A STUDY ON SEMANTIC UNDERSTANDING FOR AUTONOMOUS DRIVING

SHIYU MENG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering

A Study on Semantic Understanding for Autonomous Driving

Shiyu Meng

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

July 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Meng Shiyu
_____ (Name of student)

Abstract

Reliable and explainable autonomous driving systems must simultaneously achieve accurate perception, robust localization, and trustworthy decision-making in complex and dynamic environments. These core capabilities are essential not only for ensuring driving safety and efficiency, but also for enhancing user trust and system transparency. Semantic understanding serves as the bridge between perception and cognition, empowering autonomous systems to infer dynamic entities, contextual dependencies, and holistic scene semantics beyond mere sensory interpretation. This thesis presents a series of progressive contributions that advance semantic understanding in autonomous driving through innovations in BEV perception, multi-modal fusion, interpretable decision-making, and cross-modality place recognition.

We begin by addressing the need for dense BEV moving-obstacle segmentation using cost-effective visual sensors. Among various BEV perception tasks, moving-obstacle segmentation is very important, since it can provide necessary information for downstream tasks, such as motion planning and decision making. In general, existing LiDAR-based methods often suffer from sparsity and hardware cost limitations. To this end, we propose a semantics-assisted segmentation framework that utilizes multi-camera visual inputs and temporal semantic cues to

generate dense BEV maps of moving obstacles, enabling vision-based dynamic perception without relying on 3-D LiDAR information.

To further improve segmentation performance in challenging scenarios, we extend this effort with a BEV multi-modal moving-obstacle segmentation framework. Recognizing the complementary strengths of LiDAR and image-based depth estimation, we introduce DPMoSeg, a novel architecture that integrates sparse 3-D point clouds to generate dense depth information through a sparse-dense attention mechanism. Therefore, DPMoSeg produces more accurate and complete BEV segmentation results. Our hybrid design bridges the gap between low-cost visual sensors and high-fidelity geometric cues.

Further, we explore interpretable decision-making to improve the transparency of autonomous driving behavior. While many learning-based solutions offer accurate performance for vehicle behavior, they often lack human-oriented explanations, reducing user trust and hindering widespread adoption. To resolve this, we propose a unified framework that couples vehicle behavior prediction with natural language-based interpretation. This is achieved via a self-supervised, class-agnostic object segmentor and semantic-aware fusion, enabling decision outputs that are both effective and explainable, without requiring extra annotations.

The final part of the thesis addresses the challenge of cross-modal place recognition, which is vital for localization in GPS-denied or degraded conditions. We propose a unified framework that matches real-time RGB images with pre-built LiDAR maps by transforming point clouds into range-view images. A Transformer-Mamba Mixer module is designed to model both intra-modal and inter-modal dependencies. Furthermore, a semantic-promoted descriptor enhancer is introduced to embed high-level scene context. The framework is trained under a contrastive

learning paradigm to optimize cross-modal similarity learning. Experimental results on multiple benchmarks demonstrate its competitive performance against state-of-the-art methods.

In summary, this thesis presents a set of novel and practical frameworks that address key challenges in perception, decision-making, and localization for autonomous driving. By leveraging visual information, semantic understanding, and multi-modal integration, our methods contribute to the development of more cost-effective and robust autonomous systems.

Keywords: BEV Perception, Obstacle Segmentation, Deep Learning, Place Recognition, Cross modality, Explainable Study, Decision-Making, Autonomous Driving

Publications

1. Meng, Shiyu, Yi Wang, Huaiyuan Xu, and Lap-Pui Chau. "Contrastive learning-based place descriptor representation for cross-modality place recognition." Information Fusion (2025): 103351.

2. Meng, Shiyu, Yi Wang, Yawen Cui, and Lap-Pui Chau. "Foundation model-assisted interpretable vehicle behavior decision making." Knowledge-Based Systems (2025): 113868.

3. Meng, Shiyu, and Yuxiang Sun. "Semantic-MoSeg: Semantics-Assisted Moving-Obstacle Segmentation in Bird-Eye-View for Autonomous Driving." IEEE Transactions on Intelligent Transportation Systems 26, no. 7 (2025): 9251-9262.

4. Meng, Shiyu, Yi Wang, and Lap-Pui Chau. "Depth-powered Moving-obstacle Segmentation Under Bird-eye-view for Autonomous Driving." In 2024 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5. IEEE, 2024.

5. Xu, Huaiyuan, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. "A survey on occupancy perception for autonomous driving: The informa-

tion fusion perspective." Information Fusion 114 (2025): 102671.

6. Xu, Huaiyuan, Huaping Liu, Shiyu Meng, and Yuxiang Sun. "A novel place recognition network using visual sequences and LiDAR point clouds for autonomous vehicles." In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 2862-2867. IEEE, 2023.

Acknowledgements

I would like to take this opportunity to express my deep and heartfelt gratitude to everyone who has supported and guided me during my graduate studies and the completion of this thesis.

First and foremost, I am deeply indebted to Professor Chau Lap-Pui, my chief supervisor, for his exceptional guidance, unwavering support, and constant encouragement throughout my graduate studies. His profound knowledge, visionary insight, and high standards of academic excellence have been a continual source of inspiration. I am especially grateful for his patience, his trust in my work, and his thoughtful mentorship, all of which have played a pivotal role in shaping my academic journey and personal growth. I would also like to thank Professor Wang Yi and Professor Sun Yuxiang for their generous support, insightful comments, and kind assistance throughout the study. They provided thoughtful feedback on my manuscripts and offered valuable suggestions on the design and presentation of experiments. Their involvement has been a great encouragement, and I am truly grateful for their help.

My heartfelt appreciation goes to my family, whose unconditional love and encouragement have been the foundation of all my accomplishments. To my parents, thank you for your endless support, sacrifices, and for always believing in

me. Your strength and devotion have been a constant source of motivation.

I would also like to express my deep appreciation to all the members of the laboratory, who have provided a stimulating and friendly environment in which to learn and grow. I am thankful for the many meaningful discussions, the collaborative spirit, and the willingness to share knowledge and ideas. To specific, I would like to acknowledge Feng Yuchao, Ma WeiXin, Deng Kunyuan, Yao Lei, Xu Huaiyuan, Cui Yawen, Su Yuejiao, Zhang Yi, Chen Junliang, Pan Jianhong, Lian Jiawei, Wang Xiaoqi, and Ruan Jianyuan, whose companionship and intellectual input have made my research journey both productive and enjoyable.

I would also like to extend my sincere thanks to the members of my thesis committee for their valuable time, constructive criticism, and insightful suggestions, which have greatly improved the quality and depth of this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Autonomous driving has undergone rapid development in recent years, showing great promise in transforming modern transportation systems by reducing traffic accidents, improving travel efficiency, and enabling mobility for all. Fueled by breakthroughs in deep learning and large-scale annotated datasets, autonomous vehicles (AVs) have achieved remarkable progress. This progress encompasses several critical tasks, including perception, planning, decision-making, and control, as shown in Fig. 1.1. End-to-end learning frameworks, advanced neural architectures, and powerful backbone models pretrained on large-scale vision and language corpora have significantly enhanced the capabilities of AV systems in complex driving scenarios.

Despite these advancements, achieving safe, robust, and generalizable intelligent vehicles in complex and dynamic real-world environments remains a grand challenge. Traditional modular pipelines are increasingly being replaced or aug-

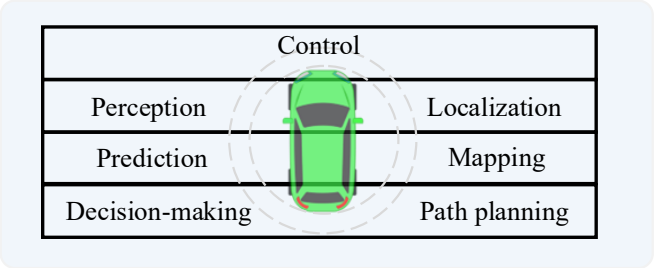

Figure 1.1: The key tasks in autonomous driving systems

mented by end-to-end approaches. Modern autonomous driving systems must go beyond surface-level perception and develop a deep semantic understanding of ego-vehicle surroundings. Specifically, they must be capable of perceiving dynamic agents, reasoning about temporal changes in the environment, making context-aware and interpretable decisions, and localizing themselves accurately and robustly under diverse and adverse conditions. These requirements demand multi-task, multi-modal, and end-to-end frameworks that are data-efficient, explainable, and scalable. It is essential to explore end-to-end semantic understanding for autonomous driving.

The first challenge tackled in this thesis is moving-obstacle segmentation under bird's-eye view (BEV). Moving obstacles such as vehicles and pedestrians pose greater risk to safety than static ones, and detecting them accurately is a prerequisite for downstream tasks such as motion planning and collision avoidance. Most of the recent research based on camera front-view focuses on the segmentation of the state of moving vehicles. Among those solutions, multiple RGB images, optical flow, or depth images with continuous timestamps are usually as the input to solve the segmentation of moving objects task. However, for the front view part, there are object occlusions in most cases, which add to the burden of immediate

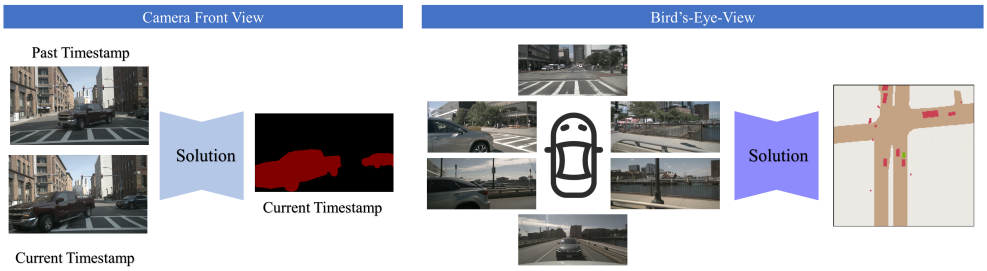

Figure 1.2: Camera-front view vs bird's-eye-view. The left figure shows 2 front-view images with continuous timestamps, and the right figure shows an example of the moving-obstacle segmentation task under BEV.

discrimination, as shown in Fig. 1.2. While LiDAR-based methods can project 3D segmentation results into BEV, the post-projected outputs are often sparse due to the inherent limitations of point clouds. To address these issues, we first propose a semantics-assisted BEV moving-obstacle segmentation network that solely relies on cost-effective visual cameras. Our approach leverages temporal visual semantics from multiple surrounding views to generate dense and usable BEV representations in real time, thereby reducing hardware dependencies and enhancing scalability.

