

Title： Research on Multi-modal Sensor Fusion for Object Detection in Autonomous Driving

研究生：鍾昕燁

本論文業經審查及口試合格特此證明

This is to certify that the Doctoral Dissertation of JHONG,SIN-YE

論文考試委員：

has passed the oral defense under the decision of the following committee members

楊家輝　　詹寶珠

夏至賢　　

陳永耀　　

指導教授 Advisor(s)：　夏至賢

單位主管 Chair/Director：

（單位主管是否簽章授權由各院、系（所、學位學程）自訂）

(The certificate must be signed by the committee members and advisor(s). Each department/graduate institute/degree program can determine whether the chair/director also needs to sign.)

2024/11/18

Research on Multi-modal Sensor Fusion for Object Detection in Autonomous Driving

Sin-Ye Jhong

A Dissertation submitted to the Faculty of the
Department of Engineering Science
In partial fulfillment of the requirements for the Degree of
Doctor of Philosophy
College of Engineering
National Cheng Kung University Tainan, Taiwan
November, 2024

Approved by:

Yung-Yao Chen　　Jaufer Yang

Chih-Hsien Hsia　　Pau-Choo Chung

Chin-Feng Lai

Dissertation Advisor:　　Department Chairman:

Chin-Feng Lai　　Chin-Feng Lee

Chih-Hsien Hsia

摘要

自動駕駛系統在實際應用中面臨諸多挑戰，包括光線不足、惡劣天氣，以及複雜的交通情境。如何克服這些挑戰，並實現安全、便利且穩定的自駕技術，是當前重要的課題。物件偵測作為系統中關鍵的技術之一，在感知環境扮演重要的角色，因為準確的物件偵測是保障行車安全的基礎。然而，目前許多偵測方法依賴單一感測器資訊，導致在動態且不可預測的場景中難以穩定運作。為了解決這一問題，基於多模態感測器融合的物件偵測技術被受關注，期待透過不同感測器資訊的整合來提升系統的感知能力。

在本論文中，我們提出了兩個新的多模態感測器融合框架來解決當前在自駕系統中物體偵測技術的問題。第一個框架專注於二維物件偵測任務，透過融合可見光影像及熱影像感測器資訊，來提升系統在不同環境條件下的表現。第二個框架針對三維物件偵測任務，結合影像及點雲模態中的語意與深度資訊，來強化對遠距離、小型及遮擋物件的偵測能力。此外，本研究建置一台實驗車，除了收集真實行駛資料外，也將技術整合並實現於車載平台，以驗證所提出框架的可行性。最後，藉由多個具代表性的公開資料集及實際應用場景測試的全方面實驗證實，所提出的解決方案相較於現有方法具備更高的偵測性能與及時的執行速度，這使其能更有效應對現實駕駛環境中的各種挑戰，為未來更安全及可靠的自動駕駛系統開發奠定基礎。

關鍵字：自動駕駛系統、物件偵測、異質感測器融合、多模態學習、語意指導、視覺-語言指導

ABSTRACT

Autonomous driving systems face numerous challenges in real-world applications, including low-light conditions, adverse weather, and complex traffic scenarios. Overcoming these challenges to achieve safe, convenient, and stable autonomous driving technology is a critical issue. Object detection, one of the key technologies in autonomous systems, plays a crucial role in environmental perception, as accurate detection is essential for ensuring driving safety. However, many current detection methods rely on single-sensor information, making them less stable in dynamic and unpredictable environments. To address this problem, multi-modal sensor fusion for object detection has gained attention, aiming to improve system perception by integrating data from various sensors.

In this dissertation, we propose two new multi-modal sensor fusion frameworks to address current challenges in object detection for autonomous driving systems. The first framework focuses on 2D object detection, enhancing performance under varying environmental conditions by fusing data from visible and thermal sensors. The second framework is designed for 3D object detection, combining semantic and depth information from both image and point cloud modalities to improve detection of distant, small, and occluded objects. Furthermore, we developed an experimental vehicle that not only collected real-world driving data but also integrated and implemented these technologies on an in-vehicle platform to validate the feasibility of the proposed frameworks. Finally, comprehensive experiments conducted on multiple representative public datasets and real-world scenarios demonstrate that the proposed solutions outperform existing methods in both detection accuracy and real-time execution, providing a foundation for the development of safer and more reliable autonomous driving systems in the future.

Keywords: Autonomous driving systems; Object detection; Heterogeneous sensor fusion;

Multimodal learning; Semantic guidance, Vision-language guidance

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my advisor, Professor Chin-Feng Lai, whose expertise, guidance, and patience have been invaluable throughout my doctoral journey. I would also like to thank my co-advisor, Distinguished Professor Chih-Hsien Hsia from National Ilan University. As my first mentor in research, Dr. Hsia guided me from my undergraduate project through my master's thesis and into my doctoral dissertation. His encouragement and guidance were like a lighthouse, always pointing me in the right direction whenever I felt lost.

I also wish to express my gratitude to Professor Yung-Yao Chen from National Taiwan University of Science and Technology for his support during my R&D alternative military service. Dr. Chen provided research resources and numerous opportunities to collaborate with industry, enabling the findings of this thesis to be transformed into practical applications.

Their valuable feedback and collaborative mentorship helped sharpen my thinking and elevate my work. I feel truly fortunate to have studied under their guidance.

I am also deeply thankful to my committee members: Professor Chin-Feng Lai from National Cheng Kung University, Distinguished Professor Chih-Hsien Hsia from National Ilan University, Distinguished Professor Jar-Ferr Yang, *IEEE Fellow*, from National Cheng Kung University, Distinguished Professor Pau-Choo Chung, *IEEE Fellow*, from National Cheng Kung University, and Professor Yung-Yao Chen from National Taiwan University of Science and Technology. Their insightful comments, challenging questions, and thoughtful suggestions greatly refined my research.

To my colleagues and friends, Hsin-Chun Lin and Si-Yu Lu, thank you for the stimulating discussions, collaborative spirit, and moral support throughout this journey. The exchange of ideas and shared experiences made this process more fulfilling.

My sincere appreciation goes to my family: my parents, Shih-Lin Chung and Pi-Hsiu Chen, and my sister, Yi-Ting Jhong. Their unwavering support, care, and encouragement throughout my PhD years allowed me to focus wholeheartedly on completing my dissertation.

Finally, I thank all those who have directly or indirectly contributed to the completion of this thesis. Your support means more to me than words can express.

Dr. Sin-Ye Jhong

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 Background and Motivation

The rapid growth of urban environments poses critical challenges in managing traffic congestion and ensuring road safety on a global scale. The increasing number of vehicles on the road, driven by expanding cities and rising populations, exacerbates these issues, leading to significant delays, elevated levels of air pollution, and a higher incidence of traffic-related fatalities. According to the World Health Organization, road accidents rank among the leading causes of death worldwide, with millions of fatalities reported annually [1], [2]. A large proportion of these accidents are attributed to human error, which can result from factors such as distraction, fatigue, or impaired judgment. Addressing these challenges is a priority for researchers and policymakers alike, as they strive to enhance road safety and optimize traffic flow. One promising approach to these problems is the development of autonomous driving technologies, which aim to reduce traffic accidents and congestion by eliminating human error [3].

Autonomous vehicles (AVs) offer a potential solution by automating critical driving functions, including navigation, obstacle detection, and lane maintenance. By doing so, they promise to improve the efficiency and safety of traffic networks. However, achieving these goals requires the development of highly advanced perception systems capable of accurately interpreting complex driving environments and making real-time decisions. A key aspect of these systems is object detection, which enables AVs to identify and track other road users, such as vehicles, pedestrians, and cyclists. Inaccurate or unreliable object detection can lead to serious safety incidents, such as collisions, emphasizing the critical need for robust perception systems in AVs [4].

The performance of AV object detection systems is heavily dependent on the sensors used. While visible-spectrum cameras are a popular choice due to their affordability and ability to capture detailed semantic and geometric information [5], as shown in Figure 1. However, they suffer from significant limitations in challenging conditions such as low-light environments or adverse weather (e.g., fog, rain, or snow) [6]. These environmental factors reduce the visibility of objects and hinder detection accuracy. Moreover, visible cameras are inherently limited in their ability to provide depth information, which is crucial for accurate 3D object detection and localization [7].

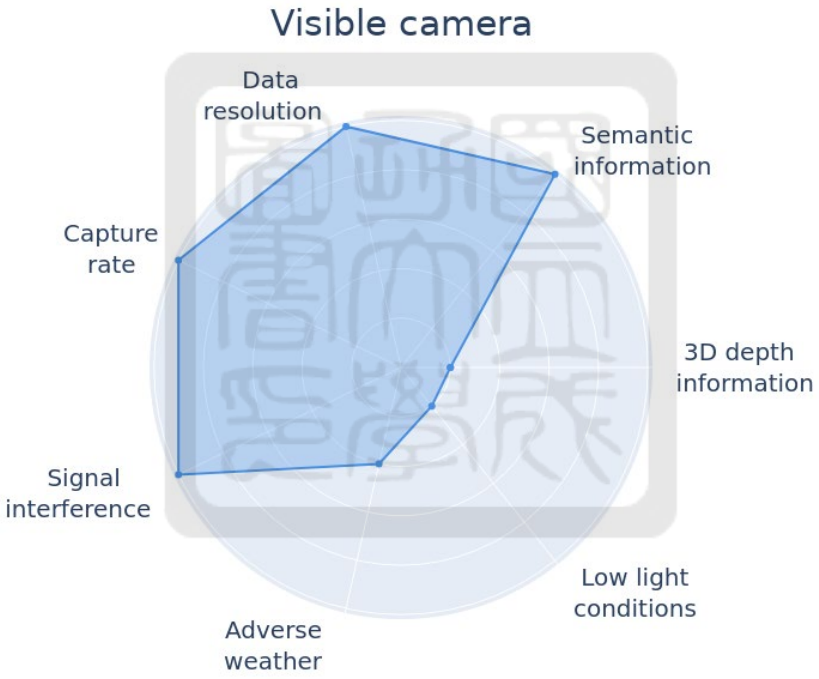


Figure 1. Sensor characteristics of visible camera.

Urban driving environments introduce further complexities. While motorways typically feature well-defined lanes and predictable vehicle movement, urban settings are far more dynamic and heterogeneous, involving various road users such as pedestrians, cyclists, and static objects. Occlusions, where objects block the view of others, further complicate

detection tasks. Objects also vary widely in size and distance, from large vehicles like trucks to smaller entities such as dogs, generating diverse sensor readings [8]. Therefore, effective AV perception systems must meet several stringent criteria: they must be accurate, capable of operating in real-time at high speeds, and robust in the face of environmental and situational variability.

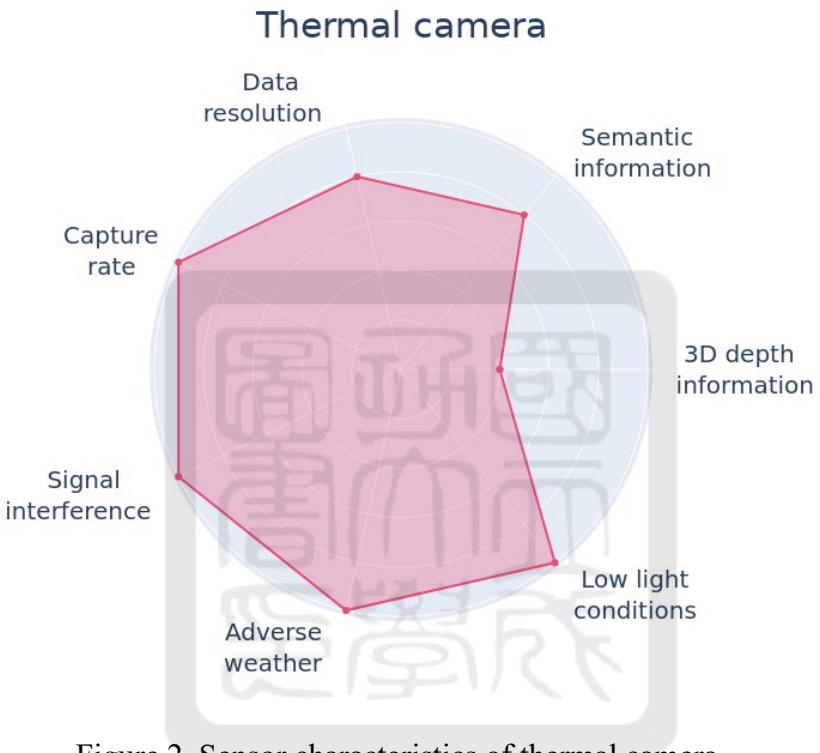


Figure 2. Sensor characteristics of thermal camera.

To overcome the limitations of individual sensors, AV systems often employ a multi-modal approach, integrating data from multiple sensor types to improve overall performance. The goal of multi-modal sensor fusion is to leverage the complementary strengths of each sensor to enhance accuracy and robustness [1], [5]. The most commonly used sensors in AVs, such as thermal cameras, LiDAR, and radar, each provide distinct advantages. For example, as shown in Figure 2, thermal cameras perform reliably in both daytime and nighttime conditions and under various weather scenarios. They are particularly useful for providing

2D semantic information, making them suitable for object detection tasks in low-level applications like ADAS. However, like visible cameras, they lack the ability to capture depth information directly.

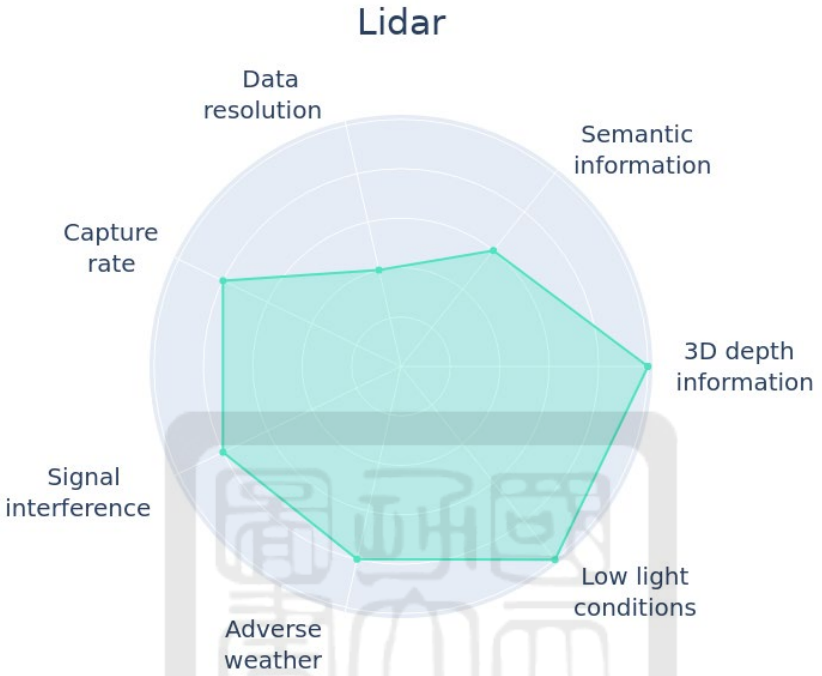


Figure 3. Sensor characteristics of lidar.

LiDAR (Light Detection and Ranging), as depicted in Figure 3, provides precise depth measurements by emitting laser beams and detecting their reflections. LiDAR systems are more resilient to adverse lighting and weather conditions than visible cameras, although they have their limitations. For example, LiDAR struggles with object classification due to its inability to capture fine textures, and its performance deteriorates with increased object distance and occlusion.

Radar (Radio Detection and Ranging), shown in Figure 4, adds another layer of data by providing 3D depth and velocity information. While radar is resistant to lighting and weather issues, its low resolution and susceptibility to interference limit its effectiveness for high-

precision tasks, particularly object classification. Radar's coarse features make it difficult to integrate with other sensor modalities in tasks that require high accuracy.

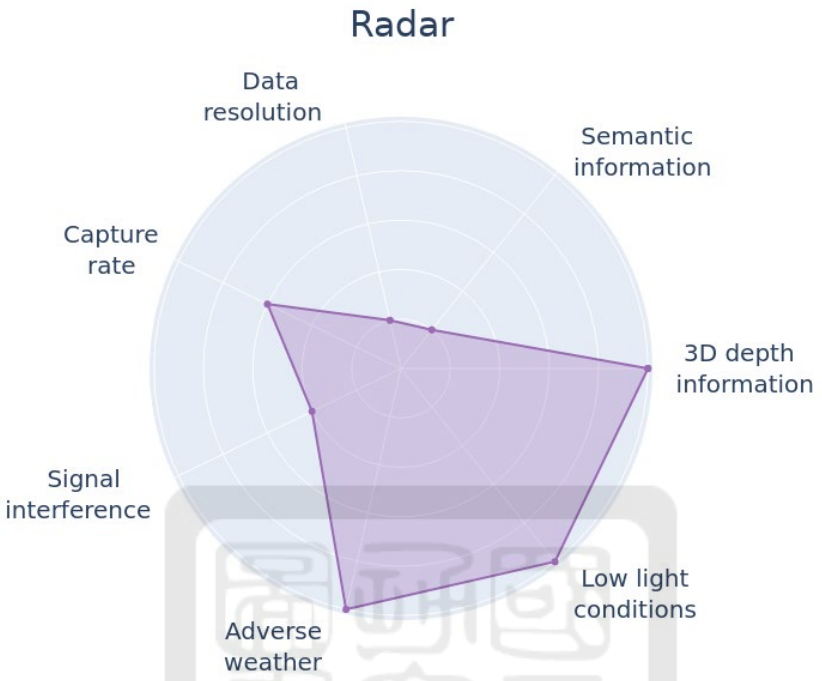
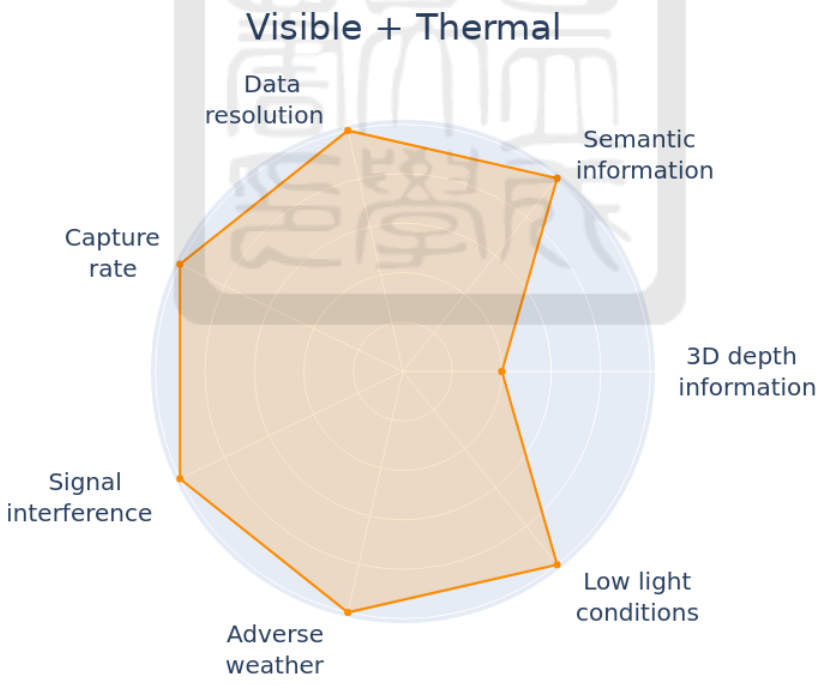


Figure 4. Sensor characteristics of radar.

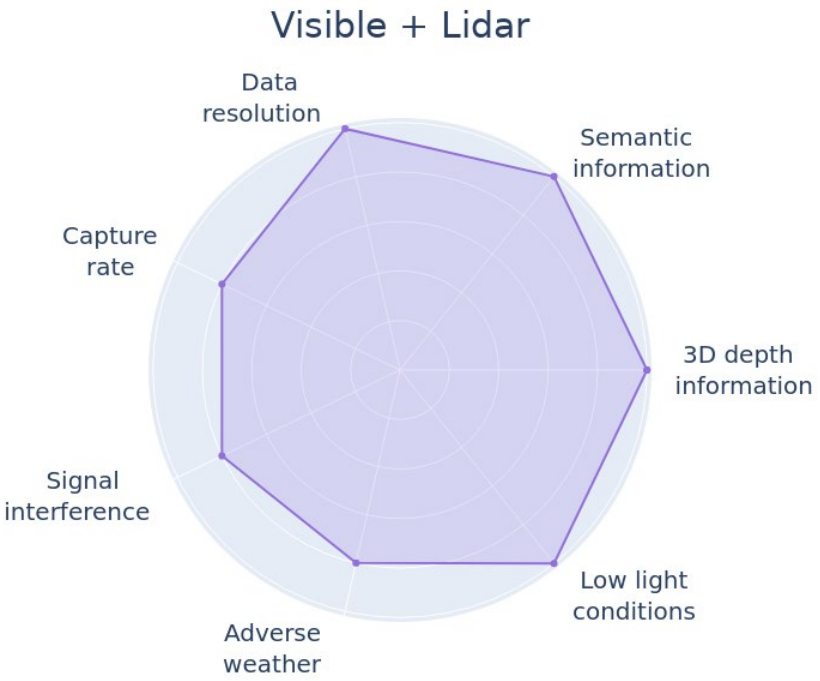
Given the strengths and weaknesses of each sensor modality, designing effective fusion architectures is a major challenge. Visible cameras, when combined with thermal, LiDAR, or radar sensors, offer the potential to overcome the limitations of individual sensors. In the context of AVs, object detection tasks are generally divided into 2D and 3D detection, each of which plays a key role in different levels of decision-making. 2D object detection is essential in low-level ADAS systems (L1 and L2), where real-time responses are crucial for managing traffic [9]. Conversely, 3D object detection is pivotal for higher-level autonomous driving tasks (L3 and beyond), where spatial awareness is required for the accurate positioning and orientation of objects within the environment [1].