Despite the advantages of using cameras, purely vision-based methods often suffer from the lack of explicit depth information, which is crucial for accurate localization and reasoning of obstacle. Therefore, this motivates our second contribution: a dense depth-powered multi-modal segmentation framework that combines the complementary strengths of LiDAR and camera inputs. By introducing a novel sparse-dense attention mechanism, we effectively fuse geometric and semantic cues to produce more complete and reliable BEV segmentation maps. This multi-modal solution bridges the depth perception gap in vision-only methods, delivering both precision and density in dynamic obstacle understanding.

Building upon robust perception capabilities, the third part of this thesis addresses the need for trustworthy and interpretable decision-making in autonomous driving. Current end-to-end learning paradigms for behavior prediction often operate as black boxes, making it difficult to understand or verify the systems decisions. We propose a novel unified framework that not only predicts driving behavior from egocentric visual inputs but also simultaneously generates human-centric natural language interpretations. To this end, we design a self-supervised, class-agnostic object segmentor and a semantic extractor for the framework whose fused output is processed via a self-attention mechanism to obtain globally contextualized features. So, the dual-output design improves both the transparency and reliability of autonomous decisions without requiring additional supervision.

Lastly, we address the challenge of robust and scalable localization through cross-modality place recognition. Place recognition either relies on visual appearance which is sensitive to environmental variations, or on LiDAR geometry, which requires expensive hardware. We present a cross-modal retrieval approach that aligns acquired RGB images with previously generated LiDAR point cloud maps. A unified place descriptor is learned via a Siamese framework, where visual images are aligned with LiDAR representations transformed into range-view form. To overcome modality discrepancies, we introduce a Transformer-Mamba Mixer and a semantic-promoted descriptor enhancer, enabling fine-grained metric learning. Our contrastive learning strategy facilitates robust place recognition across modalities, enhancing global localization under diverse and changing conditions.

In summary, this thesis presents a series of novel, interconnected solutions aimed at enhancing semantic understanding in autonomous driving. From dense moving-obstacle segmentation to cross-modal localization and interpretable be-

havior prediction, our work collectively contributes to the development of safe, efficient, and explainable autonomous driving systems.

1.2 Structure of the Thesis

The chapters of the thesis are organized as follows: Chapter 1 gives the introduction, background of the research, and presents the structure of the whole thesis.

Chapter 2 shows a literature review of the research about the semantic segmentation, BEV generation and perception, vehicle behavior decision, interpretability-aware learning, and place recognition.

Chapter 3 introduces the methodology of our proposed Semantic-MoSeg framework with multi-camera visual inputs and temporal semantic cues, as well as the experimental results and discussion of the proposed network.

Chapter 4 introduces DPMoSeg, a novel architecture that integrates sparse 3-D point clouds and dense visual features through a sparse-to-dense attention mechanism to enhance depth perception. It also presents the experimental results and discussion of the proposed network.

Chapter 5 proposes a unified framework that couples vehicle behavior prediction with natural language-based interpretation, and presents the experimental results and discussions for the components.

Chapter 6 presents our cross-modality place recognition methodology, the experimental results, and the discussion of our proposed network.

Chapter 7 presents the conclusions of the thesis and our future work.

Chapter 2

Literature Review

2.1 Semantic Segmentation

2.1.1 Semantic Image Segmentation

Semantic image segmentation aims to assign a class label to each pixel in a visual image, enabling a dense understanding of visual scenes. Most existing approaches adopt the widely used encoder-decoder architecture, where the encoder extracts high-level features using a deep backbone network, and the decoder progressively recovers spatial resolution to generate dense segmentation maps. Yu *et al.* [108] proposed BiSeNet, which introduces a spatial path to preserve spatial resolution and a context path to capture high-level semantic information. Zhang *et al.* [115] designed an asymmetric encoder-decoder architecture based on convolutional neural networks (CNNs) and incorporated dilated convolutions to address urban scene segmentation efficiently. Choi *et al.* [20] proposed a segmentation framework that utilizes height attention to enhance performance by selectively

emphasizing features based on the vertical position of pixels, which is particularly beneficial for structured road scenes. Yu *et al.* [107] proposed a convolutional neural network for semantic segmentation, where feature aggregation was directly supervised to enhance discrimination between intra-class and inter-class contextual information. Badrinarayanan *et al.* [3] introduced SegNet, an encoder-decoder architecture specifically designed for semantic segmentation, where the decoder utilizes pooling indices from the encoder for efficient upsampling. Lo *et al.* [63] proposed an efficient segmentation framework by integrating asymmetric convolutional structures and dilated convolutions to enlarge the receptive field. Gao *et al.* [27] employed two parallel convolutional layers with different dilation rates to effectively expand the field of view, thereby enhancing segmentation accuracy. Xiao *et al.* [102] proposed BASeg, a CNN-based architecture that incorporates three parallel streamssemantic, boundary, and aggregationto extract boundary-aware features and improve segmentation performance.

2.1.2 Moving-obstacle Segmentation

In contrast to semantic segmentation, moving-obstacle segmentation focuses on identifying the motion state of objects, i.e., distinguishing between static and moving obstacles, rather than assigning detailed semantic class labels. It can thus be regarded as a higher-level binary segmentation task that emphasizes dynamic scene understanding. Depending on the sensing modality, existing approaches can generally be categorized into two groups: vision-based methods and LiDAR-based methods.

LiDAR-based Methods: Chen *et al.* [15] converted 3D point cloud data into

range images and employed SalaNext for moving-obstacle segmentation. Mersch *et al.* [70] proposed a 4D neural network based on the MinkowskiEngine for point-wise segmentation of moving obstacles, incorporating a Bayesian filter for post-processing the predictions. Kim *et al.* [43] also utilized range image projections of LiDAR point clouds and developed a CNN-based network that integrates both semantic and motion features. Chen *et al.* [16] presented an offline approach that combines LiDAR odometry with a clustering algorithm for dynamic object removal and automatic labeling of moving obstacles.

Visual Camera-based Methods: Siam *et al.* [88] proposed a moving-obstacle segmentation solution based on ShuffleNet. However, their approach was limited by the use of a dataset containing only a few hundred images. Vertens *et al.* [98] designed a deep CNN architecture that employed FlowNet 2.0 to distinguish moving obstacles by explicitly capturing motion cues. Rashed *et al.* [81] combined optical flow and LiDAR flow, derived from both camera and LiDAR sensors, to capture detailed motion information, and subsequently proposed a CNN-based framework for object detection. Notably, their method focused exclusively on moving vehicles, excluding moving pedestrians. Liu *et al.* [59] generated depth images from LiDAR point clouds and utilized a CNN-based network for segmentation. Although effective, this approach incurs higher computational costs due to its reliance on LiDAR-derived depth images compared with vision-only methods. Kumar *et al.* [48] introduced a multi-task network employing a shared encoder for fisheye images to simultaneously perform semantic segmentation and motion detection.

Another line of visual-based approaches leverages background modeling algorithms to segment moving objects [91]. However, these methods typically assume

a static camera setup, making them unsuitable for autonomous driving scenarios.

2.2 BEV Generation and Perception

Current methods for BEV map generation can generally be categorized into two types: geometry-based and geometry-free approaches. Among the geometry-based methods, a classical technique is Inverse Perspective Mapping (IPM) [57], which projects perspective-view images into the BEV space. However, the effectiveness of IPM is constrained by its reliance on the ground-plane assumption. In contrast, geometry-free methods aim to learn the mapping from perspective-view to BEV without relying on explicit camera calibration. For example, Jiang *et al.* [37] proposed a Transformer-based architecture with attention modules to implicitly learn this transformation. Pan *et al.* [73] utilized multilayer perceptrons along with an aggregation module to convert visual feature maps into the BEV space. Roddick *et al.* [82] employed a stack of dense Transformer layers to convert perspective-view features into BEV representations. Philion *et al.* [76] proposed a method that learns a depth feature map and projects the volumetric grids into BEV using camera intrinsic parameters. Lu *et al.* [65] introduced an end-to-end framework for generating BEV occupancy maps using monocular camera inputs. Similarly, Can *et al.* [9] utilized a Transformer-based deep neural network to learn structured representations in BEV space. Dwivedi *et al.* [22] proposed a deep network to produce semantic BEV maps by lifting 2D semantic features. Zhou *et al.* [120] designed an attention-based model to implicitly learn the transformation from multiple camera views to a unified map-view representation.

Due to the practical benefits of BEV maps, image-based scene understanding

in the BEV domain has recently attracted considerable attention. However, most existing research primarily focuses on semantic segmentation tasks, which aim at class-level categorization [34, 61, 55]. These approaches emphasize identifying object categories rather than determining their motion status, which is essential for dynamic scene understanding. Furthermore, many of these methods rely on both past and future frames as input, making them less applicable to real-time autonomous driving scenarios where only current-frame information is available.

2.3 Vehicle Behavior Decision

Vehicle behavior decision-making is a critical component of autonomous driving systems. Broadly, existing approaches can be categorized into two paradigms: pipeline-based and end-to-end frameworks [93, 19, 14]. Pipeline-based systems decompose the overall task into sequential sub-modules, such as semantic segmentation [58, 121] and behavior decision-making, each handled by specialized algorithms. These systems typically depend on hand-crafted intermediate representations and structured reasoning. However, the sequential nature of pipeline architectures makes them vulnerable to error propagation, where inaccuracies in early modules can cascade and amplify through subsequent components, ultimately compromising the reliability of the final decision. In contrast, end-to-end approaches leverage visual input to directly predict control commands for the ego vehicle [72]. While this paradigm simplifies the architecture and reduces potential error accumulation, it often suffers from a lack of transparency and interoperability, which are key factors in the deployment of trustworthy and accountable autonomous systems.

2.4 Interpretability-aware Study

While autonomous vehicles today are capable of accurate environmental perception without human intervention, their decision-making processes often remain opaque and difficult for humans to interpret. This has led to growing interest in interpretability for driving systems [5, 49]. The explainability of deep learning-based models is increasingly regarded as a critical factor for the widespread deployment and societal acceptance of autonomous vehicles. Numerous studies have aimed to enhance the interpretability of such models. For instance, Kim *et al.* [44] proposed using attention maps to highlight salient regions in visual inputs, with masked areas serving as visual explanations. However, attention maps primarily emphasize regions of high visual saliency and often fail to provide a holistic understanding of the environment in relation to vehicle behavior. Zeng *et al.* [112] introduced cost-volume computation as an interpretable intermediate representation. Xu *et al.* [104] proposed a two-branch architecture leveraging a pre-trained object detection backbone to improve model interpretability. However, this approach depends on additional detection annotations, significantly increasing annotation costs. Likewise, Mori *et al.* [71] incorporated attention and localization modules to generate textual captions for vehicle behavior explanations. Despite these efforts, many existing approaches, particularly those centered on attention mechanisms, offer limited transparency and often fail to convey intuitive, human-understandable reasoning behind driving decisions. Therefore, we would like to predict the natural-language explanations directly for the ego-vehicle behavior.