As shown in Figure 5, the characteristics of visible-thermal, visible-lidar, and visible-

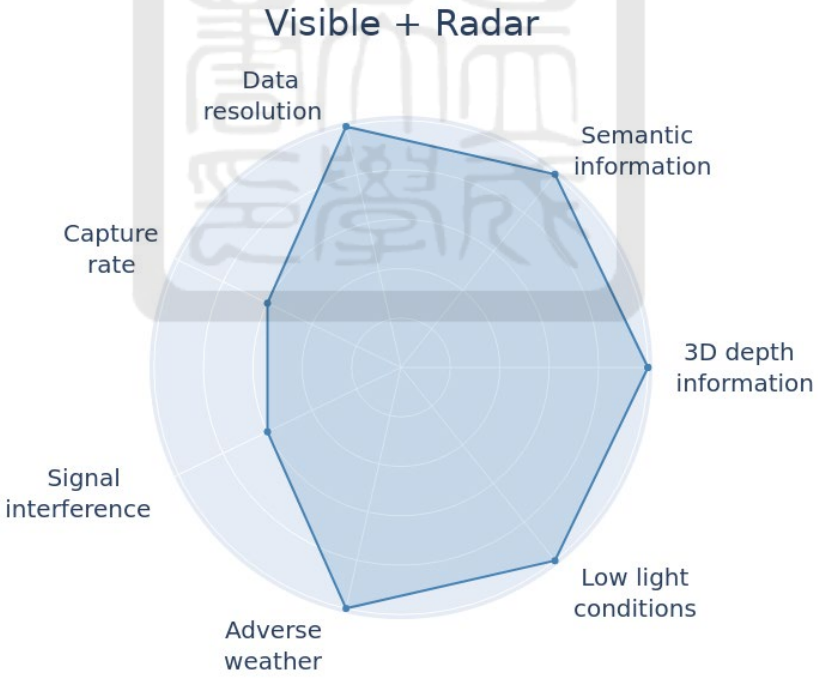
radar sensor fusion systems differ in their applications [3], [5]. Visible-thermal sensor fusion compensates for the limitations of visible cameras in adverse conditions, making it well-suited for 2D detection tasks in ADAS [10]–[12]. In contrast, visible-LiDAR fusion excels in 3D object detection tasks due to LiDAR's ability to provide depth information, making it ideal for higher-level autonomous driving applications [13]–[15]. While visible-LiDAR fusion can also support 2D object detection, its higher cost and the challenges associated with directly integrating point cloud features with image data make it less favorable for this task compared to visible-thermal fusion. In addition, studies have shown that visible-radar sensor fusion is less effective than both visible-thermal and visible-LiDAR fusion for 2D and 3D object detection tasks due to the sparse and coarse nature of radar features, as well as its susceptibility to interference [9], [16].



(a)



(b)



(c)

Figure 5. Comparison of key characteristics in multi-sensor fusion methods. (a) visible-thermal fusion, (b) visible-lidar fusion, and (c) visible-radar fusion.

In light of these challenges and opportunities, this dissertation investigates multi-modal sensor fusion techniques for improving object detection in autonomous driving. Specifically, it focuses on designing and evaluating fusion architectures for 2D object detection using visible-thermal sensors and for 3D object detection using visible-LiDAR sensors. Through this work, the aim is to contribute to more robust and efficient AV perception systems, benefiting both academic research and real-world applications.

1.2 Research Challenges

Despite advancements in sensor fusion techniques, effectively integrating visible-thermal and visible-LiDAR data for 2D and 3D object detection remains a complex challenge. This section examines the inherent obstacles in both visible-thermal-based 2D detection and visible-LiDAR-based 3D detection, focusing on the need for robust fusion strategies in real-world applications.

In the case of visible-thermal sensor fusion for 2D object detection tasks, conventional fusion approaches often fail to deliver consistent performance, especially when applied across diverse environmental conditions [17]. It is therefore essential to develop more advanced fusion methods that fully capitalize on the complementary nature of visible and thermal imagery. Some recent studies have explored the use of illumination-aware networks [10], [11], [18], which dynamically select the most reliable modality based on the prevailing lighting conditions. While these methods show promise, they often struggle in more challenging environmental scenarios, such as rain, fog, or snow, where errors in modality prioritization can occur. Moreover, many models overlook the quality of thermal images, particularly when the temperature contrast between objects and their surroundings is low,

leading to degraded object detection performance.

In response to these limitations, recent research has shifted toward adaptive fusion strategies that employ advanced architectures such as self-attention mechanisms and transformers. These models dynamically integrate visible and thermal information based on learned representations, offering an improvement over static fusion strategies [10]–[12], However, transformer-based approaches tend to rely heavily on deep learning models without leveraging prior knowledge that could further enhance their performance. Additionally, processing raw modality data with transformers can introduce redundant information, which increases computational complexity without necessarily boosting detection accuracy.

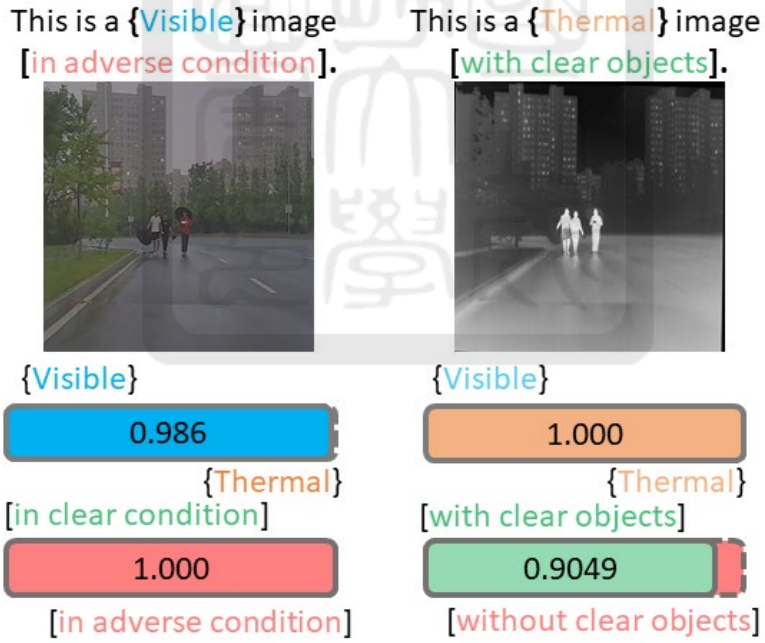


Figure 6. CLIP's assessment of visible and thermal image quality. In the left panel, visible images taken under adverse conditions are shown, and CLIP identifies the modality with 98% confidence before accurately classifying the image as exhibiting "adverse conditions." The right panel displays thermal images captured under

clear conditions, with CLIP identifying the modality type with 100% confidence and correctly classifying the image quality as showing a “clear object.” These results highlight CLIP’s ability to precisely distinguish modality quality in varying scenarios, reinforcing its effectiveness in multispectral object detection.

An ideal fusion strategy would not only align and extract modality-specific features but also optimize the complementarity between them, thereby improving object detection accuracy. Recent advances in vision-language foundation models (VLFMs), such as contrastive language-image pretraining (CLIP) [19], provide a promising solution by capturing rich semantic information across multiple modalities. These models enable the detection system to effectively determine which modality to prioritize under different conditions, minimizing biases and reducing the risk of overreliance on a single modality. As illustrated in Figure 6 CLIP demonstrates robust performance in identifying high-quality data within both visible and thermal images, providing a more contextually aware and reliable fusion process than illumination-aware networks, which may falter in severe weather conditions or when faced with low-quality thermal data.

In the domain of 3D object detection using visible-LiDAR sensor fusion, numerous multimodal approaches have been explored, each presenting distinct challenges. Early methods, such as those in [20], [21], relied on 2D region proposals to construct 3D bounding boxes using frustum techniques. However, the dependence on 2D detection inherently limits their ability to fully harness the geometric information critical for accurate 3D localization. This limitation highlights a fundamental challenge in integrating depth and geometric detail when relying on 2D-driven processes. To address these shortcomings, more recent approaches [22]–[24] have turned to fusing RGB image data with bird's-eye view (BEV) representations generated from voxelized LiDAR point clouds. While this fusion leverages

complementary sensor data, it introduces new challenges, particularly the misalignment of features caused by voxelization and projection, leading to diminished 3D localization accuracy. The difficulty of aligning spatial information between these modalities underscores a key challenge in multimodal sensor fusion.

Point-wise methods [13], [25], [26] have emerged as a potential solution by directly projecting raw point cloud data onto the image plane, enabling the extraction of semantic features from corresponding pixels. This approach preserves fine geometric structures and establishes robust point-pixel relationships. However, it encounters limitations due to the sparse nature of point cloud data, which constrains the amount of semantic information available for feature extraction. Additionally, interactive multimodal learning frameworks [25], [27], which aim to integrate different modalities, are hampered by resolution mismatches between image and point cloud data, resulting in inefficiencies that degrade the overall performance of the fusion process.

Both 2D and 3D object detection tasks demand innovative fusion techniques that address environmental variability, spatial misalignment, and computational efficiency. For 2D detection, recent advances in adaptive transformers and vision-language models offer dynamic prioritization of sensor data, mitigating the limitations of static fusion methods. In 3D detection, resolving the misalignment issues in voxelized data and enhancing point-wise methods to handle sparse inputs will be key to improving accuracy. Ultimately, refining these approaches requires reducing computational complexity and leveraging prior knowledge for more robust, real-world sensor fusion solutions.

1.3 Research Contribution

This dissertation makes significant contributions to the field of multimodal sensor fusion for object detection in autonomous driving by addressing the limitations of existing methodologies and introducing novel frameworks designed to enhance both 2D and 3D detection accuracy. The focus of these contributions is on improving detection performance, particularly under adverse environmental conditions, and providing solutions to the challenges of sparse and noisy sensor data that are common in real-world autonomous driving scenarios.

One of the primary contributions of this dissertation is the development of the vision–language-guided adaptive cross-modal fusion (VL-ACFDet) framework. This innovative approach leverages VLFMs to guide the fusion of visible and thermal sensor data, improving the complementarity between these two modalities. The VL-ACFDet framework introduces two major components: (1) an adaptive cross-contextual attention module, which dynamically aligns and fuses features from both visible and thermal data streams, and (2) a vision–language-guided channel attention transfer module, which utilizes semantic information derived from VLFMs to enhance object detection accuracy. Extensive experiments conducted on the M^3FD dataset [28] and a newly developed dataset demonstrated that VL-ACFDet consistently outperforms current state-of-the-art (SOTA) methods, particularly in challenging weather conditions such as rain, fog, and snow.

For 3D object detection, this dissertation addresses the challenge of sparse point clouds and inefficient point selection through the introduction of the Semantic-guided and Density-aware Fusion (SD-AFDet) framework. This approach improves the fusion of LiDAR and camera data by enriching raw LiDAR points with semantic possibility features from visual data and implementing a novel point sampling algorithm to optimize key feature selection

in the 3D backbone. Furthermore, SD-AFDet includes a density-aware detection head, which adjusts the position of artificial points based on point cloud density, improving the aggregation of geometric and semantic features. Experiments on the KITTI dataset [29] and the newly created dataset show that SD-AFDet excels in detecting distant and small objects, which are often difficult to capture in sparse point clouds.

These contributions provide distinct advancements in sensor fusion techniques, each tailored for different applications within autonomous driving, enhancing both 2D and 3D object detection in specific operational contexts. The VL-ACFDet framework for 2D object detection and the SD-AFDet framework for 3D object detection work synergistically to provide a comprehensive solution for real-time perception systems in autonomous vehicles. By combining innovative feature alignment strategies, knowledge distillation [30] techniques, and advanced sampling mechanisms, this work not only addresses current limitations in sensor fusion but also establishes a foundation for future research, ultimately contributing to the development of safer and more reliable autonomous driving technologies.

CHAPTER 2. RELATED WORKS

2.1 Visible-based 2D/3D Object Detection

Recent advancements in visible-spectrum object detection have primarily led to the development of two-stage and one-stage models for 2D detection tasks. Two-stage models, such as the widely recognized R-CNN family, prioritize detection accuracy by refining bounding box proposals through multiple stages. A prominent example is Faster R-CNN [31], which introduced the region proposal network (RPN) to enhance the generation of region proposals, delivering strong performance in visible-spectrum images. Variants of Faster R-CNN have also been adapted to other modalities, such as thermal imaging, for specialized tasks like nighttime pedestrian detection. However, the computational demands of two-stage models pose a challenge for real-time applications in autonomous driving, where low latency is crucial.

In contrast, one-stage models like the YOLO (You Only Look Once) series [32] focus on striking a balance between detection speed and accuracy, making them more suitable for real-time applications. YOLOv5 [33], for example, effectively balances processing speed and precision, rendering it a popular choice for both single-spectral and multispectral object detection tasks [10], [12], [34], [35]. Nevertheless, visible-spectrum models remain limited under challenging environmental conditions, such as low light or adverse weather. Thermal-based models, while advantageous in certain conditions, face difficulties when the temperature contrast between foreground and background objects is minimal, leading to reduced detection performance.

As the demand for greater environmental awareness in higher-level autonomous driving applications grows, researchers have increasingly turned their attention to 3D object

detection using visible cameras. A key drawback of visible cameras is their inability to capture depth information directly. To address this, various methods have been developed. For example, Reading et al. [36] generated pseudo point clouds from dense depth maps, leveraging sparse LiDAR data to create a more complete depth representation. Other techniques [37], [38] leverage the spatial correlations between 2D image projections and 3D structures to improve object detection accuracy. Furthermore, researchers such as [39] have introduced methods that augment feature maps by predicting depth distributions, improving object localization in the absence of direct depth sensing. However, these image-based approaches still fall short when compared to multimodal fusion techniques, which integrate data from multiple sensors and provide superior accuracy in 3D localization tasks.

2.2 Visible-thermal-based 2D Object Detection

Visible-thermal-based 2D object detection, commonly referred to as multispectral object detection, aims to combine data from different spectral bands to improve detection accuracy and robustness under various environmental conditions. By fusing visible and thermal data, these methods provide enhanced performance, particularly in scenarios where either modality alone may be insufficient. Depending on the integration stage, fusion methods are typically classified as early, mid, or late fusion, or they are grouped based on specific fusion strategies, such as addition, averaging, or concatenation.

Early fusion strategies integrate visible and thermal data at the input level, enabling simultaneous multimodal processing. However, this approach can introduce noise and redundancy, degrading the quality of the feature representations [40]. Late fusion, also known as decision-level fusion, occurs after feature extraction and maintains the independence of each modality [41]. However, this strategy tends to be computationally

expensive, as it requires fully developed dual-stream networks. For example, MSCoTDet [40] uses a late fusion approach, involving the training of two separate detection networks and leveraging large language models for decision-making, leading to a highly resource-intensive process.

Mid-fusion approaches offer a balance by integrating modality-specific features during the feature extraction phase. These approaches often use dynamic fusion weights to enable more adaptive combinations of features from both modalities. Static fusion methods, such as addition, averaging, or concatenation, lack the flexibility to adjust to varying environmental conditions [42], [43]. Illumination-aware networks have also been introduced to modulate the fusion process based on lighting conditions [18], [44]–[46]. For instance, MBNet [18] incorporates a subnetwork that estimates visible image illumination and adjusts the fusion accordingly. However, relying solely on illumination-based mechanisms proves insufficient, as many other factors influence the reliability of each modality.

Recent work has shifted toward adaptive fusion mechanisms that leverage advanced architectures, such as convolutional networks or transformers, to model dynamic interactions between the two modalities [10]–[12], [18], [43]. For example, CFR [18] enhances multispectral features using a sequence of simple fusion modules based on convolutional layers, although it does not incorporate prior knowledge. CFT [12] applies self-attention mechanisms to learn intermodal correspondences adaptively, but it lacks specialized fusion modules, limiting the full exploitation of multispectral data and reducing the fusion efficiency.

2.3 Visible-lidar-based 3D Object Detection

Advances in 3D object detection have increasingly relied on multimodal fusion,

particularly integrating data from visible cameras and LiDAR sensors, which play a crucial role in autonomous driving. Visible-LiDAR fusion methods are commonly categorized into cascading approaches, multi-view fusion techniques, and point-wise fusion strategies, each with its unique strengths and challenges.

Cascading methods process 2D and 3D detections in separate stages. For instance, [20] projects 2D detections into frustum spaces to generate 3D bounding boxes, which are subsequently refined. Techniques like [21] improve upon this by segmenting the frustum space and extracting region-specific features. Despite these refinements, cascading approaches are inherently limited by the performance of their independent 2D and 3D detectors, which constrains their ability to fully harness the benefits of multimodal fusion.

Multi-view fusion approaches aim to integrate features from different perspectives—such as range view (RV), bird's-eye view (BEV), and camera view (CV)—to provide a more comprehensive understanding of the scene. Methods like [22] and [23] enhance detection accuracy by combining region of interest (ROI) features from BEV and CV in a region proposal network (RPN). More advanced techniques, such as [24], employ attention-based dynamic feature extraction architectures to integrate information from multiple views and raw point clouds. Additionally, [47] introduces a gating mechanism to reduce noise in the BEV and CV feature streams. However, these methods are still hindered by significant feature misalignment across different views, which affects the overall accuracy of the fusion process.

Point-wise fusion methods directly project raw LiDAR point cloud data onto the 2D image plane, facilitating the fusion of geometric and semantic features. This method enhances the alignment between 2D image data and 3D point cloud features, improving detection accuracy. For example, [13] and [26] use 2D segmentation to extract semantic features that augment the LiDAR point cloud data, while LI-fusion [27] refines this process

by employing a cross-learning mechanism to capture both semantic and geometric features. In addition, [25] introduces a feature-alignment framework that reduces the loss of high-dimensional information when projecting 3D point clouds into the 2D space. However, the low resolution of LiDAR point clouds remains a critical limitation, restricting the fusion of rich semantic information and ultimately constraining detection performance.

Despite the significant progress in visible-LiDAR-based 3D object detection, challenges such as feature misalignment, limited resolution in LiDAR data, and the computational complexity of multimodal fusion continue to hinder the effectiveness of these methods. Addressing these challenges will require the development of more robust fusion techniques that can better align spatial information, reduce noise, and optimize the integration of geometric and semantic features across modalities, paving the way for more accurate and reliable 3D object detection in autonomous systems.

CHAPTER 3. PROPOSED VL-ACFDET FRAMEWORK FOR 2D OBJECT DETECTION

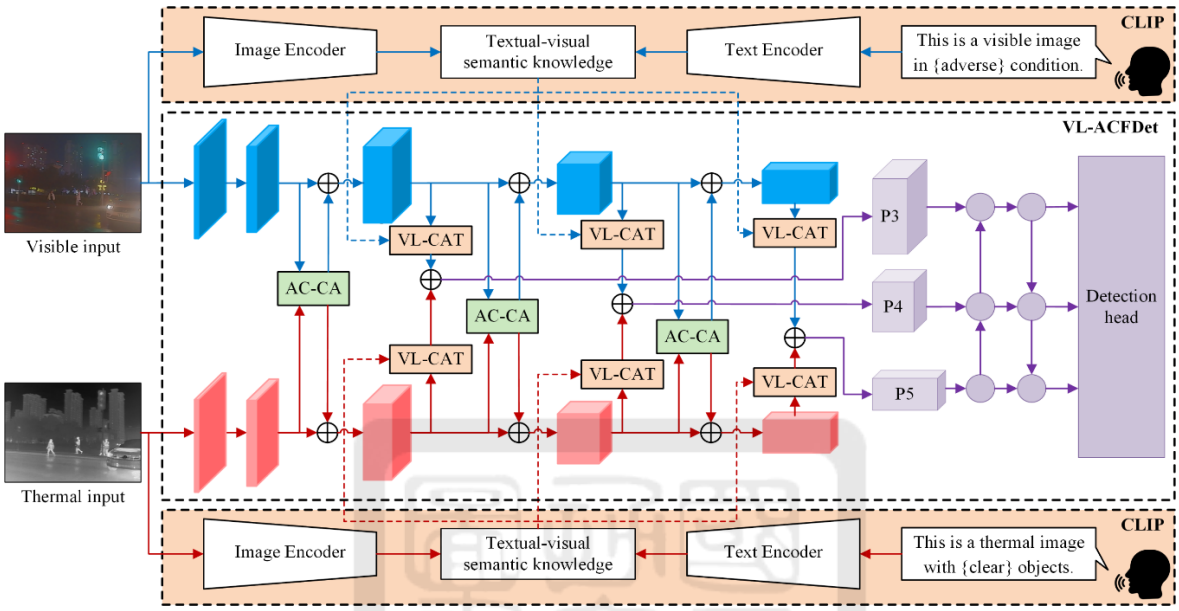


Figure 7. Overview of the VL-ACFDet Framework. The framework features a dual-stream architecture that separately processes visible and thermal inputs through distinct branches. Each branch is augmented with semantic knowledge from CLIP. The VL-CAT and AC-CA modules selectively fuse and enhance features from both modalities, significantly boosting object detection accuracy. These enriched features are then processed through multiple levels of a feature pyramid network, leading to the final detection outputs. Importantly, the CLIP-based enhancements are applied only during training, ensuring no additional computational overhead during inference.