2.5 Place Recognition

2.5.1 Single-modality Place Recognition

Single-modality place recognition can be broadly categorized into vision-based and LiDAR-based localization approaches. In the area of vision-based place recognition, Arandjelovic *et al.* [2] proposed NetVLAD, a CNN-based architecture that integrates a generalized VLAD layer to perform robust visual place recognition. Yu *et al.* [110] introduced SPE-VLAD, an end-to-end framework that incorporates a spatial pyramid structure into the VLAD layer and utilizes an improved triplet loss to enhance visual localization performance. Berton *et al.* [7] presented CosPlace, which formulates place recognition as a classification task and leverages contrastive learning to address large-scale recognition challenges effectively. Ali-Bey *et al.* [1] proposed an aggregation-based method, MixVPR, that fuses backbone feature maps with global contextual relationships to produce robust global descriptors.

For the LiDAR-based solutions, most methods proposed to fully extract and learn the statistical features of geometric distributions. For example, Uy *et al.* [96] introduced PointNetVLAD, a network that combines PointNet [77] with NetVLAD [2] to directly extract global descriptors from raw point clouds for 3D place retrieval. Ma *et. al* [68] proposed a real-time solution, named OverlapTransformer, using range images and a yaw-invariant framework with attention structures to perform place recognition. Hui *et al.* [36] developed a pyramid point cloud Transformer structure integrated with a VLAD layer to learn spatial relationships and produce robust global descriptors purely from point cloud data. Li *et al.* [52] presented a lightweight network that fuses semantic and geometric information to

enable effective LiDAR-based place recognition.

2.5.2 Cross-modality Place Recognition

In addition to single-modality approaches, extensive efforts have been made to explore multi-modality place recognition, which leverages the complementary characteristics of different sensor modalities. For example, Bernreiter *et al.* [6] proposed to extract a multi-modality descriptor based on spherical projection and CNN architecture for place recognition. Wang *et al.* [100] introduced PRFusion, which integrates visual and LiDAR data through a global fusion mechanism, manifold metric attention, and a neural diffusion module. However, cross-modality place recognition remains a more challenging yet promising direction due to the inherent modality gap between visual and LiDAR data. To bridge this gap and enhance localization performance, various solutions have been proposed. Zhao *et al.* [118] designed an transformer-based architecture that extracts cross-modal features from captured images and 3D information to produce robust global descriptors. Shubodh *et al.* [87] developed LIP-Loc, a contrastive learning-based framework designed to address the cross-modal localization task. Zheng *et al.* [119] introduced a solution that reconstructs LiDAR via a depth estimation network, enabling recognition of visual inputs within point cloud maps. Li *et al.* [54] introduced VxP, a unified place recognition framework that voxelizes point clouds and projects both voxel representations and image features into a common embedding space.

2.6 Summary

This chapter presented a systematic review related to the core components of this thesis. We began by examining recent advances in semantic segmentation, including image-level segmentation, moving obstacle segmentation. Following this, we discussed BEV generation and perception techniques, with a focus on monocular BEV map construction and BEV-based visual understanding. Moreover, we present related works for the vehicle behavior decision and interpretability-aware study. Furthermore, we reviewed on place recognition methods, encompassing both single-modality and cross-modality approaches. Together, these lines of research form the basis for the methodologies proposed in this thesis.

Chapter 3

Semantics-assisted Moving-obstacle Segmentation in Bird-Eye-View for Autonomous Driving

3.1 Introduction

In recent years, autonomous driving technologies have made significant strides. For the widespread deployment of autonomous vehicles, accurate and reliable detection or segmentation of moving obstacles in dynamic traffic environments is a fundamental requirement. This capability is crucial, as the perception of moving obstacles, such as vehicles in motion or walking pedestrians, provides essential information for a variety of downstream tasks, including motion planning [21, 84], decision-making [32, 30], and visual localization [113, 106].

Moving-obstacle segmentation involves identifying obstacles and distinguishing their motion states (i.e., static or moving). Existing approaches typically per-

form this task using either 2D perspective-view images or 3D point clouds. Some methods [75, 29] operate in an offline manner by processing entire sequences of frames, while others rely on prior information such as manually annotated masks for the initial frame. In this work, we focus exclusively on online methods that do not depend on any manually labeled masks during inference. Compared to segmentation in the front-view perspective, BEV representations offer a more intuitive and task-relevant format, particularly since many downstream modules, such as motion planning and trajectory prediction, operate on BEV maps. To perform moving-obstacle segmentation in BEV, one common approach is to leverage existing 3D LiDAR-based methods [15, 4] to first generate segmentation results in the point cloud space, and subsequently project them into the BEV domain. However, a key limitation of this strategy lies in the inherent sparsity of point clouds, which often results in BEV maps that are incomplete and contain numerous holes or gaps. Such sparse representations significantly hinder the effectiveness of downstream modules. Moreover, LiDAR sensors are expensive and impose substantial computational demands, limiting their practicality for large-scale or cost-sensitive deployments.

To address the aforementioned limitations, this paper proposes a novel approach that generates dense BEV maps using images captured by low-cost visual cameras. Specifically, our network takes as input visual data from six surrounding cameras mounted on the ego-vehicle and directly produces dense BEV moving-obstacle segmentation maps. The reason for us to use six cameras is that the field-of-view (FOV) of the cameras can be complemented by each other, and hence, our method can produce 360° ego-centric BEV maps. Note that in some literature, *moving obstacle* is also termed as *moving object*. Since this paper focus on traffic

environments for autonomous driving research, the term *moving obstacle* would be more appropriate. In the following text, we do not discriminate them. Moreover, there exist some methods using visual images to generate semantic BEV maps [76, 9]. Our method differs fundamentally from existing approaches by incorporating motion state discrimination (e.g., static or moving), which is absent in conventional semantic-only segmentation methods. In addition, some methods [65, 73] can only produce cone-shaped BEV maps due to the use of visual information from one camera view.

The principal contribution of this work is the development of an online method capable of producing dense BEV segmentation results specifically for moving obstacles. Our method does not rely on future captured images other than the current and previous images. To the best of our knowledge, this is the first work that produces moving-obstacle segmentation results in BEV using sequential images from multiple vehicle-mounted cameras. In addition, our method incorporates movable features to bring the motion and movable circumstances of the obstacles by proposing a movable-obstacle segmentation auxiliary task and showing that the movable information could further benefit the moving-obstacles segmentation performance.

3.2 The Proposed Framework

3.2.1 The Overall Architecture

Fig. 3.1 shows the structure of our network Semantic-MoSeg. It can be seen that it mainly consists of four modules: feature extractor, geometry-guided BEV generation (G2BG) module, moving-obstacle segmentation module, and auxiliary task:

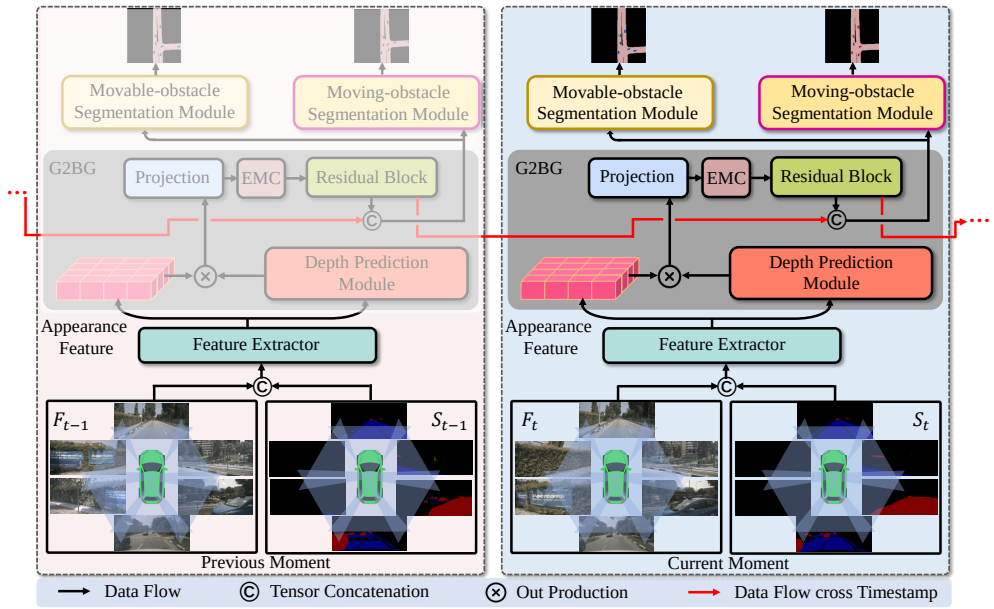

Figure 3.1: The overall structure of our proposed network. F and S represent the vision images and corresponding semantic prior information, respectively. EMC represents ego-motion compensation. The ■ color in the moving-obstacle segmentation map represents moving obstacles. The ■ color in the movable-obstacle segmentation map represents the movable obstacles. The ■ and ■ colors represent drivable areas and the ego-vehicle, respectively.

movable-obstacle segmentation module. To be specific, our network takes as input two sets of six images from the surrounding cameras mounted on the ego-vehicle that are captured at different moments (the time interval is fixed) and directly generates moving-obstacle segmentation maps and movable-obstacle segmentation maps of the current moment.

As shown in Fig. 3.1, semantic segmentation is first performed on six input images from the surrounding cameras using [17] to generate semantic segmentation maps $S_t = \{S_t^1, ..., S_t^n\}$ at the current moment t, then visualize the semantic segmentation maps in 3-channel color images. This segmentation maps information is seen as our semantic prior information. The input images and the corresponding semantic segmentation maps are concatenated to form 6-channel images. Sec-

ondly, a feature extractor is adopted to extract the visual features from the 6-channel images from the current and previous moments. All the visual images share the same feature extractor module. Thirdly, the G2BG module transforms the perspective view features to BEV and produces BEV feature maps for the consecutive two moments. Finally, the moving-obstacle segmentation module and the movable-obstacle segmentation module are employed to produce the segmentation results. The movable and moving-obstacle segmentation maps are both with the size of N_c*H*W, N_c is the number of classes, the H and W are the sizes of the BEV feature map.

3.2.2 Feature Extractor

As mentioned above, the input data to the feature extractor are the concatenated 6-channel images from F_t and S_t at current and previous moments. We choose EfficientNet-B4 [94] as our backbone for the feature extractor due to its lightweight architecture. Specifically, we adopt EfficientNet-B4 with the random initialization. We modify the input channel of the first layer so that the feature extractor can take as input the concatenated images. The concatenated images are finally downsampled with a factor of 8.

3.2.3 Geometry-Guided BEV Generation (G2BG)

Here, the G2BG module is proposed to generate dense moving-obstacle features under BEV. We follow [76] to transform perspective-view features to BEV feature maps by predicted depth distributions. Firstly, depth prediction is conducted based on the features produced by the feature extractor, which is shown in Fig. 3.2. We

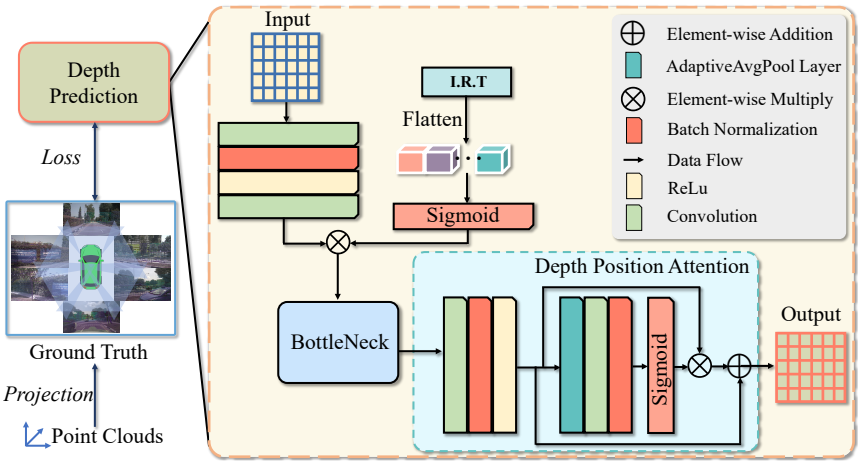

Figure 3.2: The structure of our proposed depth prediction module. The ground truth for depth prediction is the sparse depth from LiDAR point clouds, which are overlaid on the RGB images for visualization.

simply flatten the camera intrinsic matrix (3*3) and extrinsic matrix (3*3 rotation and 3 * 1 translation) into a vector with the length of 21. To further enhance the depth prediction, we design a depth position attention (DPA) module. In the DPA module, a convolutional kernel of size 3*3 is employed. Specifically, the sparse depth data from the LiDAR point clouds is utilized as the supervision signal to train the depth prediction network. With the predicted depth, we project the perspective-view features to BEV features.

Secondly, we conduct ego-motion compensation (EMC) on the BEV features. Since the differences between consecutive images result from both ego-vehicle motion and dynamic object movement. We adopt the vehicle pose between the two moments to conduct EMC. The idea is to transform all the previous features into the coordinate system at the current moment.

Finally, the coarse BEV feature maps are further refined by exploiting the temporal information between the two moments. A residual module is proposed to

find the difference between the coarse BEV feature maps at the two moments:

$$\hat{bev}_t = bev_t + bev_t * bev_{t*1}, \tag{3.1}$$

where bev_t is the output representation from the residual block, $\hat{bev}_t$ denotes the difference. The bev_t is doubled to avoid $\hat{bev}_t$ to be zero. Then, we concatenate the residuals calculated from the previous moment as the output of the G2BG. The idea behind the residual module is that the static parts would become smaller after subtraction, which in turn amplifies the differences between the moving part and static part.

3.2.4 Moving-obstacle Segmentation

We design the moving-obstacle segmentation module using a modified DenseNet [35], atrous spatial pyramid pooling, and skip connections. At the end of the module, a moving-obstacle segmentation head is added, which is a 2-D convolutional layer, to provide dense pixel-wise prediction. The BEV map range is $100m$*$100m$ with a unit length of $0.5m$. So, the resolution of the BEV map is $200 * 200$.

3.2.5 Auxiliary task: Movable-obstacle Segmentation

The movable-obstacle segmentation module is designed to endow the network with the ability to inherently and implicitly learn what kind of obstacles are possible to move. The class-agnostic movable-obstacle segmentation task is introduced here to enhance our moving-obstacle segmentation performance since moving objects should be movable objects (e.g., pedestrians, vehicles). In this task, both pedes-

trians and parked cars are considered the same class. Endowing the network with the ability to segment movable obstacles should increase the moving-obstacle segmentation performance. Note that movable is still a concept of semantics. It does not involve motion states. In this task, we only focus on binary segmentation, that is, two classes of movable (e.g., vehicles and pedestrians) and unlabelled background. We adopt the same network as the moving-obstacle segmentation module and share the same parameters between them. The only difference is that we replace a movable-obstacle segmentation head with the moving-obstacle segmentation head to achieve the binary movable segmentation.

3.2.6 Loss Functions

Binary cross-entropy (BCE) loss is employed for both moving and movable segmentation, as well as the depth distribution prediction. The three losses are denoted as $\mathcal{L}_{moving}$, $\mathcal{L}_{movable}$, and $\mathcal{L}_{depth}$. We follow [53] to compute the depth loss. The total loss is:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{moving} + \beta \cdot \mathcal{L}_{movable} + \gamma \cdot \mathcal{L}_{depth}, \tag{3.2}$$

Different weights are assigned to different losses. The parameters α, β, and γ are weights that can be learned to balance the three losses.

3.3 Training Strategies and Evaluation Metrics

3.3.1 Datasets

To evaluate performance of the proposed method, we conduct experiments on the public large-scale nuScenes [8] and Lyft [41] datasets. The nuScenes dataset in-

cludes $1,000$ scenes with visual images from six vehicle-mounted cameras and point clouds from a 3-D LiDAR sensor. Among the scenes, 850 scenes are given ground-truth annotations. The moving-obstacle ground-truth labels are generated by filtering the attributes and projecting the provided ground-truth bounding boxes into BEV to obtain 2-D polygons. The moving obstacles in nuScenes have the attributes of `pedestrian.moving` and `vehicle.moving`. All the images in the same scene are used either for training, validation, or testing. We divide the 850 scenes into three subsets: 550 for training, 150 for validation, and 150 for testing. The Lyft Perception dataset includes multi-view images captured by rooftop-mounted cameras on the ego-vehicle. From the 180 scenes with annotated ground truth, 36 are randomly selected for evaluation. Importantly, as the shuffling is performed at the scene level, the temporal continuity within each image sequence is preserved.

3.3.2 Implementation Details

We implement our network with PyTorch and the PyTorch-Lightning [74] library. The resolution of the input images is resized to 224*480 for the experiments. The proposed network processes a sequence of 6 camera images and outputs BEV maps with 200*200 resolution at 50cm unit length in both the x and y directions. If not stated differently in the experiments, the time interval for our experiment is the time between two adjacent key frames.

Adam optimizer with decoupled weight decay [64] are adopted for training. The initial learning rate is 10^{-3}. In particular, we train all the experiments under mixed precision. The training data are randomly shuffled before each epoch. How-

Table 3.1: The ablation study results of different feature extractors.

Extractor	Movable		Moving	
	IoU %	Precision %	IoU %	Precision %
Efficient-B0	37.62	49.97	32.54	43.74
Efficient-B1	38.60	52.08	33.33	45.12
Efficient-B2	38.45	51.25	33.72	46.91
Efficient-B3	39.49	53.26	34.76	49.06
Efficient-B4	39.57	53.08	35.08	48.61
ResNet-18	36.09	49.22	29.75	43.51

ever, since the shuffling is performed on entire sequences of consecutive frames, the temporal order within each image sequence is preserved. A MultiStepLR scheduler is employed to decay the learning rate throughout training.

3.3.3 Evaluation Metrics

We adopt evaluation metrics for quantitative evaluation: $\mathrm{precision}$ and intersection-over-union (IoU) [23]:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.3}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. Note that our metrics are calculated with respect to the ground truth at the current moment.

Table 3.2: The ablation results of different module components in G2BG.

Variants No.	Timestamp	Geometry-aware	Pose	Residual Block	Movable		Moving	
					IoU %	Precision %	IoU %	Precision %
Variant-1	2	✗	✗	✗	38.03	51.37	31.09	43.26
Variant-2	2	✓	✗	✗	38.75	51.83	32.36	44.30
Variant-3	2	✗	✓	✗	39.41	53.37	33.38	45.80
Variant-4	2	✗	✗	✓	37.55	49.31	33.03	43.79
Variant-5	2	✓	✓	✗	39.78	53.99	34.09	47.37
Variant-6	2	✓	✗	✓	39.33	52.95	34.23	47.72
Variant-7	2	✗	✓	✓	39.80	53.93	34.74	47.32
Variant-8	2	✓	✓	✓	39.57	53.08	35.08	48.61

3.4 Ablation Study

We create several variants of our network to validate the effectiveness of our design. We train all the variants up to 30 epochs and report the best results on the nuScenes dataset for moving-obstacle segmentation.

3.4.1 Ablation on Feature Extractor

We adopt the ResNet and EfficientNet. The EfficientNet has variants from B0-B7, where B5-B7 contains more parameters, which increases network performance but also increases the computational cost. To trade off performance and computational cost, we conduct the ablations using EfficientNet B0-B4. From Tab. 3.1, the variant with EfficientNet-B4 gets the best performance. So, unless otherwise specified, EfficientNet-B4 is adopted as the feature extractor for our experiments.

3.4.2 Ablation on Ego-motion Compensation

This ablation study is to demonstrate that introducing ego-motion compensation could increase the moving-obstacle segmentation performance. Tab. 3.2 shows the results. It can be observed that with the ego-motion compensation, the IoU of

the moving-obstacle segmentation has been improved with 0.85%, and the precision of segmenting moving obstacles has been improved with 0.89%.

3.4.3 Ablation on Geometry Awareness

This ablation study is designed to demonstrate the benefits of introducing the camera's intrinsic/extrinsic parameters. The results are shown in Tab. 3.2. It can be illustrated that the camera parameters are beneficial to the moving-obstacle segmentation performance. From Variant-6, it can be found that with the intrinsic and extrinsic parameters, the IoU of our method has been improved by 0.34%. It can also improve precision by 1.29% in the movable-obstacle segmentation.

3.4.4 Ablation on Residual Block

This ablation study aims to demonstrate the effectiveness of the residual block. Tab. 3.2 shows the experimental results. It can be seen that the residual layer is positive on moving-obstacle segmentation based on the residuals of the time-series feature maps. From Variant-5, we can learn that with the residual operation, the IoU of our network has been improved by 0.99%, and the precision has been improved by 1.24%. It can be observed that with the amplified difference between the static and moving obstacles, it is easier for the network to find the moving obstacles.