The proposed VL-ACFDet framework, illustrated in Figure 7, enhances 2D object detection by leveraging the complementary strengths of visible and thermal data through a mid-fusion strategy within an extended dual-stream architecture based on YOLOv5. This

architecture facilitates multi-level adaptive cross-modal fusion during the feature extraction phase, leading to improved object detection accuracy across various environmental conditions. The VL-ACFDet framework introduces two novel modules: the Adaptive Cross-Contextual Attention (AC-CA) module and the Vision-Language Channel Attention Transfer (VL-CAT) module, both of which are integrated into the backbone network to optimize feature fusion.

The AC-CA module addresses the inherent heterogeneity between visible and thermal features by selectively extracting the most relevant information while filtering out redundancies. This module capitalizes on the contextual information from each modality, enhancing the discriminative properties of the fused features. The VL-CAT module, on the other hand, utilizes knowledge distillation from the CLIP model to guide the network's attention, focusing on key features and advantageous modality information during the fusion process. Both modules are integrated into the final layers of the visible and thermal branches of the backbone network, and the enhanced features are subsequently processed by a feature pyramid network to generate the final object detection results. The following sections delve into the detailed operation of these modules.

3.1 Adaptive Cross-contextual Attention Module

Traditional multispectral fusion methods often rely on the comprehensive extraction of multimodal features, typically using robust Transformer architectures to capture complementary information. However, these methods frequently introduce redundant information due to the differences between visible and thermal sensors, leading to increased computational costs. Additionally, modality biases can hinder the use of fine-grained semantic details essential for accurate object classification. The AC-CA module, shown in

Figure 8, addresses these challenges through three key components: the Similarity Feature Selection (SFS) block, the Contextual Feature Extractor (CFE) block, and the Cross-Modal Attention Fusion (CAF) block.

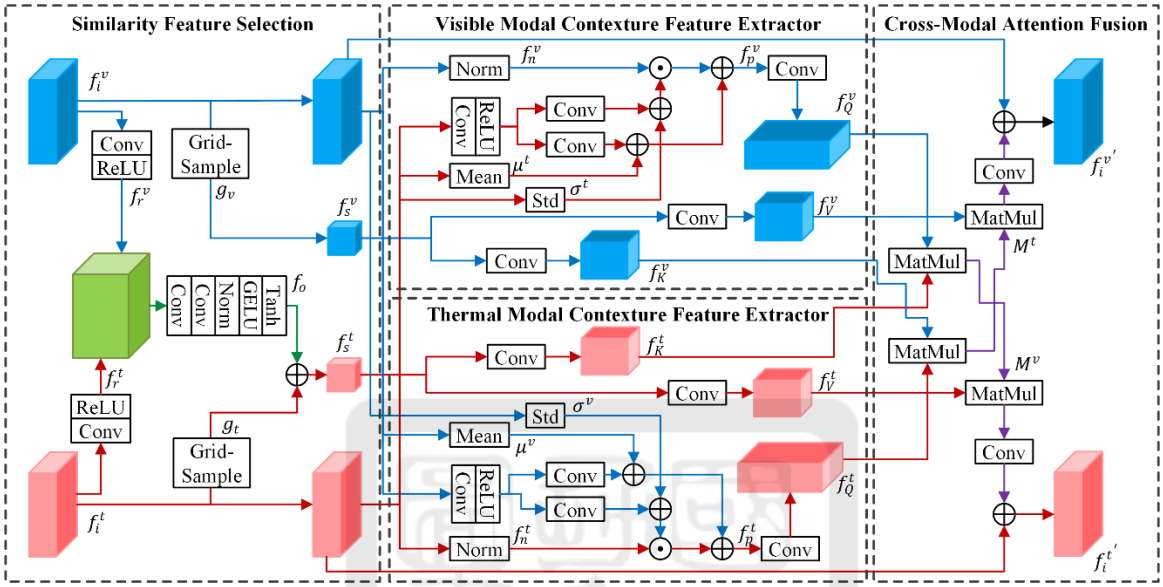


Figure 8. The AC-CA module. This module comprises three essential components: the Similarity Feature Selection (SFS) block, the Contextual Feature Extractor (CFE) block, and the Cross-Modal Attention Fusion (CAF) block. The SFS block ensures precise alignment of visible and thermal features by predicting spatial shifts and applying bicubic interpolation, facilitating the selection of the most pertinent features. The CFE block then normalizes these aligned features to reduce modality biases and re-projects them, optimizing their effectiveness for cross-modal fusion. Finally, the CAF block integrates features from both modalities using cross-modal similarity matrices, resulting in a unified and robust feature representation.

1) SFS block: In multispectral object detection, raw modality features often contain significant redundancies, which complicate the fusion process and inflate computational

demands, particularly in Transformer-based architectures. Inspired by [18], the SFS block mitigates this issue by predicting shifts between visible and thermal features, selecting only the most relevant and similar features, and reducing the influence of misaligned redundant features. Let $f_i^v \in R^{C\times H\times W}$ (visible modality) and $f_i^t \in R^{C\times H\times W}$ (thermal modality) denote the input feature maps.. These feature maps are transformed through convolutional layers and ReLU activation, resulting in refined feature representations $f_r^v \in R^{C\times H\times W}$ and $f_r^t \in R^{C\times H\times W}$. The transformed features are concatenated along the channel dimension and compressed back to their original size using a 1×1 convolution operation, producing the compressed feature $f_c \in R^{C\times H\times W}$. The deviation network D is then employed to predict the offset $f_o \in R^{2\times H\times W}$, reflecting their spatial correspondence between the two modalities. To ensure stability and predictability of the offsets, we apply a hyperbolic tangent function, defined as:

$$f_o = \mathrm{Tanh}(D(\mathrm{Concat}(f_r^v;\ f_r^t))). \tag{1}$$

Reference grids $g_v \in R^{C\times\frac{H}{s}\times\frac{W}{s}}$ and $g_t \in R^{C\times\frac{H}{s}\times\frac{W}{s}}$ are generated for the visible and thermal branches, respectively, where s is the stride factor. The thermal grid is aligned with the visible grid using the predicted offsets, and aligned features are obtained through bicubic interpolation:

$$f_s^t = \varphi_{\mathrm{Bic}}(f_i^t, g_t + f_o) \in R^{C\times\frac{H}{s}\times\frac{W}{s}}, \tag{2}$$

$$f_s^v = \varphi_{\mathrm{Bic}}(f_i^v, g_v) \in R^{C\times\frac{H}{s}\times\frac{W}{s}} \tag{3}$$

where $\varphi_{\mathrm{Bic}}(\cdot\,;\cdot)$ represents the bicubic interpolation sampling function.

2) CFE block: After aligning the features, the CFE block extracts the contextual information required for effective cross-modal fusion. Due to significant differences in modality biases, directly interacting features using traditional self-attention mechanisms often results in suboptimal performance. Inspired by [48]–[50], we apply a modality normalization technique to align feature distributions across the two modalities. Let μ_v, σ_v, μ_t, and σ_t, denote the means and standard deviations for visible and thermal modality features, respectively. The features are normalized as:

$$f_n^t = \frac{f_i^t - \mu^v}{\sigma^v},\ f_n^v = \frac{f_i^v - \mu^t}{\sigma^t}, \tag{4}$$

where f_n^t and $f_n^v \in R^{C\times H\times W}$ are the normalized modality features. These features are re-projected onto their respective distributions using three 3×3 convolutional layers, producing context features suitable for cross-modal fusion. The query, key, and value features are computed as follows:

$$f_Q^t = W_Q^t * f_n^t, f_K^t = W_K^t * f_s^t, f_V^t = W_V^t * f_s^t, \tag{5}$$

$$f_Q^v = W_Q^v * f_n^v, f_K^v = W_K^v * f_s^v, f_V^v = W_V^v * f_s^v, \tag{6}$$

where W_Q, W_K, and W_V denote the convolution operations for generating the query, key, and value features.

3) CFE block: The cross-modal attention similarity matrices are computed to integrate features from visible and thermal modalities. The query and key features from different

modalities are used to calculate the cross-modal similarity matrices M^t and M^v, defined as:

$$M^t = \text{softmax}\left(\frac{\text{MatMul}\left(f_Q^t, T(f_K^v)\right)}{\sqrt{d}}\right) \in R^{HW \times HW}, \tag{7}$$

$$M^v = \text{softmax}\left(\frac{\text{MatMul}\left(f_Q^v, T(f_K^t)\right)}{\sqrt{d}}\right) \in R^{HW \times HW}, \tag{8}$$

where $T(\cdot)$ denotes matrix transposition, $\text{MatMul}(\cdot,\cdot)$ represents matrix multiplication, and d is the dimensionality of the query/key vectors. These matrices capture cross-modal relationships, and the final fused features are computed by combining input features with the output of matrix multiplication between the similarity matrices and the value features:

$$f_i^{t'} = f_i^t + \text{conv}\left(\text{MatMul}(M^v, f_V^t)\right), \tag{9}$$

$$f_i^{v'} = f_i^v + \text{conv}\left(\text{MatMul}(M^t, f_V^v)\right). \tag{10}$$

The AC-CA module dynamically models and facilitates the exchange of contextual information between visible and thermal modalities, reducing cross-modal biases and improving the quality of the fused features.

3.2 Vision–language-guided Channel Attention Transfer Module

Traditional multispectral detection methods often rely on illumination-aware networks and basic parameters to evaluate the relative importance of visible and thermal modalities. However, these approaches are susceptible to environmental factors such as rain, fog, or snow, which can lead to incorrect prioritization of low-value features, thereby negatively

affecting detection performance. To address this issue, we propose the VL-CAT module, as illustrated in Figure 9. This module enhances the multispectral detection model by leveraging rich semantic information from VLFMs, such as CLIP, to transfer both textual and visual semantic features into the detection pipeline. This process reduces the influence of low-semantic features while emphasizing high-semantic ones, improving detection robustness across varying conditions.

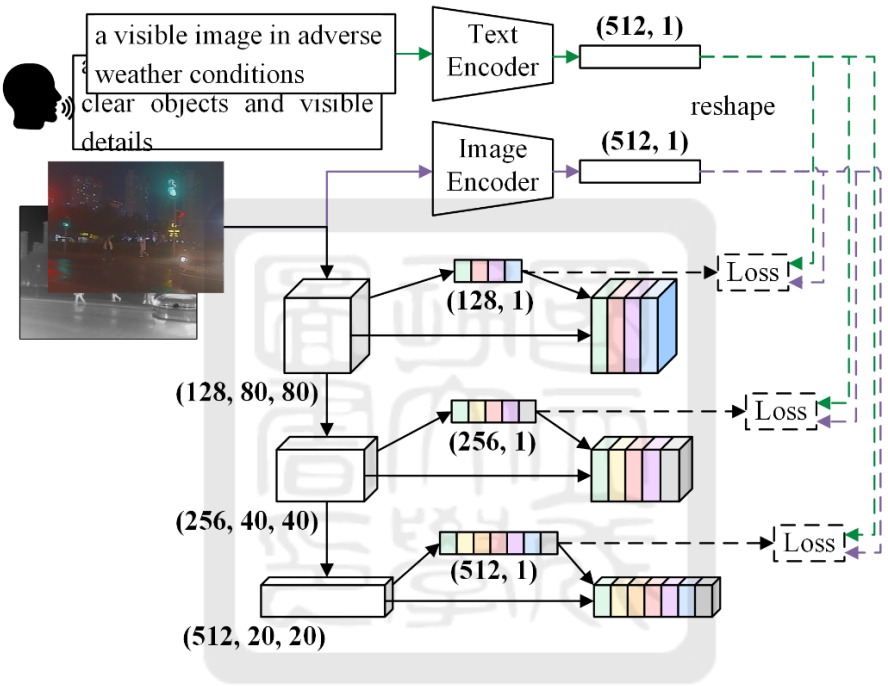


Figure 9. The VL-CAT module. This module utilizes CLIP's text and image encoders to extract rich semantic features from multispectral images. These semantic features are essential in guiding the generation of channel-wise attention features within the backbone network. The module then refines these attention features by employing L2-norm losses, comparing them against the extracted semantic features to highlight the most critical details. Once refined, the enhanced features are reintegrated into the network, allowing the model to better focus on high-semantic information, ultimately enhancing multispectral object detection performance.

The VL-CAT module first processes multispectral images through CLIP, where the image encoder extracts visual semantic features $I_s^t \in R^{C\times 1}$ and $I_s^v \in R^{C\times 1}$ for the thermal and visible images, respectively. Additionally, prompts tailored to modality quality assessment are used to extract textual semantic features T_s^t and T_s^v, which provide context, such as "a visible image in adverse weather conditions" or "a thermal image without clear objects and visible details."

Within the backbone of the detection model, additional branches are introduced in the last three layers. Each branch consists of a Max Pooling operation and fully connected layers to compute channel-wise attention [51]–[53]. For a feature map $f \in R^{C\times H\times W}$, the channel-wise attention feature map f_c is computed as:

$$f_c = \sigma\Big(\mathrm{MLP}\big(\mathrm{MaxPool}(f)\big)\Big) \in R^{C\times 1}, \tag{11}$$

where σ represents the sigmoid activation function, MLP denotes the fully connected layer operation, and MaxPool is the max pooling operation. This process generates three channel-wise attention features for the visible modality $f_{c_l}^v$, $f_{c_m}^v$, $f_{c_s}^v$ and three for the thermal modality $f_{c_l}^t$, $f_{c_m}^t$, $f_{c_s}^t$. Inspired by knowledge distillation, we introduce L2-norm loss functions to facilitate the transfer of semantic features from CLIP to the channel-wise attention features. For a given visible channel-wise attention feature f_c^v, the textual and visual feature transfer losses are defined as:

$$L_T = \| f_c^v - T_s^t \|_2^2, \tag{12}$$

$$L_I = \| f_c^v - I_s^t \|_2^2. \tag{13}$$

These loss functions measure the distance between the extracted channel-wise attention features and the corresponding textual and visual semantic features from CLIP. The total textual and visual feature transfer losses are averaged and combined to produce a final loss term that encourages the alignment of semantic information with channel-wise attention. The learned features, now rich in semantic information, are re-weighted into the backbone via attention maps to minimize the impact of low-semantic features and enhance high-semantic ones, improving the overall detection accuracy.

An important advantage of this approach is that CLIP is only used during the training phase. All layers of the CLIP model are frozen, and no additional data or computational burden is introduced during inference, ensuring that the system remains efficient in real-time applications.

3.3 Joint Learning of Detection and Transfer Losses

During training, the proposed VL-ACFDet model follows the procedure described in CFT [12], designed as an end-to-end 2D object detection framework. The detection loss function at each stage includes classification loss L_c, objectness confidence loss L_o, and Complete Intersection over Union (CIOU) [55] loss L_{CIOU} for bounding box regression. The classification loss L_{c} is computed as:

$$L_{\mathrm{c}} = \sum_{i,j} I_{ij}^{o} \sum_{i,j} P_c(i,j,c) \log\left(\widehat{P_c}(i,j,c)\right), \tag{14}$$

where i and j index the grid cells and bounding boxes, respectively. The indicator function I_{ij}^{o} equals 1 when an object is present in the ith grid cell and is predicted by the jth bounding box; otherwise, it equals 0. $P_c(i,j,c)$ denotes the predicted probability that the jth bounding

box in the ith grid cell belongs to class c. This loss function measures quantifies the model's accuracy in classifying objects within the predicted bounding boxes. The objectness confidence loss L_o is defined as:

$$L_o = \lambda_o \sum_{i,j} I_{ij}^{o} \left(P_o(i,j) - \widehat{P_o(i,j)}\right)^2 + \lambda_n \sum_{i,j} I_{ij}^{n} \left(P_o(i,j) - \widehat{P_o(i,j)}\right)^2, \quad (15)$$

where λ_o and λ_n are balancing factors for losses associated with cells containing objects and those without. Here, we set λ_o=1 and λ_n=0.5. Finally, we aggregate the detection loss components with the textual and visual feature transfer losses L_T and L_I to form the overall detection loss $L_{\mathrm{VL-ACFDet}}$, expressed as:

$$L_{\mathrm{VL-ACFDet}} = \lambda_o \cdot L_o + \lambda_c \cdot L_c + \lambda_{\mathrm{CIOU}} \cdot L_{\mathrm{CIOU}} + \lambda_T \cdot L_T + \cdot \lambda_I L_I, \quad (16)$$

where λ_o, λ_c, λ_{CIOU}, λ_T, and λ_I are weight factors. By employing joint learning, we optimize both the primary detection task and the auxiliary knowledge transfer task simultaneously, thereby enhancing the model's ability to perform complementary fusion and improving overall detection performance.

CHAPTER 4. PROPOSED SD-AFDET FRAMEWORK FOR 3D OBJECT DETECTION

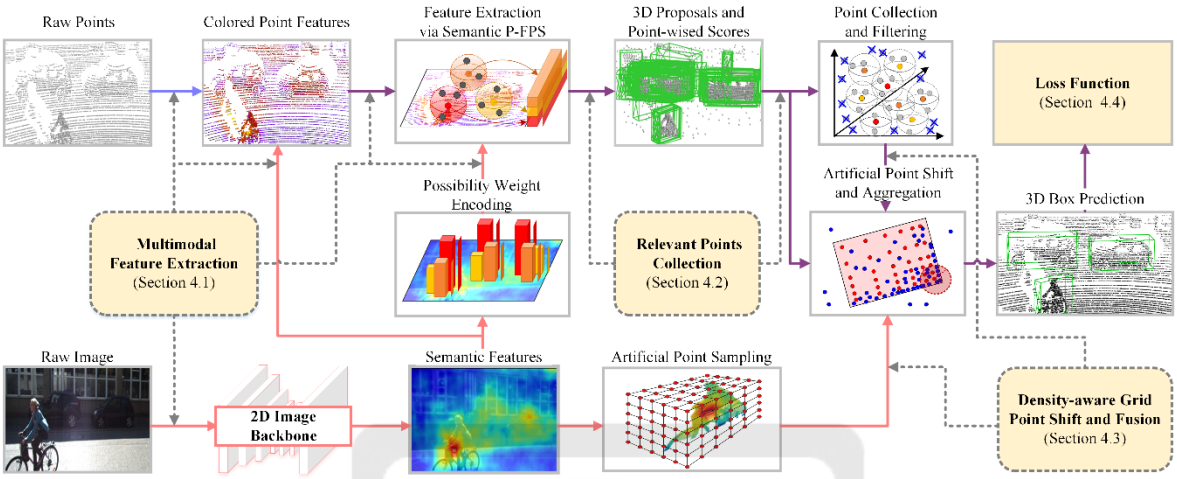


Figure 10. Illustration of SD-AFDet framework. Different color-coded lines representing the flow of point data (blue), image data (red), and fusion features (purple). Each phase of the model contributes to the goal of enhancing 3D object detection by systematically processing the input data. In the initial phase (Section 4.1), the raw point cloud is augmented with semantic information obtained from a 2D segmentation network, addressing issues like shape ambiguity. This semantic information is encoded into possibility weights. To enhance the feature extraction capabilities of the 3D backbone, we employ P-FPS (Point-based Farthest Point Sampling), which improves the selection of points used in subsequent steps. Moving to the next stage (Section 4.2), predicted bounding boxes and the results from P-FPS are used, guided by the predicted scores, to collect more foreground points. This ensures that relevant points are retained while maintaining point diversity, setting the stage for more accurate refinement. In the subsequent stage (Section 4.3), artificial points are generated for the purpose of semantic feature sampling. Their spatial positions are shifted based on the local point density within the proposals, allowing the framework to maximize the

aggregation of useful point features from the scene. Finally, in the last phase (Section 4.4), the model computes the total loss, optimizing its performance through end-to-end training.

The SD-AFDet framework (illustrated in Figure 10) is designed to address two persistent challenges in autonomous driving: inefficient point sampling and sparse point clouds, both of which limit the accurate detection of small and distant objects. To mitigate these limitations, SD-AFDet introduces two key innovations that enhance both point cloud augmentation and point selection.

The first innovation is the 2D segmentation network, which creates a pixel-wise feature map encoding the probability of object presence. This feature map serves a dual purpose: filtering out noise in raw images and enriching point cloud data. By projecting points from 3D space onto the 2D feature map, the network extracts semantic information that enhances object differentiation. This method is particularly effective in reducing false positives, especially when objects with similar shapes are present.

The second core innovation is the P-FPS algorithm, which improves traditional point sampling techniques. Conventional methods often discard significant foreground points and retain less relevant background points, hampering the detection of small objects. In contrast, P-FPS uses a guided weighting mechanism, informed by geometric and semantic features, to prioritize the retention of relevant foreground points while preserving overall point cloud diversity. This innovation is critical for enhancing the detection of smaller objects, which are typically underrepresented in raw point clouds.

In addition, SD-AFDet employs a density-aware artificial point shift and fusion (DAPSF) to further improve object detection. DAPSF generates artificial points within 3D region proposals, augmenting the sparse point cloud and acting as proxies for sampling semantic features. This process enables the model to better capture the geometric structure

of the scene, significantly improving detection accuracy, especially for distant and tiny objects in real-world autonomous driving scenarios. The subsequent sections provide a detailed breakdown of each component of the SD-AFDet framework, explaining how these innovations collectively enhance 3D object detection performance.