3.4.5 Ablation on Depth Supervision

This ablation study aims to demonstrate the depth supervision performance. The results are shown in Tab. 3.3. As aforementioned, the spare depth ground truth

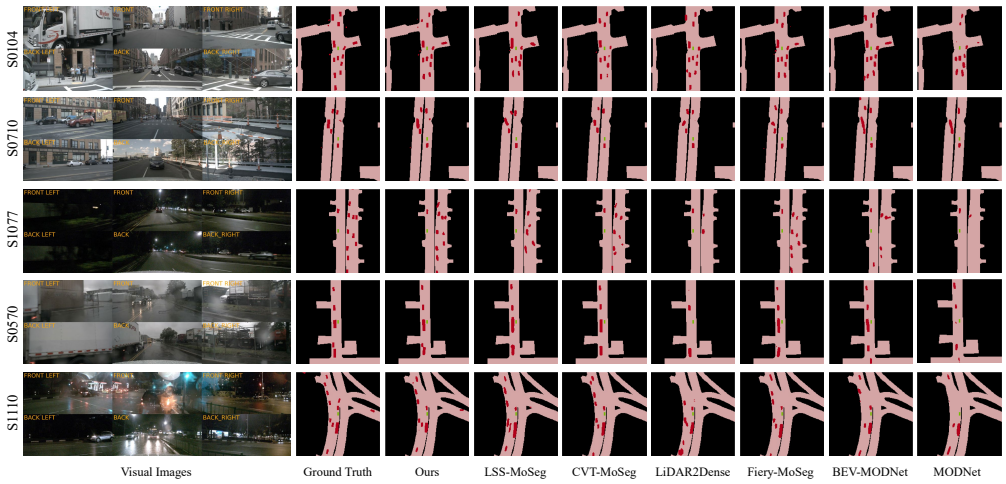

Figure 3.3: Sample qualitative demonstrations for moving-obstacle segmentation under different weather and lighting conditions. The ■, ■, and ■ pixels represent moving obstacles, ego-vehicle, and drivable areas.

Table 3.3: The ablation results of the depth supervision and auxiliary task.

Supervision	Sub-task	Movable		Moving	
		IoU %	Precision %	IoU %	Precision %
✗	✗	-	-	32.16	44.10
✗	✓	38.84	51.38	33.94	47.48
✓	✓	39.57	53.08	35.08	48.61
✓	✗	-	-	33.78	46.48

is obtained from the LiDAR point cloud. It can be seen that with the depth supervision, the IoU of our network has been improved by 1.14%, and the precision of segmenting moving obstacles has been improved by 1.13%. This indicates that depth supervision could further benefit the moving-obstacle segmentation performance.

Table 3.4: The ablation results on different ranges.

Timestamp	Range	Movable		Moving	
		IoU %	Precision %	IoU %	Precision %
2	(-25m, 25m, 0.25m)	53.05	68.51	45.05	59.70
2	(-40m, 40m, 0.4m)	40.68	52.84	35.91	47.04
2	(-50m, 50m, 0.5m)	39.57	53.08	35.08	48.61

3.4.6 Ablation on Auxiliary Task

This ablation study is to demonstrate whether adding auxiliary-task, movable-obstacle segmentation, is helpful for boosting the performance of our moving-obstacle segmentation. The results are displayed in Tab. 3.3. It can be illustrated that the auxiliary task to learn the movable features in the surroundings is helpful in improving moving-obstacle segmentation performance. This is reasonable, as moving obstacles must inherently be movable.

3.4.7 Ablation on Detection Ranges

This ablation study is to demonstrate the robustness and performance of our network on the different detection ranges. Here, we have three different settings: $100m$*$100m$ resolution at $50cm$ unit length, $80m$*$80m$ resolution at $40cm$ unit length, and $100m$*$50m$ resolution at $25cm$ unit length. Tab. 3.4 shows the results. It can be seen that our network performs robustly at different detection ranges.

3.4.8 Ablation on Semantic Prior Information

This ablation study is to demonstrate the benefit brought by the semantic prior information. The results are displayed in Tab. 3.6. The variant Visual-only refers

to using the captured visual multi-view images as input. The variant Prior-only refers to using the semantic-prior information as input. The variant Element-wise summation refers to using the addition between visual multi-view images as input. It can be observed that the moving-obstacle segmentation performance is gradually enhanced as the input information is gradually enriched. Feature concatenation achieves higher performance gains compared to element-wise summation. The reason may be due to that the concatenation is helpful for our method to adaptively learn feature maps. The incorporation of semantic information could make the network segment moving obstacles more easily.

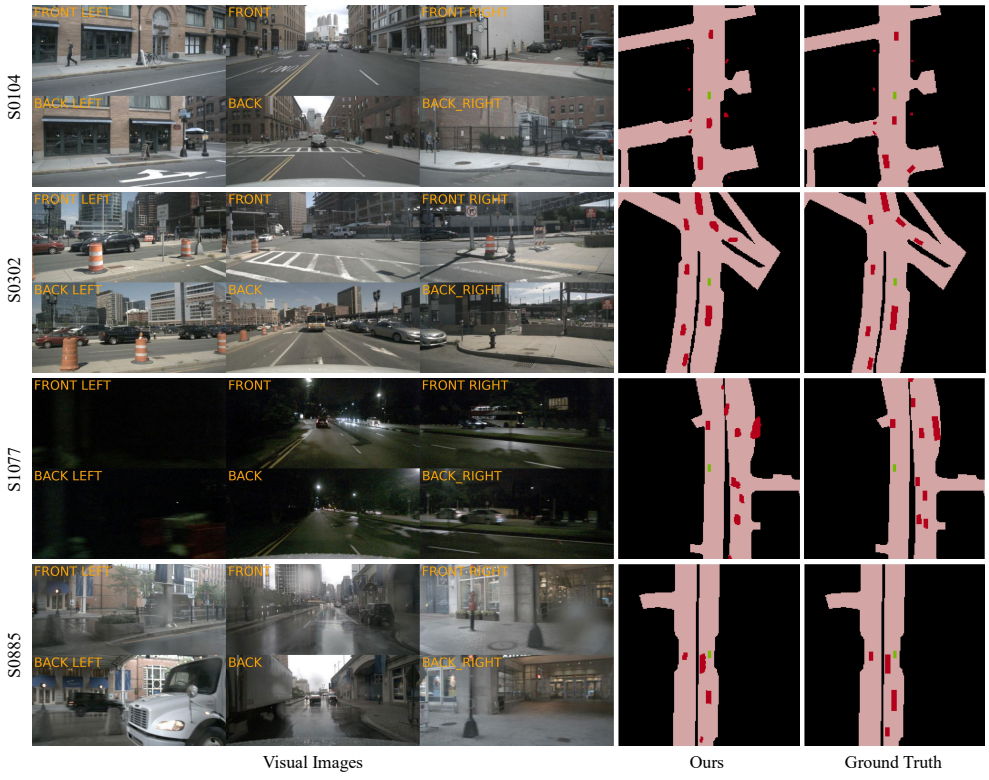

Figure 3.4: Sample qualitative demonstrations of our network for moving-obstacle segmentation on the nuScenes dataset. The ■ pixels represent the moving obstacles under BEV. The ■, and ■ pixels represent ego-vehicle, and drivable areas, respectively.

Table 3.5: Comparative results of different methods on the nuScenes dataset.

Timestamp	Baseline Type	Moving	
		IoU (%)	Precision (%)
2	BEV-MODNet	22.17	30.07
2	SMSnet	12.18	14.97
2	MODNet	13.68	17.14
2	LSS-MoSeg	24.60	37.34
2	CVT-MoSeg	27.41	37.81
2	Fiery-MoSeg	29.89	40.08
2	LiDAR2Dense	25.13	47.42
2	Semantic-MoSeg (ours)	35.08	48.61

Table 3.6: The ablation results on semantic prior information.

.

Timestamp	Input Type	Movable		Moving	
		IoU %	Precision %	IoU %	Precision %
2	Visual-only	39.14	53.52	32.44	44.60
2	Prior-only	37.98	51.13	31.98	44.31
2	Element-wise addition	39.13	53.16	34.30	47.60
2	Semantic-MoSeg (ours)	39.57	53.08	35.08	48.61

In summary, the proposed network consistently achieves superior performance compared to other variants, validating the effectiveness of its overall design. The observed improvements in IoU and precision across various configurations further indicate that each component contributes positively, highlighting the advantage of incorporating all proposed modules.

3.5 Comparative Experiments

3.5.1 The Overall Results

Here, we compare our method with existing methods (i.e., BEV-MODNet [80], SMSnet [98], and MODNet [89]) and create several baselines (i.e., LSS-MoSeg, CVT-MoSeg, and Fiery-MoSeg) for comparison. For all the compared methods, the output layers are modified to generate moving-obstacle segmentation maps in BEV. The detailed descriptions for the compared methods are listed as follows:

- **BEV-MODNet**: This method [80] is proposed for BEV moving object detection based on the front-view images. We first generate the optical flow for the multi-view visual images. Then we re-implement the model based on the architecture and adopt the multi-view visual images as inputs.

- **SMSnet**: This method [98] is designed to detect moving objects in a perspective view based on two consecutive time-stamp inputs. We adopt the multi-view visual images as inputs and project the detection results via the IPM algorithm.

- **MODNet**: The MODNet method [89] is proposed for perspective-view detection based on the visual image and its corresponding optical flow. We generate the optical flow for the visual images and project the segmentation results to BEV via the IPM algorithm.

- **LSS-MoSeg**: This baseline is based on the LSS method [76]. The input of LSS is the monocular images at a single moment. Since moving-obstacle segmentation requires sequential data, we change the time interval length to

2 moments. We name this baseline as LSS-MoSeg.

- **CVT-MoSeg**: This baseline is based on CVT [76] approach. We change the time interval of the CVT method to 2. We name this baseline as CVT-MoSeg.

- **Fiery-MoSeg**: This baseline is based on Fiery [34] method. The original Fiery is for semantic segmentation with sequential data. Since the future information could not be used for online applications, such as autonomous driving. So, we omit the sub-module using future features in [34]. We name this baseline as Fiery-MoSeg.

- **LiDAR2Dense**: This baseline is based on PointPillars [50]. It adopts the same moving-obstacle segmentation head as ours to generate dense predictions. Here, we only use the 3-D coordinates of the point clouds as input. The other information, such as intensity, is discarded. Since our output is dense BEV maps for moving obstacles, we name this baseline as LiDAR2Dense.