4.1 Multi-modality Feature Extraction

1) Point-wise Augmentation: The SD-AFDet framework (illustrated in Figure 10) utilizes two parallel pipelines to perform multi-modality feature extraction. The image pipeline applies a 2D segmentation model [56] followed by a softmax function to generate a semantic feature map $S \in \mathbb{R}^{H \times W \times C}$, where each class (e.g., cars ρ^c, pedestrians ρ^p, cyclists ρ^r, and the background ρ^b) is represented by its respective likelihood. Concurrently, the geometry pipeline projects raw LiDAR points onto the image space to extract corresponding semantic information from the feature map. This process is mathematically represented as:

$$s\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = KR_t P, \tag{17}$$

where K and R_t represent camera parameters, and $P = \{p_i = (x_i, y_i, z_i)\}_{i=1\sim N} \in \mathbb{R}^{N \times 3}$ is the raw point cloud set. This projection results in an augmented point set $AP = \{x_i, y_i, z_i, \rho^c, \rho^p, \rho^r, \rho^b\}_{i=1\sim N} \in \mathbb{R}^{N \times 7}$, where raw point coordinates are concatenated with point-wise possibilities from the possibility map. These enriched points contain valuable semantic information, enabling the model to differentiate between objects with similar shapes and minimize false detections. This process is detailed in Algorithm 1.

Algorithm 1: Multi-modal Feature Extraction Procedure

Input:
3D coordinates of the point cloud $P = \{p_i \in \mathbb{R}^{N\times 3} \mid i = 1,\ldots, N\}$
RGB Image $I \in \mathbb{R}^{H\times W\times 4}$
Sample point count M

Output:
Indices of the selected points $F_{index} = \{f_i \in \mathbb{R}^{M} \mid i = 1,\ldots, M\}$

Initialization:
F_{index}, which will store the final selected point indices, is initialized as an empty set.
tmp, an array of length N, , is initialized with $+\infty$ to track the minimum distance between selected and unselected points.
$record$ an array of length N, marks is initialized to zero to track selected points.

$S = SoftMax(SEGMENT(I)) \in \mathbb{R}^{H\times W\times 4}$
$P^{xy} =$ PROJECT$(P, K, R_t) \in \mathbb{R}^{N\times 2}$
$P_s = S[P^{xy}[:,0], P^{xy}[:,1]] \in \mathbb{R}^{N\times 4}$
$B = \lambda_c \rho^c + \lambda_p \rho^p + \lambda_r \rho^r \in \mathbb{R}^{N\times 1}$
for $i = 1$ *to* M:
  if $i = 1$
    f_i = **argmax**(B)
  else
    $Dis = tmp * B^{\omega}$
    f_i = **argmax**$(record = 0)$
  end if
  $F_{index} \leftarrow F_{index} \cup f_i$
  $record[f_i] \leftarrow record \cup f_i$
  for $j = 1$ *to* N
    $dis = ||p_i - p_j||$
    $tmp[j] = min(tmp[j], dis)$
  end for
end for
return $\boldsymbol{F_{index}}$

2) Semantic-guided Point Sample: Once augmented, the points are processed by the 3D backbone for feature extraction. Traditional downsampling techniques, such as Farthest Point Sampling (FPS), aim to achieve an even distribution of points. However, FPS often oversamples background points, which make up the majority of the point cloud, while

undersampling critical foreground objects. This imbalance, coupled with outlier points, significantly hampers the detection accuracy of small and distant objects [59].

To mitigate the issue of reduced model generalization due to insufficient point diversity, a straightforward approach is to predict confidence scores for each point and select the top k points with the highest scores. However, this method often compromises point diversity, leading to suboptimal generalization during testing. To overcome this limitation, we propose the P-FPS algorithm, which increases the likelihood of sampling foreground points while preserving diverse and important foreground information..

In the P-FPS algorithm, possibility values for different object classes, such as cars ρ^c, pedestrians ρ^p, and cyclists ρ^r , are encoded into a possibility weight matrix $B \in \mathbb{R}^{N\times 1}$. To ensure that the sampling process prioritizes foreground points without losing diversity, the algorithm retains the basic structure of the traditional FPS method. However, instead of relying solely on distance metrics, the possibility weight $B^{\omega} \in \mathbb{R}^{N\times 1}$ is introduced to guide the sampling, with a control factor ω adjusting the level of influence from the possibility values.

Table 1. Average points and pixels by depth range and object category.

Depth Range	Number of Points / Pixels		
	Car	Ped.	Cyc.
Near (0~30m)	403 / 22,392	162 / 9,633	160 / 9,643
Mid. (30~50m)	41 / 1,964	18 / 759	25 / 913
Far (50~70m)	12 / 768	9 / 332	10 / 306

In addition, balance coefficients λ_c, λ_p, and λ_r are applied to adjust the importance of different object classes. These coefficients are determined based on the representation of each object class in the point cloud, with higher values assigned to classes that have fewer points, such as pedestrians and cyclists. For our experiments, we used $\lambda_c = 1$, $\lambda_p = 2$, and $\lambda_r = 2$, ensuring a higher sampling rate for smaller objects. Background points ρ^b are

excluded from the sampling process to avoid introducing noise. Table I presents the weights applied to each object class, calculated in inverse proportion to the number of points associated with the object in the point cloud.

4.2 Relevant Point Collection

To enhance the ability of the refinement network to learn more valuable features, the SD-AFDet framework employs a relevant points collection (RPC) mechanism based on the output from the RPN. This approach is designed to ensure that the most informative point features are retained for subsequent processing, thereby improving the overall detection performance. The procedure, depicted in Figure 11, consists of two main steps: proposal-oriented collection and score-oriented collection.

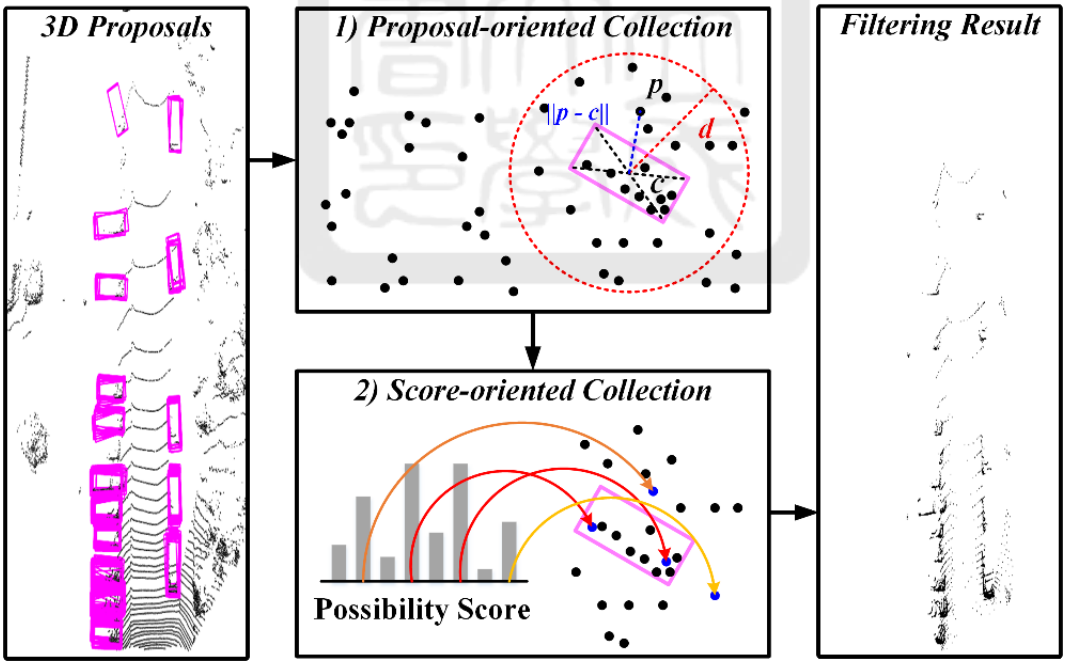


Figure 11. Relevant point collection strategy. The relevant point collection strategy employed in the SD-AFDet framework consists of two key stages, designed to optimize the set of points used for 3D object detection. The first part, potential foreground points P' are

collected based on the predicted bounding boxes. Points are selected according to their proximity to the centers of these bounding boxes, ensuring that only points within the vicinity of the proposals are considered. This method focuses the point selection on regions likely to contain objects, reducing the influence of irrelevant background points. The second part, the P-FPS algorithm refines the point set further by selecting k points from P', denoted P'', based on the confidence scores s, predicted by the 3D backbone. The confidence scores act as weights, guiding the algorithm to prioritize points with higher relevance to the detection task. This two-stage collection strategy ensures that the most relevant points are selected for the final detection, improving both the accuracy and computational efficiency of the detection process.

1) Proposal-oriented collection: As illustrated in Figure 11, the RPC process starts with a proposal-oriented method. In this step, the framework uses proposals generated by the RPN to select a pool of points. The raw points $P \in \mathbb{R}^{N\times 3}$ are filtered based on their distance to the centers of the proposed bounding boxes $C \in \mathbb{R}^{M\times 3}$, where M represents the number of proposals and N is the total number of points in the cloud.

For each point p_i the closest bounding box b_j and its center c_j are identified. A point p_i is included in the collection set P' if its distance from c_j is less than a threshold $d_j = \sqrt{(b_j^w)^2 + (b_j^l)^2} + \delta_o$, where b_j^w and b_j^l denote the width and length of the bounding box b_j, and δ_o is an offset (set to 1.5 in the experiment). The equation for this process is:

$$P' = \left\{ p_i \;\middle|\; \begin{array}{l} ||p_i - c_j|| < d_j, \\ p_i \subset P \in \mathbb{R}^{N\times 3}, c_j \subset C \in \mathbb{R}^{M\times 3} \end{array} \right\}. \tag{18}$$

This ensures that points close to the proposal boxes are selected, eliminating outliers and

irrelevant background points.

2) Score-oriented collection: After the proposal-oriented selection, the score-oriented collection refines the point set further. Using the P-FPS algorithm, *n* points are sampled from the set P', guided by the confidence scores s predicted by the 3D backbone. These confidence scores replace the possibility weights used in earlier stages of the P-FPS algorithm (as detailed in Section 4.1). This produces the final set $P'' \in \mathbb{R}^{k\times 3}$, as defined by:

$$P'' = P\text{-}FPS(P', s), \tag{19}$$

where k represents the final number of sampled points. The proposal-oriented step filters irrelevant points, while the score-oriented process ensures that foreground points are prioritized within the fixed-size point set used for training. Maintaining point diversity throughout this process helps prevent overfitting and improves detection accuracy. It's important to note that confidence scores are not directly used to select points, as preserving diversity during training is critical. In this stage, the control variable ω in this section is set to 2, balancing the effect of confidence scores on the sampling process.

4.3 Density-Aware Artificial Point Shift and Fusion Strategies

In the refinement stage of 3D object detection, we employ a point-wise fusion strategy [13], [25]–[27], [62] to extract pixel-level features from the image branch backbone, enhancing the model's capacity to distinguish between objects. However, the inherent sparsity of point clouds poses a significant challenge, leading to a low-resolution scene representation. This issue is particularly pronounced when detecting small and distant objects, as there is a noticeable mismatch between the resolution of raw points in ground

truth (GT) bounding boxes and the resolution of pixels projected across varying depth ranges, as illustrated in Table 1. This mismatch hampers the effectiveness of the camera stream in detecting smaller objects.

To mitigate this problem, we introduce artificial points that are denser than the raw point cloud and uniformly distributed within each proposal. These artificial points are projected onto the 2D space to sample relevant pixel-level features, significantly improving the feature extraction process. However, while uniformly distributed, these artificial points may fail to efficiently capture important geometric features, especially in occluded areas where points may not aggregate neighboring features effectively, reducing the overall performance of the fusion.

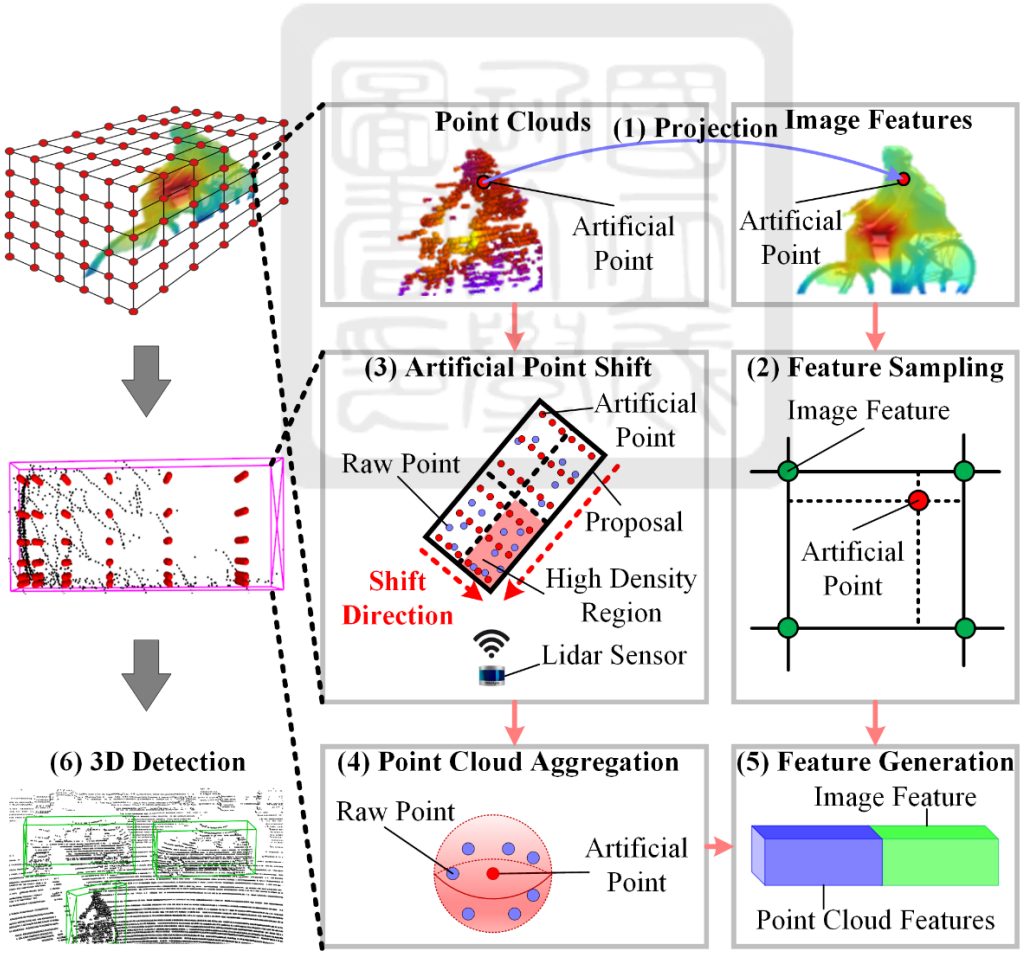


Figure 12. DAPSF strategy. In the refinement stage, the detection head leverages methods inspired by [7], [60], [61] to aggregate point features effectively. By incorporating boundary

information into the region of interest (ROI) features, the model is able to better distinguish between object categories. This approach also facilitates efficient fusion of different data types, guided by the confidence scores s predicted by the 3D backbone, thereby improving overall detection performance.

To address this issue, we propose a density-aware shift strategy that adjusts the position of artificial points, shifting them toward regions with higher point density. This strategy enhances the ability of artificial points to capture surrounding geometric features. As shown in Figure 12, each proposal is divided into four equal subregions D_1, D_2, D_3, and D_4 and point density is computed for each. Artificial points are shifted toward denser regions, using a Linear Increasing Discretization (LID) method [39], ensuring better alignment with the point cloud's dense areas. The results of this shift are visualized in Figure 13, where the artificial points are more concentrated in regions with more raw points, allowing them to aggregate richer features from the environment.

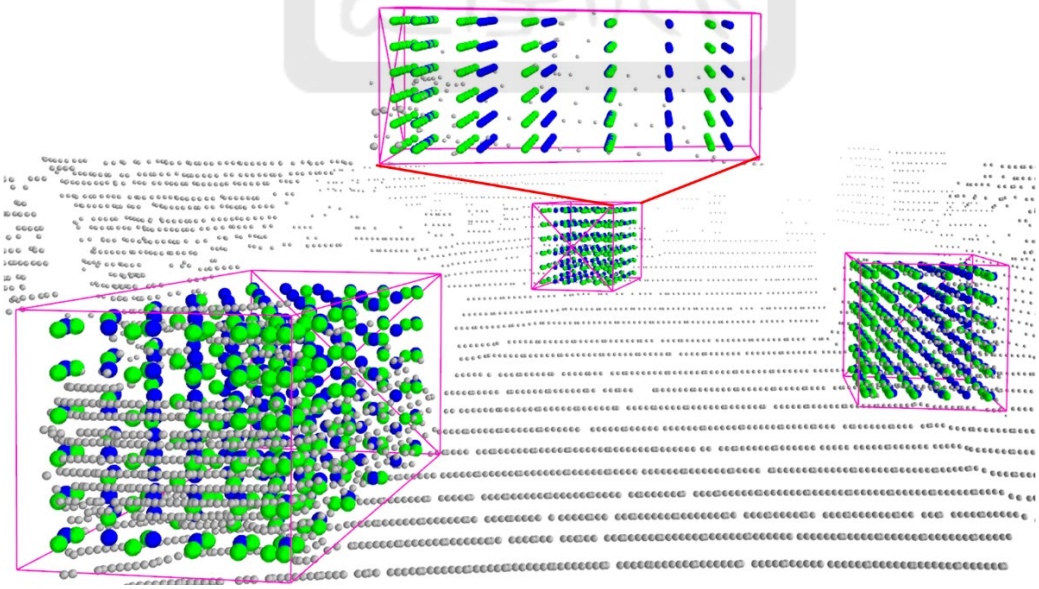


Figure 13. Visualization of artificial point shifting using LID method. The raw points (gray),

uniformly distributed artificial points (blue), and shifted artificial points (green) are shown, with a bounding box highlighting the point cloud distribution from a side view. The results indicate that the shifted artificial points are located in denser regions of the raw point cloud, demonstrating improved feature aggregation.

The shifted coordinates (x_{shift}, y_{shift}), are calculated based on the original coordinates (x_o, y_o),using:

$$x_{shift} = x_o + \frac{(x_{max} - x_{min}) \cdot \partial_x}{G(G-1)} \cdot x_i, \tag{20}$$

$$y_{shift} = y_o + \frac{(y_{max} - y_{min}) \cdot \partial_y}{G(G-1)} \cdot y_i, \tag{21}$$

where L is the discretized side length, and ∂_x, ∂_y control the displacement direction. The direction depends on the point's original location, and adjustments are made based on the grid configuration to ensure better alignment with denser regions.

$$\delta_x, x_o = \begin{cases} 1,\ x_{min},\ if\ D_o \in\ D_3,\ D_4 \\ -1,\ x_{max},\ if\ D_o \in\ D_1,\ D_2 \end{cases}, \tag{22}$$

$$\delta_y, y_o = \begin{cases} 1,\ y_{min},\ if\ D_o \in\ D_3,\ D_4 \\ -1,\ y_{max},\ if\ D_o \in\ D_1,\ D_2 \end{cases}. \tag{23}$$

This process allows the artificial points to be repositioned within the proposals, aligning them more closely with regions of higher point density, thereby maximizing their ability to group neighboring features. Once the artificial points have been shifted and their corresponding semantic features have been sampled, the geometric features are denoted as

$f_p \in \mathbb{R}^{K\times C}$, where K is the number of points and C represents the feature channels. These features are fused with the semantic features f_c using a lightweight module, which can be described by the following equation:

$$f_f = \phi(m(sum(f_p, m(f_c)))), \tag{24}$$

where m represents a MLP, while ϕ refers to a convolutional layer with ReLU activation and batch normalization. This fusion step combines the geometric and semantic features into a unified representation, denoted as $f_f \in \mathbb{R}^{K\times C}$, which is used to enhance the object detection process.

4.4 Loss Function

The SD-AFDet architecture combines a RPN and a refinement network (RCNN), with the overall loss function being the sum of the losses from both stages:

$$L_{total} = L_{rpn} + L_{rcnn}, \tag{25}$$

where L_{rpn} and L_{rcnn} represent the losses from the RPN and RCNN stages, respectively. Each stage contributes losses for both the confidence score and bounding box regression.

In the RPN stage, the confidence score loss L_{focal} addresses the issue of class imbalance between foreground and background samples by using focal loss. The bounding box regression loss L_{reg}^{rcnn} is computed using smooth L1 loss. Together, the total RPN loss is:

$$L_{rpn} = L_{focal} + L_{reg}^{rpn}, \tag{26}$$

Focal loss reduces the impact of easy-to-classify background samples, focusing on more challenging examples. The loss is defined as:

$$L_{focal} = -\alpha_t(1-p_t)^{\lambda}\, log(p_t), \tag{27}$$

where p_t is the predicted confidence score, λ controls the influence of easy samples, and α_t balances the weighting of positive and negative samples. We set, $\alpha_t = 0.25$ and $\lambda = 2$ in our experiments.

For the RCNN, the confidence score loss is computed using binary cross-entropy (BCE) loss, and the bounding box regression loss is calculated using smooth L1 loss. An additional direction loss L_{dir} is used to account for the yaw rotation of objects. The total RCNN loss is:

$$L_{rcnn} = L_{bce} + L_{reg}^{rcnn} + L_{dir}, \tag{28}$$

The 3D bounding boxes are parameterized by $(x_g, y_g, z_g, h_g, w_g, l_g, \theta_g)$ for the GT and $(x_g, y_g, z_g,\ l_g, w_g, h_g)$ for the predicted boxes. The residuals between these parameters are expressed as $\Delta t = \{\Delta x,\ \Delta y,\ \Delta z,\ \Delta l,\ \Delta w,\ \Delta h,\ \Delta\theta,\ d_m\}$ where:

$$\Delta x = \frac{x_g - x_p}{d_p}, \Delta y = \frac{y_g - y_p}{d_p}, \Delta z = \frac{z_g - z_p}{d_p}, \tag{29}$$

$$\Delta l = log(\frac{l_g}{l_p}), \Delta w = log(\frac{w_g}{w_p}), \Delta h = log(\frac{h_g}{h_p}), \tag{30}$$

$$\Delta\theta = sin(\theta_g) - sin(\theta_p)\ \ , d_m = \sqrt{h_p^2 + w_p^2}. \tag{31}$$

The smooth L1 loss is applied over the residuals to calculate the regression losses for both RPN and RCNN based on the number of positive predictions N_{pos}:

$$L_{reg} = \frac{1}{N_{pos}} \sum_i \text{Smooth}_{\text{L1}}(\Delta t). \tag{33}$$

This ensures that the predicted bounding boxes align closely with the ground truth, improving the accuracy of 3D object detection by optimizing the position, dimensions, and orientation of detected objects.

CHAPTER 5. EXPERIMENT RESULTS AND ANALYSIS

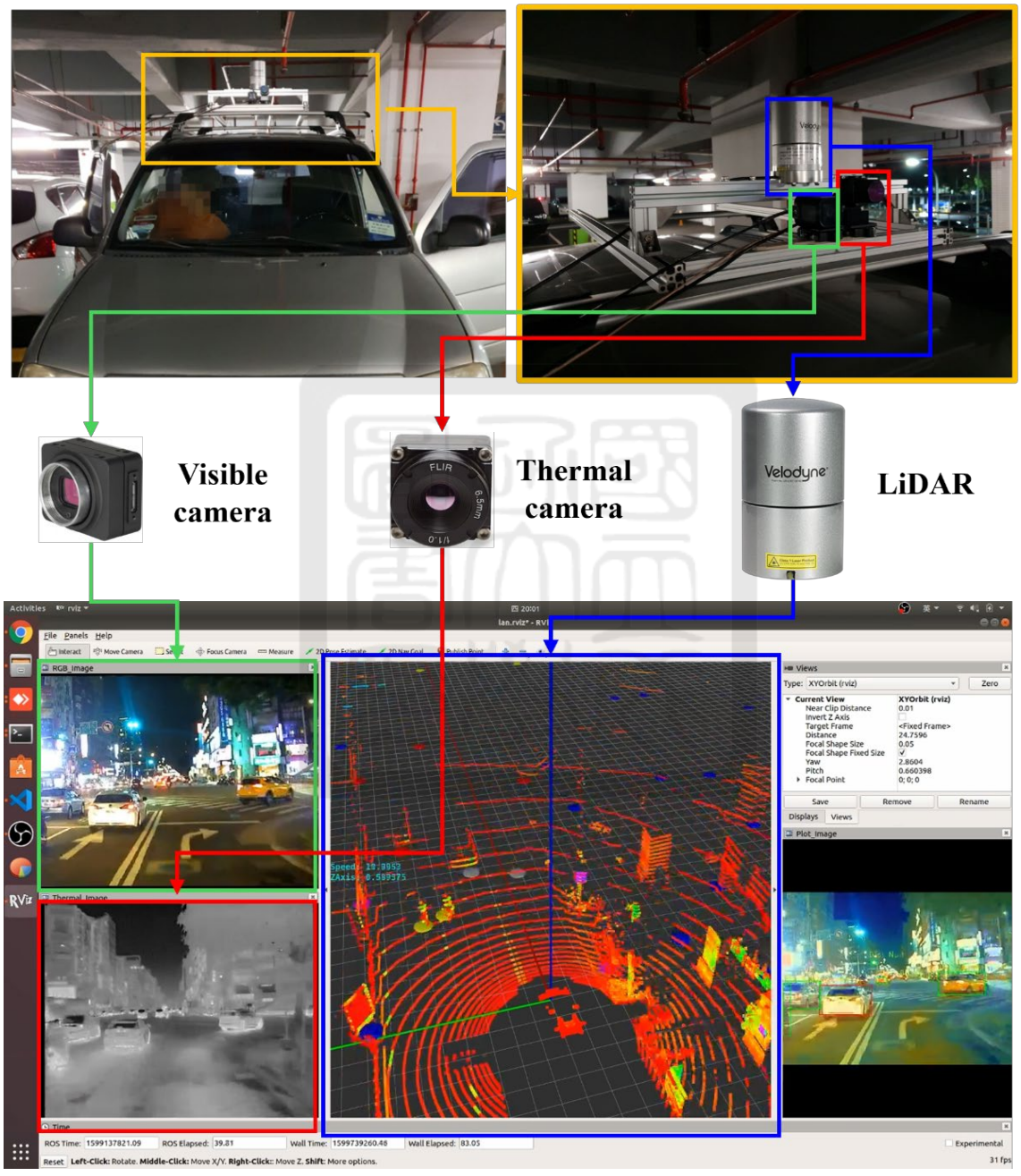


Figure 14. Sensor configuration and data fusion visualization on the experimental vehicle. The figure displays the integration of visible (green line), thermal (red line), and LiDAR

(blue line) sensors on the experimental vehicle. The visible and thermal cameras, mounted at the front of the vehicle, capture high-resolution RGB and thermal images. A Velodyne LiDAR sensor, mounted on the roof, provides 3D point cloud data. The lower section visualizes the synchronized data fusion from these sensors, showing how RGB images, thermal images, and LiDAR point clouds complement each other. The RGB view shows nighttime urban traffic, while the thermal images provide critical details in low-light conditions. The LiDAR data enhances object detection by contributing 3D spatial information, crucial for precise localization and object identification in autonomous driving systems.

This chapter details the experimental setup and performance evaluation for the proposed 2D and 3D object detection frameworks in autonomous driving. As illustrated in Figure 14, a specially outfitted experimental vehicle was used to collect real-world data across various driving environments in Taiwan, simulating conditions relevant to autonomous driving. Sensors were mounted on the roof of the vehicle, while data processing was handled by an in-vehicle computing platform located inside. The platform, powered by an Intel XEON (R) E-2288G CPU @ 3.70 GHz and equipped with a high-speed SSD (CVB-CD1024), facilitated high-speed data transfer, minimizing latency and reducing risks of desynchronization or data loss during acquisition.

To ensure high-quality image capture, the vehicle was fitted with an industrial camera (CM3-U3-31S4C-C) capable of 2048×1536 pixel resolution at 55 FPS, paired with an SV-0614V lens that provides a 54.6-degree horizontal field of view. Additionally, a Boson640 thermal imaging camera was employed, featuring a 50° horizontal field of view and a long-wave infrared sensor optimized for thermal contrast. This setup enabled precise preprocessing, especially when compensating for thermal variations over time.

The vehicle was also equipped with a Velodyne HDL-32E LiDAR, operating at 20 Hz. The data collection platform was mounted on the roof, with both the visible and thermal cameras adjusted for optimal alignment in both horizontal and vertical directions. The LiDAR was positioned on the roof to avoid obstruction from the cameras. Synchronization was achieved by using the LiDAR's sampling time as a reference, and a ROS synchronization mechanism was employed to ensure alignment between the camera and LiDAR data. After synchronization, all sensors, including the cameras and the LiDAR, operated at a frame rate of 20 FPS.

Section 5.1 introduces the datasets used for evaluation, Section 5.2 provides details on the experimental setup and the evaluation metrics used, Section 5.3 presents a comparative analysis of the proposed frameworks against existing methods, and Section 5.4 includes an ablation study, exploring the contributions of each component within the VL-ACFDet and SD-AFDet frameworks.

5.1 Dataset Collection

To evaluate the effectiveness of the proposed frameworks, we conducted experiments using three datasets: the M3FD 2D object detection dataset [28], the KITTI 3D object detection dataset [29], and a newly collected dataset from our experimental vehicle.

The M^3FD dataset is a well-established benchmark for 2D multimodal object detection, consisting of 4,200 pairs of visible and thermal frames and a total of 33,603 annotated bounding boxes. The annotations include various object categories such as people, cars, buses, motorcycles, trucks, and street lamps. This dataset covers diverse scenarios, including daytime, nighttime, and adverse weather conditions such as fog and rain. Since no official partitioning scheme is available, we adopted the train/test splits from [63], [64], which

include 3,368 training and 832 testing image pairs, with a resolution of 512×640 pixels. In addition, 10% of the training data was set aside for validation.

The KITTI dataset serves as a standard benchmark for 3D object detection, offering paired frames of visible images and LiDAR data. It consists of 7,481 frames for training and 7,518 frames for testing. Following the data split methodology used in previous studies [7], [13], [14], [25]–[27], the training data is further divided into 3,712 frames for training and 3,769 frames for validation. The dataset provides both camera images and LiDAR point clouds collected using a Velodyne HDL-64E LiDAR sensor. Objects are grouped into easy, moderate, and hard categories based on difficulty, which is determined by object size, occlusion, and truncation levels, reflecting varying detection challenges.

To further assess the robustness of the proposed frameworks, we collected a new dataset using the experimental vehicle. Following the setup from [10], this dataset was expanded to include additional day and night driving scenarios, as well as adverse weather conditions such as rain and fog. The collected data was spatially aligned and time-synchronized, with a standardized resolution of 640×480 pixels. This new dataset contains 132,905 annotated 2D and 3D bounding boxes, including 29,715 pedestrians, 76,407 cars, and 26,783 motorcycles. The data was split into training (80%) and testing (20%) sets, with 10% of the training set reserved for validation. Compared to the M3FD and KITTI datasets, this new dataset provides a broader range of challenging conditions, allowing for a more comprehensive evaluation of the proposed 2D and 3D object detection models.

5.2 Experiment Settings and Evaluation Metrics

5.2.1 2D Object Detection Task Settings

The VL-ACFDet framework utilizes a dual-stream architecture based on CFT [12] and extends the YOLOv5m backbone. Several network layers were initialized using pre-trained weights from the MS-COCO dataset, while others were randomly initialized. The framework was implemented using PyTorch and trained on an NVIDIA RTX 6000 Ada GPU with a batch size of 32. Model optimization was carried out using the Adam optimizer, with an initial learning rate of 0.01, momentum of 0.937, a weight decay of 0.0005, and a final learning rate of 0.002. The one cycle learning rate policy was applied to dynamically adjust the learning rate during training. The loss function weight factors were configured as λ_o=1, $\lambda_c = 0.5$, λ_{CIOU} =0.05, $\lambda_T = 0.1$ and, $\lambda_I = 0.1$. To enhance generalization, data augmentation was applied as described in [7], [12]. For the M3FD dataset, the best-performing model was selected after 300 epochs, resulting in a total training time of approximately 16 hours. On a newly collected dataset, the optimal model was identified after 100 epochs, with a total training time of about 7.5 hours.

The model's performance was evaluated using the mean average precision (mAP) metric at various Intersection over Union (IoU) thresholds. Specifically, mAP_{50} and mAP_{75} represent the mean AP at IoU thresholds of 0.5 and 0.75, respectively. The $mAP_{50\text{-}95}$ metric, on the other hand, reflects the mean AP across IoU thresholds ranging from 0.50 to 0.95, in 0.05 increments. Higher mAP scores indicate superior detection performance, providing a thorough evaluation of the model's capability across a wide range of detection conditions.

5.2.2 3D Object Detection Task Settings

In the SD-AFDet framework, paired raw visible camera images and LiDAR point clouds serve as inputs to the model. The point cloud is constrained to the following camera

coordinate ranges: X-axis [-40, 40], Y-axis [-1, 3], and Z-axis [0, 70.4]. Since the focus is on multi-modality fusion, we excluded points beyond the camera's field of view and downsampled the data, retaining 16,384 raw points. The backbone network was inspired by previous models [57], utilizing four set abstraction layers to downsample the point cloud into clusters of 4,096, 1,024, 256, and 64 points, followed by four feature propagation layers to restore the original resolution. For generating artificial points, each side of the discretized length L was set to 6, creating 216 artificial points per proposal. The camera stream used the Deeplabv3+ [56] pre-trained model, with categories mapped manually to the KITTI format.

For the 2D backbone, the model was first pretrained on Mapillary [65] with 300,000 training iterations and 100,000 validation iterations, using a learning rate of 0.01. It was then fine-tuned on the Cityscapes [66] dataset for 30,000 iterations with a learning rate of 0.001, followed by 10,000 iterations on the KITTI semantic segmentation task at the same learning rate. Training was conducted using SGD with a batch size of 16.

For the multi-modality model, the adam optimizer was used, dynamically adjusted using the cosine annealing strategy. The weight decay was set to 0.002, and momentum to 0.9. The model was trained with a batch size of 8 for 80 epochs, which took approximately 40 hours. During training, non-maximum suppression (NMS) was applied to filter proposals, retaining the top 9,000 highest-scoring candidates. An IoU threshold of 0.8 was used to refine bounding boxes, and in the refinement stage, we subsampled 512 proposals to focus on the most relevant detections. During testing, an IoU threshold of 0.85 was applied, with 100 proposals retained for final evaluation.

To mitigate overfitting, various data augmentation techniques were applied, including GT sampling, scaling, rotation, and scene flipping [67]. GT sampling involved cropping ground truth boxes from the dataset and placing them into point cloud frames to simulate more complex road environments. The scaling factor was randomly chosen between 0.95

and 1.05, and the rotation angle ranged from -0.78 to 0.78 radians. Flipping along the X-axis added further variability. Despite the potential inconsistencies introduced by GT sampling in camera-LiDAR models, the method from [62] effectively mitigated these challenges, ensuring smoother implementation.

To ensure fair comparison with other models, performance was measured using the official average precision (AP) protocol, which calculates AP across three difficulty levels—easy, moderate, and hard—using 40 recall points.

5.3 Quantitative Evaluation

5.3.1 Performance Comparison of VL-ACFDet with SOTA 2D Object Detection Methods

To assess the performance of the proposed VL-ACFDet framework, we compared it against several SOTA multispectral object detection methods, ensuring that all parameter settings matched their original configurations. As shown in Table 2, VL-ACFDet consistently outperformed existing models, including TarDAL [28] (early fusion) and QFDet [67] (late fusion), achieving mAP improvements of 6.17% and 4.78%, espectively, on the “ALL” subset of the M3FD dataset. Compared to other mid-fusion strategies, VL-ACFDet achieved mAP gains ranging from 1.6% to 10.1%, highlighting its superior robustness and detection accuracy under various conditions. Notably, VL-ACFDet excelled on the “Adverse” subset, which includes challenging weather conditions, outperforming illumination-aware methods like UA-CMDet [11] and achieving the highest mAP of 84.58%. This exceptional performance is largely attributed to the VL-CAT module’s use of VLFMs, which enriches semantic information, enhancing detection stability under adverse weather.

Table 2. Performance comparison of VL-ACFDET with SOTA methods on the M^3FD dataset.

Method	Year	mAP (%)				AP (%)					
		All	Day	Night	Adverse	People	Car	Bus	Lamp	Motor	Truck
UA-CMDet [11]	2022	76.50	74.36	86.13	74.29	72.79	79.90	91.68	53.14	74.45	87.04
TarDAL[a] [28]	2022	80.50	-	-	-	81.50	94.80	81.30	87.10	69.30	68.70
CFT [12]	2022	82.24	79.54	91.38	76.00	78.48	91.21	90.67	79.42	70.73	82.93
CALNet [33]	2023	81.88	79.23	91.71	77.41	73.55	88.65	88.11	84.75	78.01	78.22
QFDet [67]	2023	81.89	77.85	92.21	79.23	80.76	86.37	89.24	84.22	69.97	80.76
ADCNet [68]	2024	83.77	80.75	92.84	71.88	81.50	89.49	90.50	78.93	78.81	83.36
ICAFusion [34]	2024	84.55	82.58	93.85	79.26	79.77	91.63	91.10	81.64	74.66	88.51
Fusion-Mamba[a] [69]	2024	85.00	-	-	-	80.30	91.90	92.80	84.80	73.00	87.10
This work		**86.67**	**83.79**	**95.80**	**84.58**	**82.53**	**92.82**	**92.18**	**86.22**	**78.07**	**88.21**

[a]Results for TarDAL and Fusion-Mamba are directly cited from the original paper [69].

Table 3. Performance comparison of VL-ACFDET with SOTA methods on the newly collected dataset.

Method	Year	mAP (%)				AP (%)		
		All	Day	Night	Adverse	person	motor	car
UA-CMDet [11]	2022	67.69	61.50	69.06	57.79	66.87	69.13	67.08
CFT [12]	2022	75.91	62.75	78.38	57.86	72.51	78.86	76.36
CALNet [33]	2023	72.63	54.66	75.89	59.23	69.39	73.35	75.16
QFDet [67]	2023	74.88	64.14	77.58	55.32	70.17	78.18	76.31
ADCNet [68]	2024	70.81	61.12	77.34	58.18	67.51	75.34	69.59
ICAFusion [34]	2024	77.25	63.59	79.71	59.98	74.44	80.33	76.67
This work		**79.42**	**68.33**	**81.58**	**62.93**	**76.30**	**81.99**	**79.96**

VL-ACFDet was also tested on our newly collected dataset, which includes a more extensive and diverse set of scenarios compared to the M3FD dataset. As detailed in Table 3, VL-ACFDet consistently achieved higher mAP values across all subsets, outperforming the second-best method by 2.17% on “ALL”, 4.19% on “Day”, 1.87% on “Night”, and 2.95% on “Adverse”. These results affirm the effectiveness of VL-ACFDet in integrating multispectral data to enhance object detection and classification. Further analysis across object-specific subsets revealed that VL-ACFDet achieved high AP for detecting cars, pedestrians, and motorcycles, with APs of 76.30%, 81.99%, and 79.96%, respectively, at an

IoU of 0.5. These findings underscore the framework's robustness and adaptability in handling complex, real-world conditions.

5.3.2 Performance Comparison of SD-AFDet with SOTA 3D Object Detection Methods

Table 4. Performance comparison with sota on the KITTI test set

Method	Year	Car_{3D} AP (%)			$Pedestrian_{3D}$ AP (%)			$Cyclist_{3D}$ AP (%)			mAP (%)
		Mod.	Easy	Hard	Mod.	Easy	Hard	Mod.	Easy	Hard	
Lidar-based SOTA methods											
SASA-SSD [70]	2022	82.16	88.76	77.16	-	-	-	-	-	-	-
PointPillars [71]	2019	74.31	82.58	68.99	41.92	51.45	38.89	58.65	77.10	51.92	60.64
PointRCNN [57]	2019	75.64	86.96	70.70	39.37	47.98	36.01	58.82	74.96	52.53	60.33
DFAF3D [72]	2023	79.37	88.59	72.21	40.99	47.58	37.65	65.86	82.09	59.02	63.70
Part-A^2 [73]	2020	78.49	87.81	73.51	43.35	53.10	40.06	63.52	79.17	56.93	63.99
IA-SSD [74]	2022	80.13	88.34	75.04	39.03	46.51	35.60	61.94	78.35	55.70	64.39
PASS-PV-RCNN+[75]	2024	81.28	87.65	76.79	41.95	47.66	38.90	68.45	80.43	60.93	64.89
Lidar-camera based SOTA methods											
PI-RCNN [62]	2021	74.82	84.37	70.03	-	-	-	-	-	-	-
EPNet [13]	2020	79.28	89.81	74.59	-	-	-	-	-	-	-
3D-CVF [47]	2020	80.05	89.20	73.11	-	-	-	-	-	-	-
Pointformer [14]	2021	77.06	87.13	69.25	42.43	50.67	39.60	59.80	75.01	53.99	61.66
CL3D [15]	2022	80.28	87.45	76.21	39.42	47.30	36.97	62.02	77.33	55.52	62.50
F-ConvNet [21]	2019	76.39	87.36	66.69	43.38	52.16	38.80	65.07	81.98	56.54	63.15
EPNet++ [27]	2023	81.96	91.37	76.71	44.38	52.79	41.29	59.71	76.15	53.67	64.23
This work		**81.86**	**88.82**	**77.26**	**41.39**	**47.59**	**38.37**	**67.55**	**82.10**	**60.70**	**65.07**

A dash ('-') is used in the table to represent cases where results were either not reported by the original authors.

The proposed SD-AFDet framework was evaluated against various state-of-the-art (SOTA) methods using the KITTI testing dataset. As summarized in Table 4, SD-AFDet demonstrated substantial improvements over the baseline model [57], with AP gains ranging from 1% to 7% for cars across three difficulty levels. Additionally, SD-AFDet exhibited a 2%+ AP improvement for pedestrians at the moderate and hard levels. The most notable enhancement was in cyclist detection, where SD-AFDet achieved an AP increase of 8%–9% across all difficulty levels, illustrating its robust handling of complex conditions, such as occlusion and truncation.

In addition to outperforming the baseline, SD-AFDet surpassed other SOTA camera-LiDAR fusion approaches [13]–[15], [21], [27], [47], [62]. While SASA-SSD [70] ranked slightly higher than SD-AFDet for moderate-level car detection, SD-AFDet exhibited superior performance in most categories, especially across object types not fully addressed by SASA-SSD, raising concerns about the latter's stability across diverse tasks. Nonetheless, SD-AFDet excelled in car detection when compared to other LiDAR-based models [70]–[75], achieving an overall mean Average Precision (mAP) of 65.07% across all tasks.