As shown in Tab. 3.5, it can be seen that our Semantic-MoSeg achieves the best performance compared with all the methods, which demonstrates our superiority. Fig. 3.3 qualitatively demonstrates sample comparative results for moving-obstacle segmentation. We can see that the segmentation performance of our Semantic-MoSeg generally outperforms the other methods. This could be attributed to the effective representation of multi-view visual features and the integration of auxiliary movable-obstacle features. The LSS-MoSeg, CVT-MoSeg, and Fiery-MoSeg baselines lack movable-obstacle features, which are helpful for

moving-obstacle segmentation. Furthermore, the SMSnet and MODNet methods require post-projection and optical flow calculation, introducing intermediate errors that might affect segmentation performance. Specifically, the first row shows a cloudy scenario. It can be seen that there are moving trucks, cars, and pedestrians on the scene. Our network is the only one that can segment both moving vehicles and pedestrians. The second row shows a sunny scenario. Our network is the only one that segments moving cars and pedestrians under BEV space. For the remaining comparative methods, they both miss the pedestrians crossing the road. The third row shows a nighttime scenario. Under such light conditions, streetlights and vehicle headlights share similar visual appearances. It can be seen that our results are more complete than the other baselines. The other baselines all miss some segmentation to some degree, especially when objects are far from the ego vehicle. The fourth row shows a rainy scenario. Under such weather conditions, although the lens of the cameras are blurred, our network detects all the moving vehicles. The last row shows a scenario with on-coming and nearby headlights under a rainy night condition. Although the light is weak and there is interference from water ripples, our network can still accurately segment the moving obstacles.

3.5.2 Qualitative Demonstrations

We also present some qualitative demonstrations of our method on both moving-obstacles and movable-obstacles segmentation, which are displayed in Fig. 3.4 and Fig. 3.5. It can be shown that our Semantic-MoSeg generalizes well to unseen environments under various weather and lighting conditions, including sunny, rainy, cloudy, and nighttime.

Table 3.7: Robustness evaluation of our network with on the Lyft dataset.

Timestamp	Input Type	Movable		Moving	
		IoU %	Precision %	IoU %	Precision %
2	Semantic-MoSeg (ours)	45.11	60.28	43.19	55.43

3.6 Robustness Evaluation

3.6.1 Robustness on Different Dataset

To demonstrate the robustness of our method, we evaluate our network on another large-scale public dataset, Lyft. We get the moving-obstacle ground truth by filtering and merging the annotated attributes in the same way as nuScenes dataset. The range for segmentation is also $100m$*$100m$. Tab. 3.7 displays the results on Lyft dataset. It can see that compared to the results on nuScenes dataset, our network exhibits robust performance on another dataset. Meanwhile, our solution can perform robustness on our auxiliary task movable-obstacle segmentation. Besides, the inference speed of our model on multi-view cameras is approximately 102 ms.

3.6.2 Robustness on Different Views

To validate the robustness on different views, we randomly remove an image from a viewpoint among the six viewpoints during the test, and then apply the saved model weights to the remaining images. Fig. 3.6 shows the results. It can be observed that the performance of our network varies when dropping different viewpoints, but generally is robust when Camera-front-left, Camera-front-right, Camera-back-right, or Camera-back-right is dropped. When Camera-front or Camera-back view is dropped, the performance of our network is degraded. We conjecture the reason

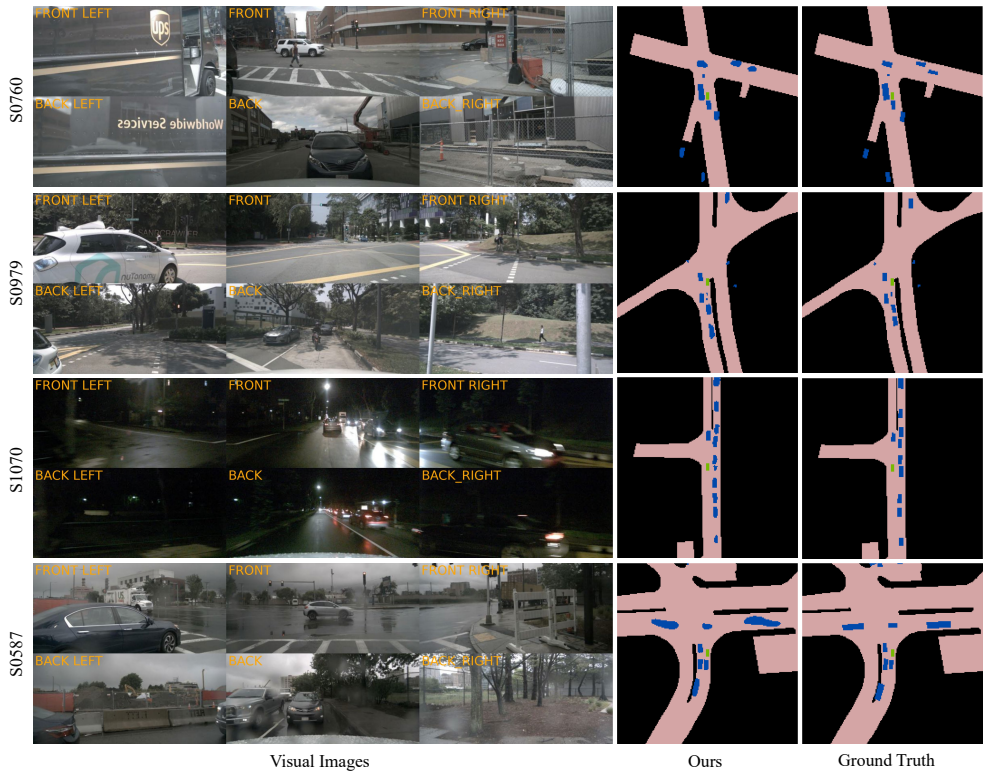

Figure 3.5: Sample qualitative demonstrations of the proposed network for auxiliary task under nuScenes dataset.auxiliary task. The ■ denotes the movable obstacles under BEV.

is that the front view encodes more information, and the back view is just the opposite of the front view. It can be seen that IoU and Precision metrics have similar performance varying trends. It can be indicated that with respect to the mIoU, it is the highest in the Camera-front-left view. And it is the highest in the perspective of the Camera-back-right in terms of the mPre.

3.7 Conclusion

In this chapter, we proposed a novel end-to-end framework for online moving obstacle segmentation in the birds-eye view, utilizing multiple images captured at dif-

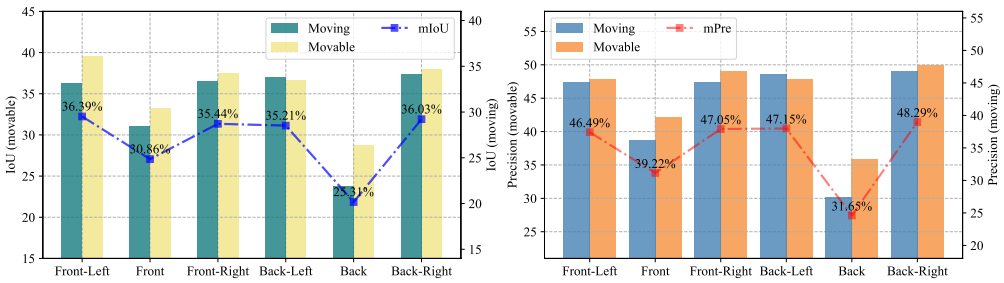

Figure 3.6: Robustness evaluation of our network with different camera views on the nuScenes dataset. *Movable* and *Moving* represent the movable-obstacle and moving-obstacle segmentation tasks.

ferent time steps. The method projects perspective-view feature maps into the BEV space by explicitly incorporating camera intrinsic and extrinsic parameters, along with prior semantic information. To further enhance segmentation performance, we introduce an auxiliary task, movable obstacle segmentation, that provides additional supervisory signals. Comprehensive experiments conducted on the public datasets validate the effectiveness and robustness of our approach, demonstrating superior performance in moving obstacle segmentation tasks.

Chapter 4

Depth-powered Moving-obstacle Segmentation Under Bird-eye-view for Autonomous Driving

The reliance on pure vision information inevitably introduces limitations, such as depth ambiguity. This chapter extends BEV dynamic perception toward a multi-modal framework by integrating LiDAR information. The proposed DP-MoSeg leverages complementary geometric cues from sparse 3-D point clouds and fuses them with dense visual representations through an attention mechanism, thereby enhancing the performance and robustness of BEV moving-obstacle segmentation.

4.1 Introduction

Nowadays, it is critical to develop a safe and trustworthy autonomous driving system. In this process, segmenting the moving obstacles in the surroundings is vital

and fundamental. This is because moving obstacles are usually accompanied by higher collision risk than static ones. It has wide real-life applications since it will elegantly provide direct, useful information for the downstream tasks. To be specific, moving obstacle segmentation can be seen as a special binary segmentation task that tries to get the obstacle states, moving or static, based on consecutive time-step input information.

Recently, the interest in moving obstacle segmentation has increased. Nowadays, some progress on moving obstacle segmentation has been mainly made in LiDAR-based [15, 92] and camera front-view based [76, 73] methods. The LiDAR-based solutions can generate sparse segmentation maps according to the captured sparse raw point clouds. The camera-based approaches usually give dense 2-D perspective-view moving-obstacle segmentation results based on semantic information and texture information.

However, on the one hand, the camera-based solutions usually lack depth information due to the characteristics of the vision sensor, even though the output representations are dense. Depth information plays a significant role in autonomous vehicles since obstacles need to be reasoned in the correct locations. On the other hand, LiDAR sensors could provide the most accurate measurements in depth and reflect the 3-D feature and location information of surrounding obstacles. Therefore, to generate dense moving-obstacles segmentation, a simple solution is to apply the information both from the visual and LiDAR information. However, directly using multi-modal information will add the burden of higher economic cost and more computation, and need to carefully design the modules to fuse the point cloud and visual information. In this paper, we propose to combine the sparse depth information obtained from LIDAR and the captured multi-view visual in-

formation to get a dense moving-obstacles segmentation map.

Moreover, obstacle occlusion is a common phenomenon under perspective view. BEV representation could alleviate the occlusion issue because it elegantly provides the grid occupancy under 3D space and provides useful, straightforward information Therefore, in this work, our target is to conduct accurate dense moving-obstacle segmentation under BEV space.

To solve the above-mentioned issues, we introduce a novel depth-powered moving-obstacle segmentation network termed DPMoSeg in this paper. To get more precise dense 3-D information, we designed a sparse-dense depth-powered attention module to produce the dense BEV features under fixed time stamps. To get more accurate moving-obstacle segmentation maps, we designed an auxiliary task, drivable-area segmentation, to boost the segmentation performance. We implement multiple baselines to perform comparative studies on the public NuScenes [8] dataset. The experimental results demonstrate our superiority.

4.2 Methodology

4.2.1 Overall architecture

The proposed network architecture, DPMoSeg, is shown in Fig. 4.1. The input is the captured visual images and the corresponding sparse depth information generated. In general, our DPMoSeg consists of a feature extractor, BEV feature representation, moving-obstacle segmentation, and drivable-area segmentation network.

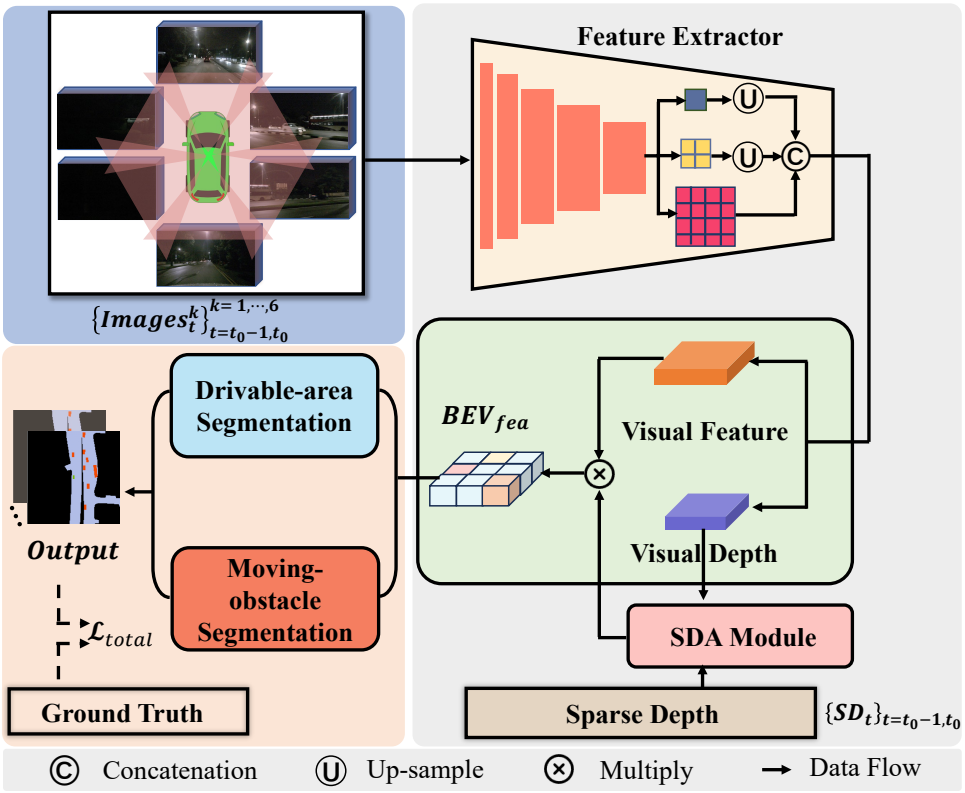

Figure 4.1: The overall architecture of our framework DPMoSeg. The sparse depth is generated from the LiDAR. The output is the moving-obstacle segmentation map under BEV. The ■ pixels represent the moving obstacles. The ■ pixels represent the drivable region. The ■ pixels refer to the ego-vehicle.

4.2.2 Depth-powered BEV Generation Module

The proposed solution is trying to get more accurate dense depth information and then get more accurate BEV feature representation. The input information of the approach is the multi-view visual information and the sparse depth generated from the LiDAR sensor.

- **2D Feature Extractor**: We modified the EfficientDet architecture [95] as our feature extractor. Specifically, we down-sample the visual information by a factor of 32 using the feature extractor and generate multi-scale feature

representation, $\mathcal{F}_8$, $\mathcal{F}_{16}$ and $\mathcal{F}_{32}$. Subsequently, we restore the feature map to one-eighth of its initial size with several up-sample operations. The output feature map of the feature extractor is denoted as $\mathcal{F}_{extract}$.

- **Sparse Depth-powered Attention-based View Transformation**: We follow the ref [76] paradigm to generate BEV feature maps, which focus on predicting a probability depth distribution from the multi-view vision images and generating a dense BEV feature maps. Due to the characteristics of the visual camera, the depth distribution generated from the multi-view vision images D_{depth}^{dense} lacks 3-D geometric information about the scenarios. The predicted depth information affects the performance of the resulting predicted BEV features. It is important to make full use of the 3D information. In this paper, we instead generate the corresponding sparse depth information by projecting the LiDAR point clouds, D_{depth}^{sparse}. Even though the generated depth information contains accurate 3-D cues, the depth information is sparse. This is because the LiDAR data only captures the foreground obstacles with a limited sampling rate. Therefore, we propose a sparse-dense attention module, dubbed SDA, to get a denser depth representation with more 3-D spatial information, $\mathcal{D}_{depth}$, which is computed as:

$$\mathcal{D}_{depth} = SDA\left(\mathcal{D}_{depth}^{dense}, \mathcal{D}_{depth}^{sparse}, \mathcal{D}_{depth}^{dense}\right) \tag{4.1}$$

The architecture is shown in Fig. 4.2. The objective of the SDA module is to utilize depth features generated from visual data as a value, combined with sparse depth information as a key and query, to acquire fused depth information features ultimately.

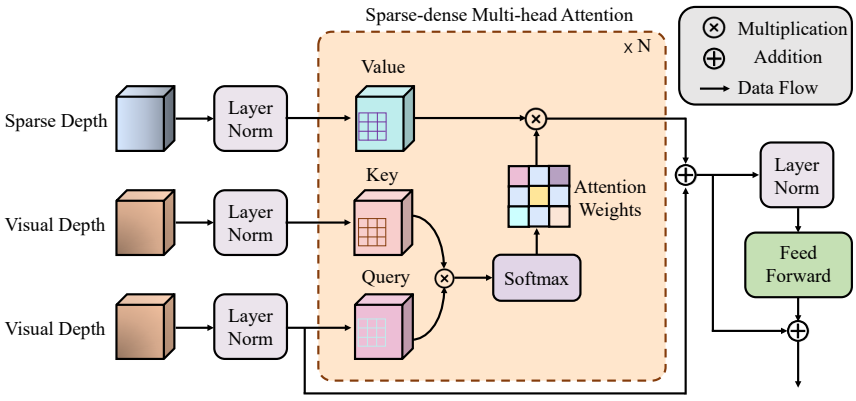

Figure 4.2: The architecture of sparse-dense depth estimation (SDA) module.

Finally, we get the continuous BEV feature representation under the fixed temporal frames, $\mathcal{F}_{BEV}$, based on the generated depth information and corresponding visual information.

- **Moving-obstacle Segmentation Module**: For the generated BEV representation, we adopt a CNN structure to get the moving obstacle scores. To be specific, we modified the ResNet18 to extract the $\mathcal{F}_{BEV}$ and adopted the upsample operation to decode the BEV feature maps. Finally, a convolutional layer will output the moving-obstacle segmentation predictions.

- **Sub-task: Drive-able Area Segmentation**: To enhance the capability of segmenting the moving obstacle, we also propose a sub-mission for segmenting the driveable area. The purpose of adding this side task is that a vehicle in motion is usually driven in a drive-able area rather than floating in the air. We add a drive-able area segmentation head to get the area segmentation score based on $\mathcal{F}_{BEV}$.

4.2.3 Loss function

We use the cross-entropy function (CE) to compute the loss. In our network, there exist two computed losses. The first part is calculated between the moving-obstacle segmentation ground truth and the proposed model predictions. The second loss is computed between the drivable area predictions and ground truth. The total loss of our solution can be calculated as Eq. 4.2.

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{CE}_{drivable\_area} + \beta \cdot \mathcal{CE}_{moseg}, \tag{4.2}$$

4.3 The Dataset and Training Details

4.3.1 Dataset Information

In our work, we apply our experiments on the public autonomous driving data set, Nuscenes [8], which is one of the largest multi-modal datasets. Nuscenes data set contains $1,000$ autonomous driving scenes with detailed annotations using six cameras, one 32-beam LiDAR, and radars. There are 850 scenes with ground truth labels, and we filter all the attributes and project the matched 3-D bounding box onto the BEV plane to obtain 2D polygons to generate moving-obstacle labels. The sparse depth information is generated from the LiDAR points clouds. We randomly disorder 850 scenarios to split the training, validation, and testing sets.

4.3.2 Experimental Setup

We implement our framework by Pytorch [74] framework. The initial learning rate for our DPMoSeg is set as 1e-3. We use the Adam optimizer and the MultiStepLR

learning rate adjusting strategy.

4.4 Comparative Study and Results

4.4.1 Comparative Study

In this section, we built several baselines to verify the effectiveness of the proposed framework. The first baseline is modified from our network. The last two baselines are built on LSS [76] and PowerBEV [61], respectively. We adopt two metrics (Precision and intersection- over-union [23]) for the quantitative evaluations.

- **DPMoSeg-sparse**: We built this baseline based on the proposed DPMoSeg. This baseline uses the sparse depth generated from corresponding LiDAR point clouds as $\mathcal{D}_{depth}$ directly. The rest of the architectural structure remains the same.

- **LSS-MoSeg**:This baseline is built based on the LSS method [76]. We use the continuous 2-timestamp captured images as input.

- **PolarBEV-MoSeg**:We built this baseline based on the PolarBEV [61]. This baseline input is single-timestamp six-view camera images. We modified it to the temporal inputs.

From Table. 4.1, we can see that our DPMoSeg outperforms the other baselines in terms of precision and IoU. As illustrated in Fig. 4.3, the qualitative comparisons also highlight the advantages of our network over existing methods. To maintain a fair comparison, we adopt the same examples from split test scenarios for the different baselines.

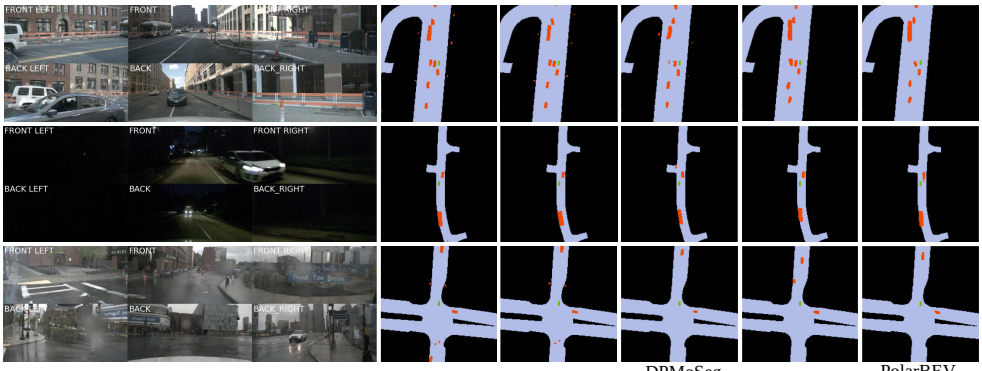

Figure 4.3: Comparative results of the baseline methods. The ■ pixels represent the predictions of the drivable area.