To further assess the model's generalization capabilities, SD-AFDet was tested on a newly collected dataset, and its performance was compared against SOTA models with publicly available codebases. Parameter settings were kept consistent with those used for the KITTI dataset, ensuring a fair comparison. As presented in Table 5, SD-AFDet maintained its strong performance, although the improvement in pedestrian detection was less significant compared to EPNet++. However, SD-AFDet excelled in other categories, particularly in motorcycle detection, where it surpassed the second-best model by more than 10%. On this new dataset, SD-AFDet achieved a commendable mAP of 63.65%, highlighting its robustness and effectiveness in real-world applications.

Table 5. Performance comparison of SD-AFDet with SOTA methods on the newly collected dataset.

Method	Year	AP (%)			mAP (%)
		Person	Motor	Car	
EPNet [16]	2020	46.55	49.25	70.33	55.38
Pointformer [14]	2021	50.80	53.27	70.13	58.06
3D-CVF [47]	2020	49.88	52.60	71.95	58.14
EPNet++ [27]	2023	53.10	55.43	71.88	60.18
This work		**52.55**	**65.91**	**72.55**	**63.65**

5.4 Ablation Studies

5.4.1 Analysis of the VL-ACFDet Framework

Table 6. Ablation study of each component of the VL-ACFDet framework on the newly collected dataset.

Method		mAP_{50}	mAP_{75}	mAP_{50-95}	Parameter	FPS
Visible only		73.40	42.70	43.30	7,018,216	250
Thermal only		65.60	28.80	34.20		
Baseline (CFT) [12]		75.91	40.77	42.19	44,506,376	99
AC-CA	VL-CAT	Proposed VL-ACFDet Method				
✓		77.89	44.91	44.98	34,701,704	124
	✓	78.39	46.14	45.87	46,114,824	95
✓	✓	79.42	47.84	47.07	36,310,152	123

1) Effectiveness Evaluation of Each Component: We evaluated the individual contributions of each component within the VL-ACFDet framework using our collected dataset. By systematically modifying the baseline architecture and integrating the proposed modules, we assessed their impact on overall performance. Comparisons were made with single image modality configurations (“Thermal-only” and “Visible-only”) using YOLOv5 and the baseline CFT model [12]. As shown in Table 6, single modality methods delivered lower performance, with the "Thermal-only" configuration performing the poorest. Although the baseline CFT model [12] demonstrated improvements over single modality setups, it struggled to perform well at higher IoU thresholds and exhibited inefficiencies due to a large number of parameters and inadequate redundancy management in raw modality features.

Replacing the standard Transformer with the AC-CA module significantly improved performance across various IoU thresholds, reduced the parameter load, and enhanced

processing speed (FPS), validating the module's capability in effective feature selection and alignment. The addition of the VL-CAT module further increased performance, with substantial mAP gains of 2.48%, 5.37%, and 3.68% at mAP50, mAP75 and mAP50-95, respectively, accompanied by a modest 4% increase in parameters. The fully integrated VL-ACFDet framework, combining both AC-CA and VL-CAT modules, achieved the highest performance levels, with mAP50 reaching 79.42%, mAP75 at 47.84, and mAP50-95 at 47.07, while maintaining an FPS of 123. These results confirm the effectiveness of the combined modules in enhancing detection accuracy while ensuring efficient processing, making VL-ACFDet highly suitable for real-time applications.

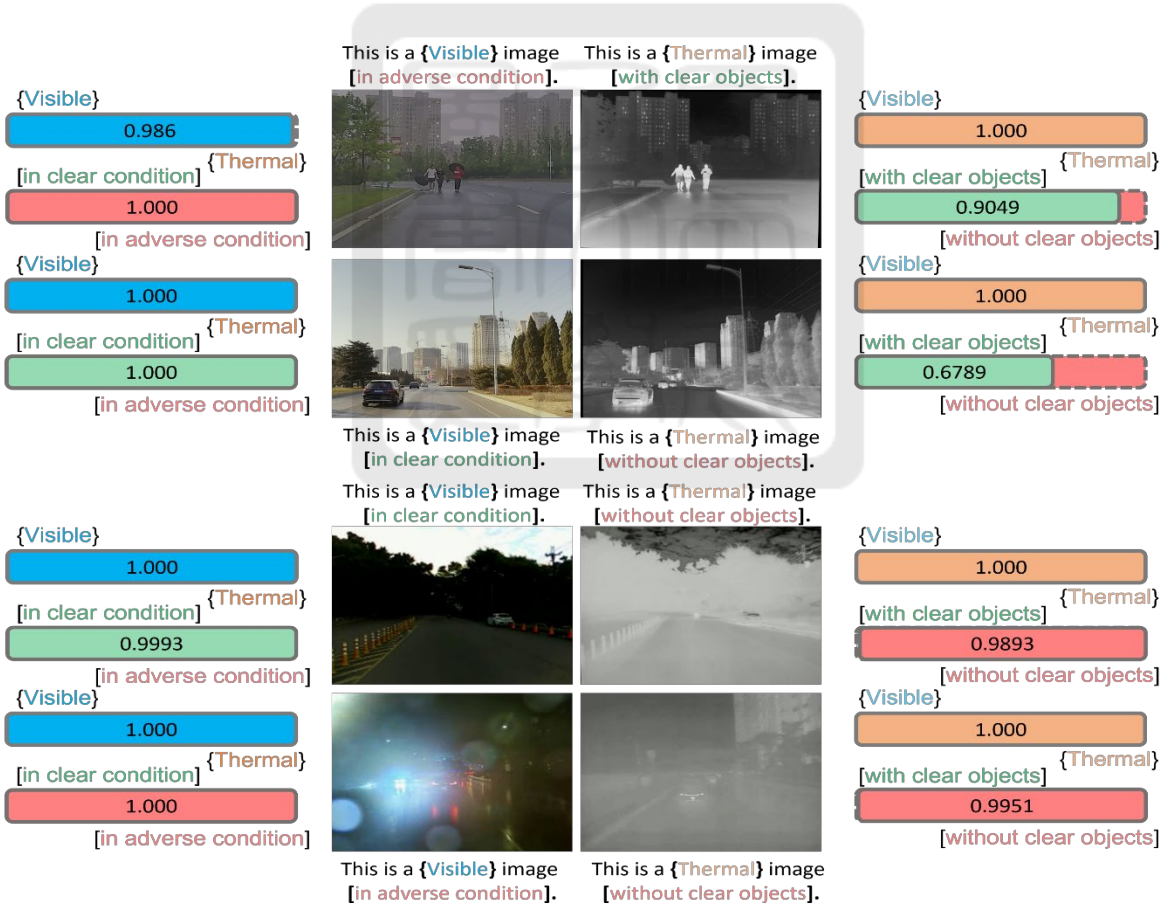


Figure 15. Modality quality assessment using CLIP across different scenarios. This figure showcases visible and thermal image pairs from two datasets, evaluated under clear and

adverse conditions using CLIP with specially crafted prompts. Visible images are categorized as “in clear condition” or “in adverse condition,” while thermal images are assessed as containing “clear objects” or “without clear objects.” The confidence scores demonstrate CLIP’s effectiveness in accurately distinguishing between clear and adverse conditions and evaluating image clarity in challenging situations, such as low contrast or fog. These assessments are pivotal in enhancing VL-ACFDet’s performance by guiding the prioritization of critical features, thereby boosting detection accuracy in complex environments.

2) Effectiveness of VLFMs in Modality Quality Assessment: To evaluate the role of CLIP in assessing modality quality, we utilized targeted prompts designed to identify the modality (visible or thermal) and evaluate image clarity. For visible images, prompts distinguished between “clear” and “adverse” conditions, while thermal images were categorized as “clear” or “without clear” objects. Figure 15 demonstrates CLIP’s ability to accurately differentiate modality quality across various scenarios, including clear and adverse conditions. This analysis emphasizes the importance of prompt engineering and demonstrates that VLFMs can effectively guide semantic evaluation, enhancing the robustness of multispectral object detection.

3) Visualization Analysis: We conducted a visualization analysis to compare VL-ACFDet with the CFT baseline [12] across the M3FD and our collected datasets. As illustrated in Figure 16, VL-ACFDet exhibited superior accuracy and stability, particularly under adverse conditions such as poor lighting and rain, where the CFT model often missed small or distant objects. In contrast, VL-ACFDet maintained strong performance by effectively fusing critical modality features, even when both modalities faced challenges, as depicted in Figure 16 (c). These visualizations highlight VL-ACFDet’s robustness and accuracy in complex,

adverse weather scenarios, demonstrating its capability to outperform existing models in real-world conditions.

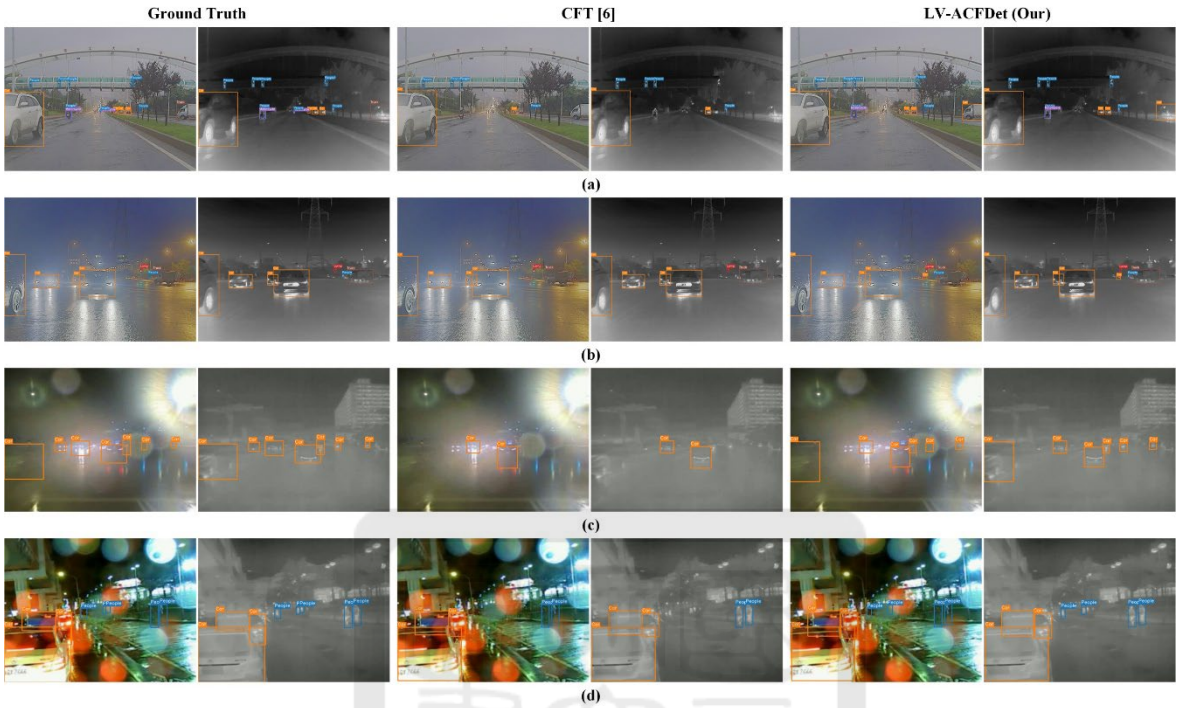


Figure 16. Comparative visualization of detection performance between the proposed VL-ACFDet framework, the CFT [12] baseline, and GT across various challenging scenarios. Subfigures (a) and (b) illustrate performance on the M3FD dataset, focusing on typical adverse conditions such as low visibility and harsh weather. Subfigures (c) and (d) are from our newly collected dataset, demonstrating the model's performance in additional complex, real-world environments, including challenging scenarios like rain and fog.

5.4.2 Analysis of the SD-AFDet Framework

1) Effectiveness Evaluation of Each Component: To evaluate the individual contributions of the key components within the SD-AFDet framework, as shown in Table 7, we conducted a detailed analysis focusing on the multi-modality feature extraction, RPC mechanism, and the DAPSF strategy. For these experiments, we used PointRCNN [57] as

the baseline model, and performance was assessed on the KITTI dataset, particularly focusing on improvements in detecting small and distant objects.

Table 7. Ablation study of individual components in the SD-AFdet on the KITTI validation dataset.

PA	P-FPS	RC	DAPSF	Car_{3D} AP (%)			$Pedestrian_{3D}$ AP (%)			$Cyclist_{3D}$ AP (%)		
				Mod.	Easy	Hard	Mod.	Easy	Hard	Mod.	Easy	Hard
				80.12	89.13	77.84	55.13	62.34	48.30	71.39	86.99	67.00
✓				79.98	88.93	77.61	57.49	64.75	51.15	73.94	89.24	69.20
✓	✓			83.07	91.85	80.6	60.18	67.77	54.31	75.20	92.87	70.04
✓	✓	✓		82.67	91.44	82.42	61.11	68.82	54.44	77.94	93.32	73.78
✓	✓	✓	✓	82.95	92.19	82.88	63.63	71.53	58.77	79.14	93.41	74.75

First, we evaluated the multi-modality feature extraction, which integrates information from both camera images and LiDAR point clouds, capturing complementary geometric and semantic features. This feature extraction process consists of two primary subcomponents: point-wise augmentation (PA) and P-FPS. PA augments the raw point cloud data with semantic information extracted from the image stream. This enhancement was particularly beneficial for distinguishing between objects, especially small and distant ones, where raw LiDAR data alone might lack sufficient detail. Incorporating PA alone resulted in noticeable performance improvements, particularly in detecting cyclists and pedestrians, as the additional semantic features provided more discriminative power.

Following that, we implemented P-FPS, which improved the efficiency of point sampling by prioritizing foreground points that are more relevant to the detection task. P-FPS further enhanced the model's ability to focus on relevant regions while maintaining diversity in the sampled points. The introduction of P-FPS led to significant gains in detection accuracy, with improvements of 2.82%, 5.49%, and 4.61% for cars, pedestrians, and cyclists, respectively, compared to the baseline model. These results demonstrate that P-FPS plays a crucial role in improving multi-modality feature extraction by intelligently

guiding the sampling process to focus on important regions within the scene.

The RPC mechanism was then assessed to evaluate its ability to select high-quality foreground points while filtering out irrelevant background data. By collecting points that are closest to the region proposals generated by the model, this mechanism reduces the amount of noise introduced into the network, leading to more efficient feature extraction. The integration of the RPC mechanism provided additional mAP gains, particularly for smaller and harder-to-detect categories like pedestrians and cyclists. Specifically, we observed a performance improvement of 0.70% for pedestrians and 1.79% for cyclists, demonstrating the effectiveness of this mechanism in improving detection accuracy for challenging object categories.

Finally, the DAPSF strategy was implemented to address the issue of point cloud sparsity, particularly in occluded regions or for distant objects. This strategy shifts the artificial points towards areas of higher point density, allowing for more effective aggregation of neighboring features. The density-aware shift helps ensure that the artificial points are positioned in regions where they can capture the maximum amount of geometric and semantic information. Incorporating this component into the model resulted in further performance enhancements, with mAP increases of 0.35% for cars, 3.18% for pedestrians, and 0.75% for cyclists. These results highlight the critical role of the DAPSF in aligning artificial points with denser areas of the point cloud, improving the model's ability to handle sparse data, especially for small and distant objects.

2) P-FPS in Multi-Modality Feature Extraction: The control factor ω is essential for fine-tuning the influence of possibility weights throughout the point selection process. By tuning ω, we can influence how much the model prioritizes foreground points based on their relevance and possibility scores. As shown in Table 8, the best results were achieved when $\omega = 10$. If ω is set too high, the P-FPS algorithm over-samples points based on their

possibility scores, which can reduce point diversity. On the other hand, if ω is set too low, the model fails to fully leverage the possibility-based weighting, diminishing its ability to prioritize foreground points effectively. The tuning of ω ensures that P-FPS strikes a balance between point diversity and the prioritization of relevant points, making it a crucial factor for enhancing detection performance.

Table 8. Hyperparameter analysis of ω in the backbone.

Method	Mod. (AP%)		
	Car_{3D}	$Pedestrian_{3D}$	$Cyclist_{3D}$
D-FPS [57]	80.12	55.13	71.39
P-FPS (ω = 0.1)	80.73	55.84	74.84
P-FPS (ω = 1)	82.82	59.46	75.85
P-FPS (ω = 10)	82.67	**63.63**	**79.14**
P-FPS (ω = 20)	**83.02**	63.39	78.57
P-FPS (ω = 30)	82.60	60.54	73.77

In addition, we compared P-FPS with other point sampling methods, such as S-FPS and SampleNet, across multiple object categories and difficulty levels. As shown in Table 9, our P-FPS method outperformed these alternative sampling techniques across all tasks, particularly at the moderate difficulty level, further validating its effectiveness in multi-modality feature extraction.

Table 9. Comparison of our point cloud sampling scheme with existing methods.

Method	Mod. (AP%)		
	Car_{3D}	$Pedestrian_{3D}$	$Cyclist_{3D}$
D-FPS [57]	80.12	55.13	71.39
SampleNet [76]	81.22	58.22	75.01
S-FPS [70]	82.95	60.16	74.47
P-FPS (Ours)	**83.07**	**60.18**	**75.20**

3) P-FPS in Relevant Point Selection: Beyond multi-modality feature extraction, P-FPS

is also crucial in the relevant point selection process. The control factor ω and the number of sampled point k both significantly impact model performance. As shown in Figure 17, the highest mAP was achieved when $\omega = 10$ and $k = 8192$. By comparing our method to alternative approaches that either use all point features or employ score-based sampling, we demonstrated that P-FPS balances feature extraction and point diversity, avoiding the pitfalls of oversampling foreground points at the cost of diversity. This balance is particularly important for maintaining high detection accuracy across diverse and complex scenes.

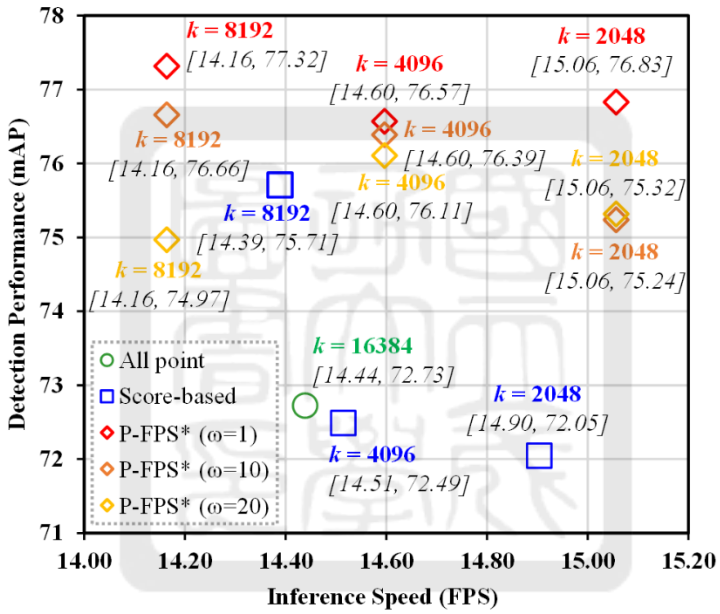


Figure 17. Model performance at varying ω values and feature point counts k.

4) Inference Speed Analysis: The inference efficiency of the SD-AFDet framework is influenced by both the feature sampling strategy and the number of sampled points. As shown in Figure 17, increasing the number of points leads to a decrease in inference speed, illustrating a trade-off between processing time and detection accuracy. For example, when 8,192 points were used during the refinement stage, the model achieved an inference speed of 14.16 FPS on a single NVIDIA RTX 3090 GPU. Although increasing the number of points

improves detection accuracy, it also results in higher latency, potentially affecting real-time performance. Nevertheless, the SD-AFDet framework maintained a satisfactory inference time, striking a balance between accuracy and speed.

Table 10. Ablation study on the components of detection head.

Method	Mod. (AP%)		
	Car_{3D}	$Pedestrian_{3D}$	$Cyclist_{3D}$
Original cloud point	82.36	55.26	75.16
Artificial point	**83.02**	58.98	73.34
Artificial point + Shift	82.59	61.11	77.94
DAPSF	82.89	**63.63**	**79.14**

Table 11. Evaluation of different depth ranges.

Method	AP_{3D} (%)					
	0m-30m			30m-70m		
	Car	Ped.	Cyc.	Car	Ped.	Cyc.
Voxel-based methods						
PointPillar [71]	84.6	56.1	64.2	42.0	7.1	35.1
SECOND [77]	87.7	50.8	66.1	45.5	13.3	38.3
VoxelRCNN [78]	89.0	54.0	70.4	50.8	12.4	34.8
Point-voxel-based methods						
PVRCNN [79]	88.8	56.8	69.3	51.1	7.4	36.1
EQ-PVRCNN [80]	**90.6**	57.9	72.5	52.7	7.1	40.3
Point-based methods						
PointRCNN [57]	87.3	51.6	69.4	43.5	5.2	35.7
EPNet++ [27]	88.9	**62.5**	67.1	51.7	13.9	36.0
Ours w/o DAPSF	87.4	57.8	73.3	49.1	9.4	35.1
Ours	88.0	61.9	**76.7**	**52.9**	**14.7**	**44.0**

5) Effectiveness of DAPSF: The DAPSF strategy was introduced to improve the detection of small and distant objects. As detailed in Table 10, DAPSF led to significant improvements, achieving mAP gains of 2.13% for cyclists and 4.60% for pedestrians at the moderate difficulty level. Furthermore, the integration of artificial point sampling improved the utilization of image features, especially for distant or small objects, leading to additional mAP gains of 2.52% for pedestrians and 1.20% for cyclists. This demonstrates the

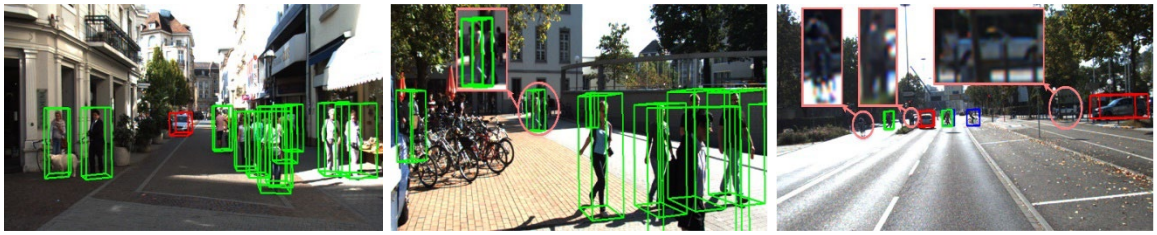
effectiveness of combining point shifting with artificial point sampling to enhance detection accuracy for challenging objects.

6) Evaluation by Distance: To further evaluate the SD-AFDet framework's performance at different distances, the dataset was divided into near distance (0m–30m) and medium-to-far distance (30m–70m) subsets. As shown in Table 11, SD-AFDet outperformed SOTA methods in detecting objects at greater distances, particularly for cars, pedestrians, and cyclists. The DAPSF strategy significantly enhanced detection accuracy for distant objects, where point cloud sparsity is a major challenge. Additional experiments were conducted to evaluate the model's performance with and without DAPSF. When DAPSF was removed, there was a notable decline in a AP across all distance ranges, highlighting its importance. For small objects such as cyclists and pedestrians at medium-to-far distances, the AP increased by 8.9% and 5.3%, respectively, when DAPSF was used.

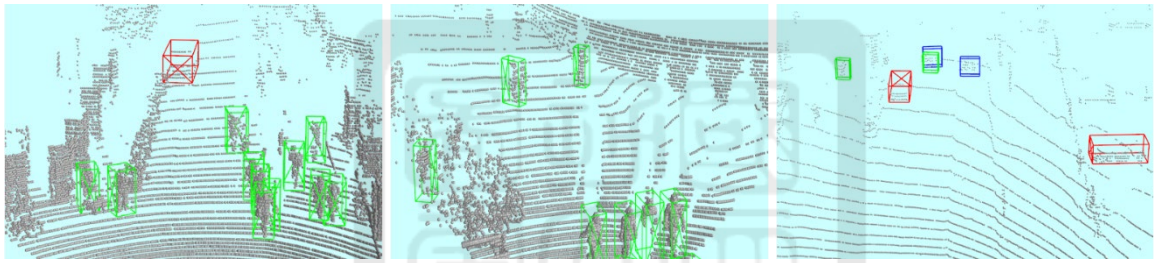
7) Visualization analysis: To assess the performance of the SD-AFDet framework, we conducted a 3D visualization analysis of its detection results. As illustrated in Figure 18, the first row (a) displays the visualization of 3D bounding boxes overlaid on the image, while rows (b)–(d) depict the detection results in 3D space. The visualizations compare the ground truth (GT) (b), the predictions of the baseline model (c), and the predictions generated by the SD-AFDet model (d). In various scenarios, the SD-AFDet framework demonstrated superior performance, particularly in detecting medium-to-far distance objects and small objects, outperforming the baseline model.

In the leftmost column of Figure 18, SD-AFDet successfully detects small objects that were missed by the baseline model. In the middle column, the model identifies pedestrians at medium-to-far distances, which the baseline failed to capture. In the right column, SD-AFDet accurately detects objects within the 30m–70m range, missing only one pedestrian, while correctly identifying a barely visible car. The missed detection was likely caused by

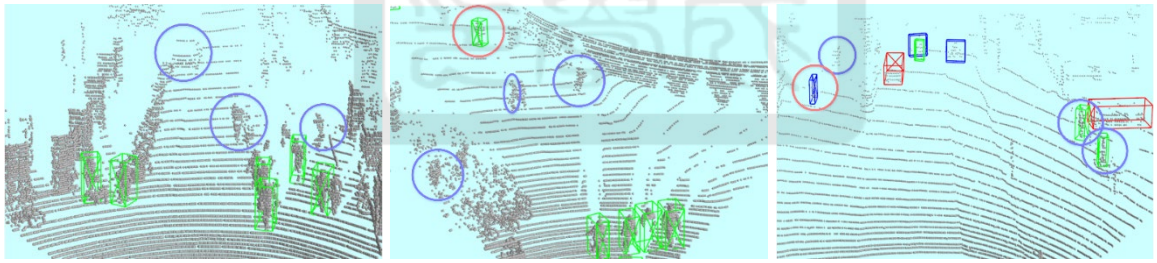
occlusion, where overlapping features from a cyclist interfered with the pedestrian detection in the camera stream. Despite this, SD-AFDet consistently demonstrated strong performance in detecting medium-to-far distance and small objects in challenging real-world environments.



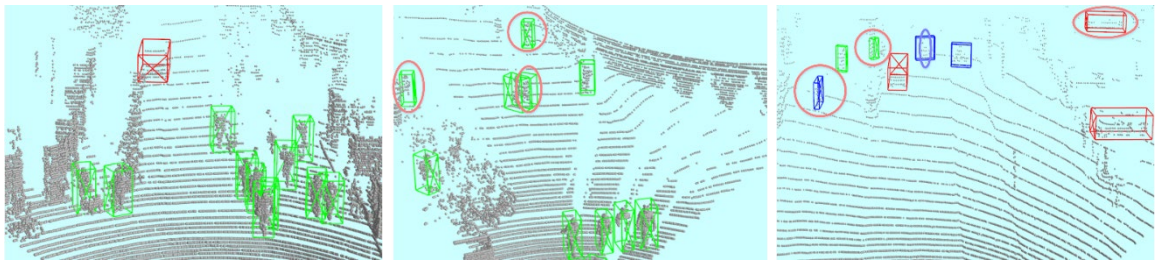
(a) GT 2D Images



(b) GT Lidar



(c) PointRCNN (Baseline) [57]



(d) SD-AFDet

Figure 18. Detection results visualization on the KITTI validation set. Pedestrians, cyclists,

and cars are shown using green, brown, and blue 3D bounding boxes, respectively. Failure cases are highlighted with blue circles, while "DON'T CARE" objects are marked with red circles.

CHAPTER 6. CONCLUSION

This dissertation has presented two innovative multi-modal sensor fusion frameworks, VL-ACFDet and SD-AFDet, aimed at enhancing 2D and 3D object detection capabilities for autonomous driving systems. VL-ACFDet leverages visible and thermal sensor data to improve 2D object detection, while SD-AFDet integrates LiDAR and visible data to achieve more stable and accurate 3D object detection. Comprehensive experiments conducted on various datasets have demonstrated the superior performance of these frameworks in terms of detection accuracy and real-time processing, contributing to more robust perception systems for autonomous vehicles.

The contributions of this research lie in addressing key challenges in autonomous driving perception, such as the limitations of single-sensor approaches and the need for efficient multi-modal integration. By effectively combining data from heterogeneous sensors, the proposed frameworks offer improved robustness in challenging environments, such as low-light conditions, adverse weather, and complex urban scenarios. These advancements lay a foundation for future developments in autonomous driving technology, with the ultimate goal of achieving safer, more reliable, and efficient transportation systems.

Looking ahead, Chapter 7 discusses potential future research directions that build upon the contributions of this work. These directions include extending fusion techniques to more complex perception tasks, exploring applications in intelligent surveillance, and leveraging emerging sensor technologies for enhanced perception. These proposed areas of future work aim to further expand the applicability and effectiveness of the multi-modal fusion frameworks presented in this dissertation.

CHAPTER 7. FEATURE WORKS

This dissertation presents a multi-modal sensor fusion framework for object detection based on the VL-ACFDet and SD-AFDet models. VL-ACFDet focuses on fusing visible and thermal camera data to achieve efficient 2D object detection, while SD-AFDet is designed for combining visible camera and LiDAR data to achieve stable 3D object detection through the integration of 3D point cloud and 2D image information. Possible future research directions are outlined below:

1) Extending Fusion Techniques to Complex and Diverse Perception Tasks: The proposed multi-modal sensor fusion frameworks efficiently utilize information from different modalities, achieving complementary advantages for 2D and 3D tasks. However, object detection serves only as the foundation for perception. Depending on application requirements, future work could explore tasks such as 2D pixel-level and 3D point-level object segmentation, multi-object tracking, or real-time decision-making. For instance, integrating segmentation capabilities would enable a more detailed understanding of the environment, facilitating the identification of object boundaries at both pixel and point levels [81], [82]. This would be particularly valuable in urban environments with dense traffic or complex pedestrian interactions, where precise localization and classification are crucial.

Extending the frameworks to support multi-object tracking could further enhance system awareness of dynamic environments, enabling more reliable predictions of object trajectories and behaviors [83], [84]. Moreover, real-time decision-making capabilities could be developed by integrating these perception modules with planning and control systems, allowing for more autonomous and adaptive responses to rapidly changing conditions [85]. Leveraging the proposed fusion techniques, these complex tasks could benefit from richer multi-modal feature representations, ultimately leading to more robust and versatile

autonomous systems.

2) Enhancing Smart Surveillance with VL-ACFDet for Multi-Modal Fusion: Beyond dynamic autonomous driving perception, static intelligent surveillance systems also represent a promising area of application for the proposed multi-modal fusion frameworks. In addition to public safety, there is increasing demand for effective, efficient, and reliable surveillance solutions for mission-critical, delay-sensitive tasks, such as battlefield monitoring and disaster response [86]. Current video surveillance systems, which primarily use visible cameras [87], are challenged by environmental factors like nighttime conditions, foggy weather, rain, or occlusions [88]. Consequently, sensor fusion approaches that combine visible and thermal cameras are gaining attention for object detection as well as image synthesis to provide better visibility across various conditions [89], [90]. Many existing fusion architectures struggle to accommodate both object detection and image synthesis due to the challenges of balancing the advantages of each modality under diverse environmental conditions [91]. The proposed VL-ACFDet, with its use of high-dimensional semantic information from vision-language models, offers a promising solution for extracting essential information from both modalities, making it suitable for both object detection and image fusion tasks. This capability positions VL-ACFDet as a powerful tool for addressing challenges in current surveillance systems, offering improved robustness and clarity in diverse conditions.

3) Leveraging SD-AFDet with Emerging Sensors for Superior Perception: For 3D object detection, LiDAR and visible cameras are currently among the most widely used sensors. However, the high cost of LiDAR restricts its use in consumer vehicles, and cameras are susceptible to challenging lighting and weather conditions [92], [93]. Recently, the emergence of 4D imaging radars has drawn attention from automakers [94]. Unlike conventional automotive radars, 4D imaging radars provide height information, maintain

stability under adverse conditions, and generate higher-resolution 3D point clouds. Although these point clouds are less dense compared to LiDAR, studies have shown the feasibility of 4D radar-based detection [95]–[97]. Given that the proposed SD-AFDet framework is particularly effective in handling sparse point clouds, future research could focus on leveraging 4D imaging radars to develop an all-weather, efficient, real-time, and cost-effective perception system. By utilizing these emerging sensors, future systems could achieve superior perception performance, paving the way for practical and scalable autonomous driving solutions.

REFERENCE

[1] X. Wang, K. Li and A. Chehri, "Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1148–1165, Feb. 2024.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby and A. Mouzakitis, "A Survey on 3D Object Detection Methods for Autonomous Driving Applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[3] D. Feng et al., "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[4] E. Yurtsever, J. Lambert, A. Carballo and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58443–58469, Mar. 2020.

[5] A. Singh, "Vision-RADAR fusion for Robotics BEV Detections: A Survey," *IEEE Intelligent Vehicles Symposium*, 2023, pp. 1–7.

[6] Y. -C. Chen, S. -Y. Jhong, and C. -H. Hsia, "Roadside Unit-based Unknown Object Detection in Adverse Weather Conditions for Smart Internet of Vehicles," *ACM Transactions on Management Information Systems*, vol. 13, no. 4, pp. 47–67, Jan. 2023.

[7] S. -Y. Jhong, Y. -Y. Chen, C. -H. Hsia, Y. -Q. Wang and C. -F. Lai, "Density-Aware and Semantic-Guided Fusion for 3D Object Detection using LiDAR-Camera Sensors," *IEEE Sensors Journal*, vol. 23, no. 18, pp. 22051–22063, Sep. 2023.

[8] X. Ma, W. Ouyang, A. Simonelli and E. Ricci, "3D Object Detection from Images for Autonomous Driving: A Survey," *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, vol. 46, no. 5, pp. 3537–3556, May 2024.

[9] Á. Takács, D. A. Drexler, P. Galambos, I. J. Rudas and T. Haidegger, "Assessment and Standardization of Autonomous Vehicles," *International Conference on Intelligent Engineering Systems*, 2018, pp. 185–192.

[10] Y. -Y. Chen, S. -Y. Jhong and Y. -J. Lo, "Reinforcement-and-Alignment Multispectral Object Detection Using Visible–Thermal Vision Sensors in Intelligent Vehicles," *IEEE Sensors Journal*, vol. 23, no. 21, pp. 26873–26886, Nov. 2023.

[11] Y. Sun, B. Cao, P. Zhu and Q. Hu, "Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[12] Q. Fang, D. Han and Z. Wang, "Cross-Modality Fusion Transformer for Multispectral Object Detection", *SSRN Electronic Journal*, Sep. 2022.

[13] T. Huang, Z. Liu, X. Chen, and X. Bai, "EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection," *European Conference on Computer Vision*, 2020, pp. 35–52.

[14] X. Pan, Z. Xia, S. Song, L. -E. Li, and G. Huang, "3D Object Detection with Pointformer," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7463–7472.

[15] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao and D. Cao, "CL3D: Camera-LiDAR 3D Object Detection With Point Feature Enhancement and Point-Guided Fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18040–18050, Oct. 2022.

[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11618–11628.

[17] H. Zhang, E. Fromont, S. Lefevre and B. Avignon, “Guided Attentive Feature Fusion for Multispectral Pedestrian Detection,” *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 72–80.

[18] K. Zhou, L. Chen and X. Cao, “Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems,” *European Conference on Computer Vision*, 2020, pp. 787–803.

[19] A. Radford et al., “Learning Transferable Visual Models from Natural Language Supervision,” *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[20] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum PointNets for 3D Object Detection from RGB-D Data,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[21] Z. Wang and K. Jia, “Frustum Convnet: Sliding Frustums to Aggregate Local Point-Wise Features for A Modal 3D Object Detection,” *IEEE International Conference on Intelligent Robots and Systems*, 2019, pp. 1742–1749.

[22] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-View 3D Object Detection Network for Autonomous Driving,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[23] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3D Proposal Generation and Object Detection from View Aggregation,” *IEEE International Conference on Intelligent Robots and System*s, 2018, pp. 1–8.

[24] G. Wang, B. Tian, Y. Zhang, L. Chen, D. Cao, and J. Wu, “Multi-View Adaptive Fusion Network for 3D Object Detection,” *arXiv preprint arXiv:2011.00652*, 2020.

[25] A. Som, H. Choi, K. N. Ramamurthy, M. P. Buman, and P. Turaga, “Pi-Net: A deep learning approach to extract topological persistence images,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 834–835.

[26] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “Pointpainting: sequential fusion for 3D object detection,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4604–4612.

[27] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang and X. Bai, “EPNet++: Cascade Bi-Directional Fusion for Multi-Modal 3D Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8324–8341, Jul. 2023.

[28] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong and Z. Luo, “Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5792–5801.

[29] A. Geiger, P. Lenz and R. Urtasun, “Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[30] Z. Li, P. Xu, X. Chang, L. Yang, Y. Zhang, L. Yao, and X. Chen, “When Object Detection Meets Knowledge Distillation: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10555–10579.

[31] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[32] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.

[33] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A Review of YOLO Algorithm Developments,” *Procedia Computer Science*, vol. 199, pp. 1066-1073, 2022.

[34] X. He, C. Tang, X. Zou, and W. Zhang, “Multispectral Object Detection via Cross-

Modal Conflict-Aware Learning," *ACM International Conference on Multimedia*, 2023, pp. 1465–1474.

[35] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan and W. Yang, "ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection," *Pattern Recognition*, vol.145, Jan. 2024.

[36] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical Depth Distribution Network for Monocular 3D Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[37] X. Chen, K. Kundu, Z. Zhang, H. Ma1, S. Fidler, and R. Urtasun, "Monocular 3D Object Detection for Autonomous Driving," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[38] G. Brazil, G. Pons-Moll, X. Liu, and B. Schiele, "Kinematic 3D Object Detection in Monocular Video," *European Conference on Computer Vision*, 2020, pp. 135–152.

[39] P. Li, H. Zhao, P. Liu, and F. Cao, "RTM3D: Real-time Monocular 3D Detection from Object Keypoints for Autonomous Driving," *European Conference on Computer Vision*, 2020, pp. 135–152.

[40] T. Kim, S. Chung, D. Yeom, Y. Yu, H. -G. Kim and Y. -M. Ro, "MSCoTDet: Language-Driven Multi-Modal Fusion for Improved Multispectral Pedestrian Detection," *arXiv preprint arXiv:2403.15209*, 2024.

[41] Y. Xiao, F. Meng, Q. Wu, L. Xu, M. He and H. Li, "GM-DETR: Generalized Multispectral Detection Transformer with Efficient Fusion Encoder for Visible-Infrared Detection," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 5541–5549.

[42] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," *IEEE International Conference on*

Image Processing, Oct. 2020, pp. 276–280.

[43] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-Aware Faster R-CNN For Robust Multispectral Pedestrian Detection," *Pattern Recognition*, vol. 85, pp. 161–171, Jan. 2019.

[44] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, Oct. 2019.

[45] H. Wang et al., "Cross-Modal Oriented Object Detection of UAV Aerial Images Based on Image Feature," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024, Art. no. 5403021.

[46] X. Yang, Y. Qian, H. Zhu, C. Wang and M. Yang, "BAANet: Learning Bi-directional Adaptive Attention Gates for Multispectral Pedestrian Detection," *International Conference on Robotics and Automation*, 2022, pp. 2920–2926.

[47] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3D-CVF: Generating Joint Camera and Lidar Features Using Cross-View Spatial Feature Fusion For 3D Object Detection," *European Conference on Computer Vision*, 2020, pp. 720–736.

[48] X. Zhang, S. -Y. Cao, F. Wang, R. Zhang, Z. Wu, X. Zhang, X. Bai and H. -L. Shen, "Rethinking Early-Fusion Strategies for Improved Multispectral Object Detection," *arXiv preprint arXiv:2405.16038*, 2024.

[49] M. Yuan and X. Wei, "C²Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[50] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention

Module," *European Conference on Computer Vision*, 2018, pp. 3–19.

[52] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[53] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[54] A. Tripathi, M. K. Gupta, C. Srivastava, P. Dixit and S. K. Pandey, "Object Detection using YOLO: A Survey," *International Conference on Contemporary Computing and Informatics*, 2022, pp. 747-752

[55] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 658–666.

[56] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *European Conference on Computer Vision*, 2018, pp. 801–818.

[57] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

[58] W. Shi and R. R. Rajkumar, "Point-GNN: Graph Neural Network For 3D Object Detection in A Point Cloud," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711–1719.

[59] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-Based 3D Single Stage Object Detector," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11040–11048.

[60] Z. Li, F. Wang, and N. Wang, "Lidar R-CNN: An Efficient and Universal 3D Object Detector," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7546–

7555.

[61] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, “VPFNet: Improving 3D Object Detection with Virtual Point Based LiDAR and Stereo Data Fusion,” *IEEE Transactions on Multimedia*, vol. 25, pp. 5291–5304, Jul. 2022.

[62] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, “PI-RCNN: An Efficient Multi-Sensor 3D Object Detector with Point-Based Attentive Cont-Conv Fusion Module,” *AAAI Conference on Artificial Intelligence*, 2020, pp. 12460–12467.

[63] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng and T. L. Lam, “Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks,” *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 4060–4067, Jul. 2023.

[64] J. Ouyang, P. Jin and Q. Wang, “Multimodal Feature-Guided Pretraining for RGB-T Perception,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 16041–16050, Sep. 2024.

[65] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, “The Mapillary Vistas Dataset for Semantic Understanding Of Street Scenes,” *International Conference on Computer Vision*, 2017, pp. 4990–4999.

[66] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[67] Y. Zhang, C. Xu, W. Yang, G. He, H. Yu, L. Yu, G. -S. Xia, “Drone-Based RGBT Tiny Person Detection,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 61–76, Apr. 2023.

[68] M. He, Q. Wu, KN. Ngan, F. Jiang, F. Meng and L. Xu, “Misaligned RGB-Infrared Object Detection via Adaptive Dual-Discrepancy Calibration,” *Remote Sensing*, vol.15,

no. 19, pp.4887, Jul. 2023.

[69] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, X. Liu, J. Zhang, G. Guo and B. Zhang, "Fusion-Mamba for Cross-modality Object Detection," *arXiv preprint arXiv:2404.09146*, 2024.

[70] C. Chen, Z. Chen, J. Zhang, and D. Tao, "SASA: Semantics-Augmented Set Abstraction for Point-Based 3D Object Detection," *AAAI Conference on Artificial Intelligence*, 2022, pp. 221–229.

[71] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang and O. Beijbom, "PointPillars: Fast Encoders for Object Detection from Point Clouds," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12689–12697.

[72] Q. Tang, X. Bai, J. Guo, B. Pan, and W. Jiang, "DFAF3D: A Dual-Feature-Aware Anchor-Free Single-Stage 3D Detector for Point Clouds," *Image and Vision Computing*, vol. 129, 2023, Art. no. 104594.

[73] S. Shi, Z. Wang, J. Shi, X. Wang and H. Li, "From Points to Parts: 3D Object Detection from Point Cloud with Part-Aware and Part-Aggregation Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2647–2664, Aug. 2021.

[74] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, and Y. Guo, "Not All Points Are Equal: Learning Highly Efficient Point-Based Detectors For 3D Lidar Point Clouds," *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18953–18962.

[75] S. Chen, H. Zhang, and N. Zheng, "Leveraging Anchor-based LiDAR 3D Object Detection via Point Assisted Sample Selection," *arXiv preprint arXiv:2403.01978*, 2024.

[76] I. Lang, A. Manor and S. Avidan, "Samplenet: Differentiable Point Cloud Sampling," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7575–7585.

[77] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely Embedded Convolutional Detection,"

Sensors, vol. 18, no. 10, p. 3337, 2018.

[78] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection," *AAAI Conference on Artificial Intelligence*, 2021, pp. 1201–1209.

[79] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1201–1209.

[80] Z. Yang, L. Jiang, Y. Sun, B. Schiele, and J. Jia, "A Unified Query-based Paradigm for Point Cloud Understanding," *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8531–8541.

[81] Y. Lv, Z. Liu and G. Li, "Context-Aware Interaction Network for RGB-T Semantic Segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6348–6360, Jan. 2024.

[82] J. Li, H. Dai, H. Han and Y. Ding, "MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving," *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21694–21704.

[83] Zhang, D. Yuan, X. Shu, Z. Li, Q. Liu, X. Chang, Z. He, and G. Shi, "A Comprehensive Review of RGBT Tracking," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–23, Jul. 2024.

[84] S. Feng, X. Li, Z. Yan, C. Xia, S. Li, X. Wang, and Y. Zhou, "Tightly Coupled Integration of LiDAR and Vision for 3D Multiobject Tracking," *IEEE Transactions on Intelligent Vehicles*, pp. 1–14, Jun. 2024.

[85] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end Autonomous Driving: Challenges and Frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, Jul. 2024.

[86] S. S. Ahmed, A. Schiessl, F. Gumbmann, M. Tiebout, S. Methfessel, and L.-P. Schmidt,

"Advanced Microwave Imaging," *IEEE microwave magazine*, vol. 13, no. 6, pp. 26–43, 2012.

[87] G. L. Charvat, L. C. Kempel, E. J. Rothwell, C. M. Coleman and E. L. Mokole, "A Through-Dielectric Ultrawideband Switched-Antenna-Array Radar Imaging System," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 11, pp. 5495–5500, Nov. 2012.

[88] N. Poredi, Y. Chen, X. Li and E. Blasch, "Enhance Public Safety Surveillance in Smart Cities by Fusing Optical and Thermal Cameras," *International Conference on Information Fusion*, 2023, pp. 1–7.

[89] X. Yi, H. Xu, H. Zhang, L. Tang and J. Ma, "Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion," *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27016–27025.

[90] X. Li, X. Li, T. Ye, X. Cheng, W. Liu and H. Tan, "Bridging the Gap between Multi-focus and Multi-modal: A Focused Integration Framework for Multi-modal Image Fusion," *IEEE Winter Conference on Applications of Computer Vision*, 2024, pp. 1617–1626.

[91] S. A. Deevi, C. Lee, L. Gan, S. Nagesh, G. Pandey and S. -J. Chung, "RGB-X Object Detection via Scene-Specific Fusion Modules," *IEEE Winter Conference on Applications of Computer Vision*, 2024, pp. 7351–7360.

[92] P. Svenningsson, F. Fioranelli, and A. Yarovoy, "Radar-PointGNN: Graph Based Object Recognition for Unstructured Radar Point-cloud Data," *IEEE Radar Conference*, 2021, pp. 1–6.

[93] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q. -L. Han and B. Zhu, "SMURF: Spatial Multi-Representation Fusion for 3D Object Detection With 4D Imaging Radar," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 799–812, Jan. 2024.

[94] L. Fan, J. Wang, Y. Chang, Y. Li, Y. Wang and D. Cao, “4D mmWave Radar for Autonomous Driving Perception: A Comprehensive Survey,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4606–4620, Apr. 2024.

[95] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij and D. M. Gavrila, “Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, Apr. 2022.

[96] B. Tan, Z. Ma, X. Zhu, S. Li, L. Zheng, S. Chen, L. Huang, and J. Bai, “3-D Object Detection for Multiframe 4-D Automotive Millimeter-Wave Radar Point Cloud,” *IEEE Sensors Journal*, vol. 23, no. 11, pp. 11125–11138, Jun. 2023.

[97] W. Xiong, J. Liu, T. Huang, Q. -L. Han, Y. Xia and B. Zhu, “LXL: LiDAR Excluded Lean 3D Object Detection With 4D Imaging Radar and Camera Fusion,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 79–92, Jan. 2024.

BIOGRAPHY

SIN-YE JHONG

EDUCATION

Visiting scholar. Electrical Engineering — 2024
University of South Florida (USF) — Florida, USA
Advisor: *Ji-En Chang*

Ph.D. Engineering Science — 2019 – 2024
National Cheng Kung University (NCKU) — Tainan, Taiwan
Advisor: *Chin-Feng Lai, Chih-Hsien Hsia*

M.S. Automation Technology — 2017 – 2019
National Taipei University of Technology (NTUT) — Taipei, Taiwan
Advisor: *Yung-Yao Chen*

B.S. Electrical Engineering — 2013 – 2017
Chinese Culture University (CCU) — Taipei, Taiwan
Advisor: *Chih-Hsien Hsia*

HONORS AND AWARDS

Best paper awards:

1. 2024 **Honorable Mention Presentation Paper Award** at International Conference on Advanced Robotics and Intelligent Systems (ARIS)
2. 2023 **Honorable Mention Paper Award** at International Automatic Control Conference (CACS)
3. 2023 **Honorable Mention Paper Award** at International Conference on Advanced Robotics and Intelligent Systems (ARIS)
4. 2022 **Honorable Mention Paper** Award at International Automatic Control Conference (CACS)

5. 2020 **Honorable Mention Paper Award** at International Conference on Advanced Robotics and Intelligent Systems (ARIS)
6. 2020 **Best Paper Award** at Cyberspace
7. 2019 **Best Paper Award** at Cyberspace
8. 2018 **Best Paper Award** at Cyberspace

Competition awards:

1. 2024 **Honorable Mention** in 20th National Electronic Design Creative Competition
2. 2023 **Honorable Mention** in National Industry-Academia Innovation Competition
3. 2023 **Second Prize Award** and **Honorable Mention** in 19th National Electronic Design Creative Competition
4. 2022 **Second Prize Award** and **Popularity Award** in Intel® DevCup Competition
5. 2022 **Second Prize Award** in 18th National Electronic Design Creative Competition
6. 2021 **Third Prize Award** and **Popularity Award** in Intel® DevCup Competition
7. 2021 **First Prize Award** in National Industry-Academia Innovation Competition
8. 2021 **Silver Award** in Information Service Innovation and Entrepreneurship Competition
9. 2021 **Excellent Award** in AIGO: AI Project-Talent Problem Solving Competition
10. 2021 **Gold Medal** in National Quemoy University Artificial Intelligence Innovation Application Competition
11. 2021 **Champion** in 2nd Autonomous Cars (Duckiebot) Thematic Competition
12. 2021 **Second Prize Award** in 17th National Electronic Design Creative Competition
13. 2020 **Special Award** in 16th National Electronic Design Creative Competition
14. 2019 **First Prize Award** in National Industry-Academia Innovation Competition
15. 2019 **Silver Medal** in Hua Nan Financial Holdings FinTech Innovation Competition
16. 2019 **Third Prize Award** in Cross-Strait Youth Innovation and Entrepreneurship Invitational Competition
17. 2019 **Honorable Mention** in Yilan Science Park Smart Industry Information Application Competition

18. 2019 **Silver Award** and **Rising Star Award** in Information Service Innovation and Entrepreneurship Competition
19. 2019 **First Prize Award** and **IEEE Special Award** in 15th National Electronic Design Creative Competition
20. 2018 **First Prize Award** in Jih Sun Bank Hackathon Competition
21. 2018 **Silver Award** and **Rising Star Award** in Information Service Innovation and Entrepreneurship Competition
22. 2018 **First Prize Award** and **Icetron Enterprises Special Award** in 14th National Electronic Design Creative Competition
23. 2018 **First Prize Award** in National Youth Creative Application Competition
24. 2017 **First Prize Award** in CCU College of Engineering Project Competition
25. 2016 **Second Prize Award** in National Industry-Academia Innovation Competition
26. 2016 **First Prize Award** in CCU EE Project Competition
27. 2015 **First Prize Award** in CCU College of Engineering Project Competition

Fellowship:

- 2024 National Science and Technology Council, The Graduate Students Study Abroad Program for "Research on Privacy-Preserving AIoT System for Facial Skin and Scalp Health Inspection", $20,000.

Scholarships:

- 2020 NCKU ES Competition Excellence Scholarship
- 2019 NTUT Competition Excellence Scholarship
- 2018 NTUT Competition Excellence Scholarship
- 2017 CCU Competition Excellence Scholarship
- 2017 CCU Academic Excellence Scholarship
- 2016 CCU Competition Excellence Scholarship
- 2016 CCU Academic Excellence Scholarship
- 2015 CCU Academic Excellence Scholarship

Journal Publications

*Corresponding author

1. **S.-Y. Jhong**, G.-T. Li, and C.-H. Hsia*, "An Edge-Cloud Collaborative Scalp Inspection System Based on Robust Representation Learning," *to appear in IEEE Transactions on Consumer Electronics*. (SCI)
2. C. -Y. Chen*, **S. -Y. Jhong**, and C. -H. Hsia, "A Domain Generalized Face Anti-Spoofing System Using Domain Adversarial Learning," *International Journal of Engineering and Technology Innovation*, vol. 14, no. 4, pp. 378-388, Sep. 2024. (EI, ESCI)
3. Y. -Y. Chen, **S. -Y. Jhong**, S. -K. Tu, Y. -H. Lin* and Y. -C. Wu, "Autonomous Smart-Edge Fault Diagnostics via Edge-Cloud-Orchestrated Collaborative Computing for Infrared Electrical Equipment Images," *IEEE Sensors Journal*, vol. 24, no. 15, pp. 24630-24648, Aug. 2024. (SCI)
4. **S.-Y. Jhong**, Y.-Y. Chen, C.-H. Hsia*, and C.-F. Lai, "iVehicles: Spatial Feature Aggregation Network for Lane Detection," *IEEE Sensors Letters*, vol. 8, no. 4, pp. 1-4, Apr. 2024. (EI, ESCI)
5. Y. -Y. Chen* and **S. -Y. Jhong**, "Multilevel Self-Training Approach for Cross-Domain Semantic Segmentation in Intelligent Vehicles," *IEEE Intelligent Transportation Systems Magazine*, vol. 16, no. 1, pp. 148-161, Jan.-Feb. 2024. (SCI)
6. Y. -Y. Chen*, **S. -Y. Jhong**, and Y. -J. Lo, "Reinforcement-and-Alignment Multispectral Object Detection Using Visible–Thermal Vision Sensors in Intelligent Vehicles," *IEEE Sensors Journal*, vol. 23, no. 21, pp. 26873-26886, Nov. 2023. (SCI)
7. Y. -Y. Chen, C. -H. Hsia*, **S. -Y. Jhong***, and C. -F. Lai, "Attention-Guided HDR Reconstruction for Enhancing Smart City Applications," *Electronics*, vol. 12, no. 22, p. 4625, Nov. 2023. (SCI)
8. **S. -Y. Jhong**, Y. -Y. Chen, C. -H. Hsia, Y. -Q. Wang and C. -F. Lai*, "Density-Aware and Semantic-Guided Fusion for 3D Object Detection using LiDAR-Camera Sensors," *IEEE Sensors Journal*, vol. 23, no. 18, pp. 22051-22063, Sep. 2023. (SCI)
9. Y. -C. Chen, H. -C. Lin, H. -W. Hwang, K. -L. Hua, Y. -L. Hsu, and **S. -Y. Jhong***, "An Edge Lidar-Based Detection Method in Intelligent Transportation System," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 4, Aug. 2023. (EI, ESCI)

10. Y. -C. Chen, **S. -Y. Jhong**, and C. -H. Hsia*, "Roadside Unit-based Unknown Object Detection in Adverse Weather Conditions for Smart Internet of Vehicles," *ACM Transactions on Management Information Systems*, vol. 13, no. 4, pp. 47-67, Jan. 2023. (EI, ESCI)
11. Y. -Y. Chen, Y. -H. Lin*, Y. -C. Hu, C. -H. Hsia, Y. -A. Lian, and **S. -Y. Jhong**, "Distributed Real-Time Object Detection Based on Edge-Cloud Collaboration for Smart Video Surveillance Applications," *IEEE Access*, vol. 10, pp. 93745-93759, Aug. 2022. (SCI)
12. **S. -Y. Jhong**, P. -Y. Yang, and C. -H. Hsia*, "An Expert Inspection System Using Deep Learning for Smart Scalp," *Sensors and Materials*, vol. 34, pp. 1265-1274, Apr. 2022. (SCI)
13. Y. -Y. Chen, **S. -Y. Jhong**, C. -H. Hsia*, and H. -L Hua, "Explainable AI: A Multispectral Palm-Vein Identification System with New Augmentation Features," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 35, pp. 112-132, Nov. 2021. (SCI)
14. **S. -Y. Jhong**, Y. -Y. Chen, C. -H. Hsia, S. -C. Lin, K. -H. Hsu, and C.-F Lai*, "Nighttime Object Detection System with Lightweight Deep Network for Internet of Vehicles," *Journal of Real-Time Image Processing*, vol. 18, pp. 1141-1155, Aug. 2021. (SCI)
15. Z. -C. Chen, **S. -Y. Jhong**, and C. -H. Hsia*, "Design of A Lightweight Palm-vein Authentication System Based on Model Compression," *Journal of Information Science and Engineering*, vol. 37, pp. 809-825, Jul. 2021. (SCI)
16. Y. -Y. Chen*, G. -Y. Li, **S. -Y. Jhong**, P. -H. Chen, C. -C. Tsai, and P. -H. Chen, "Nighttime Pedestrian Detection Based on Thermal Imaging," *Sensors and Materials*, vol. 32, pp. 3157-3167, Oct. 2020. (SCI)
17. Y. -Y. Chen*, C. -H. Hsia, **S. -Y. Jhong**, and H. -J. Lin, "Data Hiding Method for AMBTC Compressed Images," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 1-9, Nov. 2018 (SCI)

Conference Papers

1. Z. -K. Hu, **S. -Y. Jhong**, H. -W. Hwang, S. -H. Lin, K. -L. Hua and Y. -Y. Chen, "Bi-Directional Bird's-Eye View Features Fusion for 3D Multimodal Object Detection and Tracking," *International Automatic Control Conference (CACS)* ***Honorable Mention Paper Award***, 2023, pp. 1-6. (EI)

2. **S. -Y. Jhong**, C. -H. Ko, Y. -F. Su, K. -L. Hua, and Y. -Y. Chen*, “Efficient Lane Detection based on Feature Aggregation for Advanced Driver Assistance Systems,” *International Conference on Advanced Robotics and Intelligent Systems (ARIS)* ***Honorable Mention Paper Award***, 2023, pp. 1-6. (EI)
3. **S. -Y. Jhong**, H. -C. Lin, X. -X. Weng, T. -F. Xie, H. -W. Lin and Y. -Y. Chen, “A Novel Network Architecture and Training Strategies for Camera-Radar 3D Detection,” *International Conference on Consumer Electronics – Taiwan (ICCE-TW)*, 2023, pp. 411-412. (EI)
4. H. -T. Chan, Y. -W. Liao, **S. -Y. Jhong**, S. -C. Chien, K. -L. Hua, and Y. -Y. Chen, “A Skin Type Classification Method Using Mobile Device-Based Deep Learning Model,” *International Conference on Applied System Innovation (ICASI)*, 2023, pp. 199-201. (EI)
5. **S. -Y. Jhong**, Y. -Q. Wang, W. -J. Cheng, H. -W. Hwang, and Y. -Y. Chen, “LiDAR-Based Pedestrian Detection Using Multiple Features and Dimensionality Reduction Scheme,” *International Conference on System Science and Engineering (ICSSE)*, 2022, pp. 7-10. (EI)
6. S. -Y. Lu, **S. -Y. Jhong**, W. -J. Cheng, and Y. -Y. Chen, “Fast Learning-Based 3-D Lidar Localization with Multiscale Feature Recursive Matching for Autonomous Driving,” *International Automatic Control Conference (CACS)* ***Honorable Mention Paper Award***, 2022, pp. 1-6. (EI)
7. Y. -M. Zhang, S. -W. Lin, T. -H. Chou, **S. -Y. Jhong**, and Y. Chen, “Robust Lane Detection via Filter Estimator and Data Augmentation,” *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2022, pp. 291-292. (EI)
8. P. -H. Chen, **S. -Y. Jhong**, and C. -H. Hsia, “Semi-Supervised Learning with Attention-Based CNN for Classification of Coffee Beans Defect,” *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2022, pp. 411-412. (EI)
9. **S. -Y. Jhong**, P. -Y. Yang, and C. -H. Hsia, “An Attention based Expert Inspection System for Smart Scalp," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1678-1681. (EI)
10. P. -Y. Yang, **S. -Y. Jhong**, and C. -H. Hsia, “Green Coffee Beans Classification Using Attention-Based Features and Knowledge Transfer,” *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021, pp. 1-2. (EI)

11. Y. -Q. Wang, P. -H. Chen, **S. -Y. Jhong**, K. -M. Yen, and Y. -Y. Chen, "Forward Vehicle Detection Based on Thermal Vision and Convolutional Neural Network for Autonomous Vehicles," *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021, pp. 1-2. (EI)
12. **S. -Y. Jhong**, P. -Y. Tseng, N. Siriphockpirom, C. -H. Hsia, M. -S. Huang, K. -L. Hua, and Y. -Y. Chen, "An Automated Biometic Identification System Using CNN-based Palm Vein Recognition," *International Conference on Advanced Robotics and Intelligent Systems (ARIS)* ***Honorable Mention Paper Award***, 2020, pp. 1-6 (EI)
13. Q. Sun, **S.-Y. Jhong**, C.-H. Hsia, and C. Yu, "Online Social Media Interaction and Offline Protest Movement: Patterns in 2019 Hong Kong," *Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, 2020, pp. 254-259. (EI)
14. Y. -Y. Chen, **S. -Y. Jhong**, G. -Y. Li, and P. -H. Chen, "Thermal-Based Pedestrian Detection Using Faster R-CNN and Region Decomposition Branch," *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2019, pp. 1-2. (EI)

PATENTS

1. **S. -Y. Jhong**, C. -H. Hsia, "Intelligent Dandruff Detection System and Method," *US Patent*, 12086983B2, 2024.
2. **S. -Y. Jhong**, C. -H. Hsia, and C. -W. Chen, "頭髮特徵分析方法與系統," *R.O.C. Patent*, I841402, 2023.
3. **S. -Y. Jhong**, and C. -H. Hsia, "智慧頭皮屑檢測系統與方法," *R.O.C. Patent*, I823143, 2023.
4. K. -L. Hua, Y. -Y. Chen, **S. -Y. Jhong**, Y. -C. Chen, B. -L. Lin, T. -Y. Lin, C. -S. Wen, Y. -P. Wag, C. -J. Chen, T. -S. Yang, W. -H. Lu, and C. -J. Huang, "Method and System for Detecting and Analyzing Objects," *US Patent*, 11663832B2, 2023.
5. K. -L. Hua, Y. -Y. Chen, **S. -Y. Jhong**, Y. -C. Chen, B. -L. Lin, T. -Y. Lin, C. -S. Wen, Y. -P. Wag, C. -J. Chen, T. -S. Yang, W. -H. Lu, and C. -J. Huang, "偵測物件並標記距離的方法與系統," *R.O.C. Patent*, I797596, 2023.
6. K. -L. Hua, Y. -Y. Chen, **S. -Y. Jhong**, Y. -C. Chen, B. -L. Lin, T. -Y. Lin, C. -S. Wen, Y. -P. Wag, C. -J. Chen, T. -S. Yang, W. -H. Lu, and C. -J. Huang, "雙影像融合方法與裝置," *R.O.C. Patent*, I768709, 2022.

7. K. -L. Hua, Y. -Y. Chen, **S. -Y. Jhong**, Y. -C. Chen, B. -L. Lin, T. -Y. Lin, C. -S. Wen, Y. -P. Wag, C. -J. Chen, T. -S. Yang, W. -H. Lu, and C. -J. Huang, "影像物件辨識模型的訓練方法及影像物件辨識模型," *R.O.C. Patent*, I759156, 2022.
8. Y.-Y. Chen, X.-J. Peng, **S.-Y. Jhong**, and B.-Y. Chen, "利用權重參數與餘數定義隱寫資料於區塊截斷編碼影像的方法、影像壓縮裝置及電腦可讀取的記錄媒體," *R.O.C. Patent*, I643160, 2018.