Table 4.1: Comparative experimental results on several baselines.

Time Intervals	Baseline Type	IoU (%)	Precision (%)
2	DPMoSeg (Sparse)	30.65	53.04
2	LSS-MoSeg	27.24	38.56
2	Polarbev-MoSeg	26.18	38.30
2	DPMoSeg (Ours)	43.94	66.08

4.4.2 The Qualitative Demonstrations

Sample qualitative BEV moving segmentation generation results and drivable area predictions are shown in Fig. 4.4. In general, we show visualization examples of our solution under different weather and light conditions: daytime, night, and rainy days. Meanwhile, the IoU of our auxiliary task, drive-able area segmentation, is 82.30%.

4.5 Ablation Study

In this section, we built several ablation studies to demonstrate the effectiveness of our proposed framework.

Figure 4.4: Example qualitative performances of our proposed DPMoSeg. The ■ pixels represent the drivable area region. The ■ pixels show our moving obstacles prediction results.

4.5.1 Performance of Various Feature Extractors

Here, we perform the ablation study to select the optimal feature extractor for the proposed method. The results are presented in the Table. 4.2. It can be illustrated that our DPMoeg reached the best performance. Therefore, in this paper, the modified EfficientDet module is the feature extractor of our method.

4.5.2 Performance of Sparse Depth-powered Strategy

Here, we built the ablation study to display the performance of our proposed DPMoSeg. The experimental results are in Table. 4.3. The variant without SDA module refers that the visual depth generated from the images is used as $\mathcal{D}_{depth}$. We can see that more accurate depth information is predicted by the SDA model,

Table 4.2: Results of the ablation study on different extractors of the proposed method.

Time Intervals	Extractor	IoU (%)	Precision (%)
2	ResNet18	41.80	63.66
2	Efficient-B0	31.26	54.12
2	ResNet50	42.09	60.33
2	DPMoSeg (ours)	43.94	66.08

Table 4.3: Results of the ablation study on different components of the proposed method.

Time Intervals	SDA	Sub-task	IoU (%)	Precision (%)
2	✗	✓	29.91	51.87
2	✓	✗	42.90	63.93
2	✓	✓	43.94	66.08

which, in turn, produces more accurate BEV features.

4.5.3 Performance of the Incorporation of Auxiliary Task

Here, we design the experimental variant to demonstrate the performance of the auxiliary task. Table. 4.3 displays the performance of the addition of sub-task, drive-able area segmentation. It can be illustrated that drive-able area segmentation tasks can boost the performance of the moving obstacle segmentation.

4.6 Robustness for different conditions

In this section, we present our proposed DPMoSeg performance under different conditions: daytime, night, and rainy. These scenes are further organized according to the weather and light conditions in the divided test scenarios. We aligned the non-night scenes of cloudy, sunny, and cloudy weather into the daytime condi-

tions. The performance is shown in Tab. 4.4. It can be seen that our model shows robustness in moving-obstacle segmentation task under the different weather conditions.

Table 4.4: Robustness results under different weather and light conditions.

Time Intervals	Weather Type	IoU (%)	Precision (%)
2	Ours (Daytime)	43.73	65.94
2	Ours (Night)	45.85	64.13
2	Ours (Rainy)	39.58	60.44
2	Ours (Total)	43.94	66.08

4.7 Conclusion

In this chapter, we have presented a novel and effective network for autonomous driving moving-obstacle segmentation under the birds-eye view in an end-to-end online fashion. We propose solving the moving-obstacles segmentation with a sparse-dense attention module and incorporating an auxiliary task: drivable area segmentation. We demonstrate the effectiveness and superiority of our DPMoSeg through extensive experiments. Our results suggest the value of our DPMoSeg in providing moving-obstacles information for autonomous driving.

Chapter 5

Foundation Model-assisted Explainable Vehicle Behavior Decision Making

5.1 Introduction

Human-related issues, including fatigue and impaired driving, are among the leading contributors to traffic accidents, posing significant obstacles to road safety [111]. To mitigate such risks, autonomous driving technologies have been introduced as a viable means of improving operational reliability and ensuring safer road environments. Recently, the integration of deep learning techniques into autonomous driving pipelines has driven substantial progress to enhanced automation and reduced accident rates. These systems typically encompass a range of core functionalities, including semantic segmentation [66], localization [122, 46], and decision-making [79]. Among these, driving behavior decision-making plays

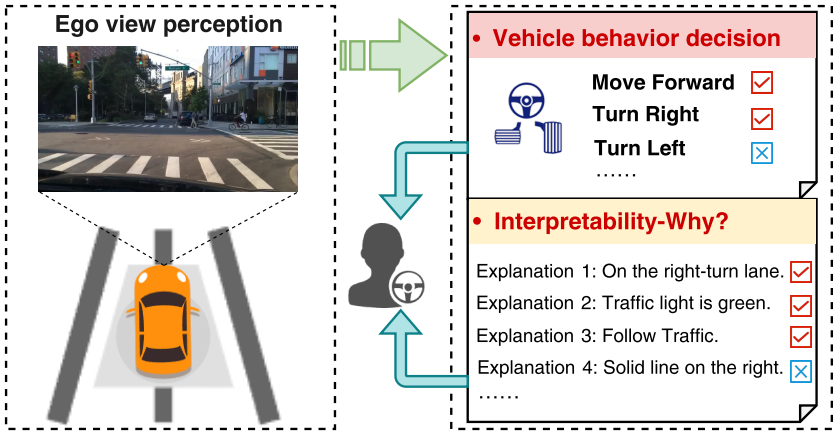

Figure 5.1: Paradigm of our framework.

a pivotal role in ensuring safe and efficient navigation, and is fundamental to the broader deployment and societal acceptance of autonomous vehicles.

Effective decision-making in autonomous driving critically depends on the systems ability to perceive and comprehend its environment. Recent deep learning-based strategies typically follow either a modular pipeline design or an end-to-end learning approach [93, 14]. In pipeline-based frameworks, decision-making is divided into multiple stages, each handled by a dedicated module in a sequential manner. The outputs from these modules are then integrated to inform the final driving decision. However, the inherent sequential dependency in these systems makes them susceptible to error accumulation, where inaccuracies in upstream modules can propagate and amplify through the pipeline, ultimately degrading the reliability of predictions. Unlike modular pipelines, end-to-end approaches learn a direct mapping from visual observations to driving decisions, thereby alleviating error propagation across intermediate stages. Many studies have focused on improving behavior prediction within end-to-end paradigms, achieving superior performance compared to traditional pipeline-based systems [12, 25].

Recent advances in deep learning architectures, including Convolutional Neural Networks (CNNs) and Transformers [90, 123], have greatly enhanced the capabilities of perception and decision-making tasks in autonomous driving. Nevertheless, the black-box nature of these models remains a significant obstacle to broader deployment, as it hinders interpretability, reduces user confidence, and complicates the understanding of model behavior. Enhancing model transparency and interpretability is crucial for the safe and reliable deployment of autonomous driving systems. Although various interpretability techniques have been proposed, such as attention mechanisms [116], saliency maps [42, 69], cost-volume analysis [112], and auxiliary task integration [39, 26], these methods often offer limited explanatory power. Moreover, they typically require extensive annotated data, making them difficult to scale to complex, real-world driving environments. Some works have investigated the use of natural language to explain ego-vehicle behaviors [45], presenting a promising direction for enhancing the interpretability of autonomous driving systems. However, such methods frequently generate lengthy or overly generic explanations, which can elevate the cognitive load of users and contribute to psychological fatigue. In addition, some studies [104, 26] have framed the interpretation of vehicle behavior as a natural language classification problem. While effective to some extent, these methods typically rely on pre-training procedures, leading to multi-stage pipelines that complicate the deployment and reduce real-time interpretability.

To address the aforementioned limitations, we present a novel end-to-end multi-task framework for interpretable vehicle behavior prediction, which incorporates a self-supervised, class-agnostic object segmentation component and a feature integration strategy. To ensure the interpretability of driving decisions is both concise

and informative, the on-hand tasks of driving decision-making and explanation are reformulated as classification problems, rather than relying on the generation of complex natural language descriptions. This approach paradigm is illustrated in the Fig. 5.1.

To be specific, our framework, termed VB-CASeg (Vehicle Behavior with Class-Agnostic Segmentation), takes visual input in the form of video clips and jointly predicts both the ego-vehicle behavior and a set of plausible, language-based explanations. The first part of the framework is a self-supervised, class-agnostic object segmentation component, which leverages a foundation model and incorporates a compact and efficient 2D adapter design. Specifically, our segmentation module embeds comprehensive object-level scene information in a self-supervised manner, without relying on any additional annotations. It enables the extraction of class-agnostic, object-aware representations that capture rich contextual cues from the surrounding environment. A semantic extractor part is employed to get hierarchical, semantic-aware feature representations. Next, we present the Class-Agnostic and Semantic Feature fusion component (CA-SF), which enhances the global contextual representations by combining early fusion techniques, self-attention mechanisms, and Fourier convolution operations. At last, behavior predictions alongside the respective human-interpretable explanations are produced jointly via specialized action prediction and explanation heads. Extensive experiments on publicly available datasets validate the effectiveness and superiority of the VB-CASeg framework on both behavior prediction and interpretable explanation prediction.

5.2 Methodology

5.2.1 Architecture

Fig. 5.2 presents an overview of our architecture. The proposed framework is designed to jointly predict vehicle behavior decisions and their respective explanations by leveraging a self-supervised class-agnostic object segmentor module and the CA-SF module. The entire process is formulated as an end-to-end multi-task learning paradigm. Specifically, our visual input is processed along two parallel streams. In one stream, the input is passed through a semantic extractor to generate semantic-aware hierarchical representations. In the other stream, both the original and augmented visual inputs are processed by a self-supervised class-agnostic object segmentor module, which captures object-level cues without relying on category-specific annotations. The class-agnostic object segmentor module generates object-aware features and segmentation predictions using SAM in conjunction with a lightweight adapter mechanism. This module is trained in a self-supervised manner to enforce consistency between the object features and their respective segmentation outputs, without requiring explicit labels. Leveraging both the semantic-aware representations and class-agnostic object representations, we introduce the CA-SF module to integrate these complementary cues. This module produces refined fused feature maps that serve as comprehensive global representations of the surrounding environment. Finally, based on the integrated global features, the model simultaneously predicts the vehicle action decision and associated explanations. In this work, both vehicle behavior prediction and interpretability generation are formulated as classification tasks.

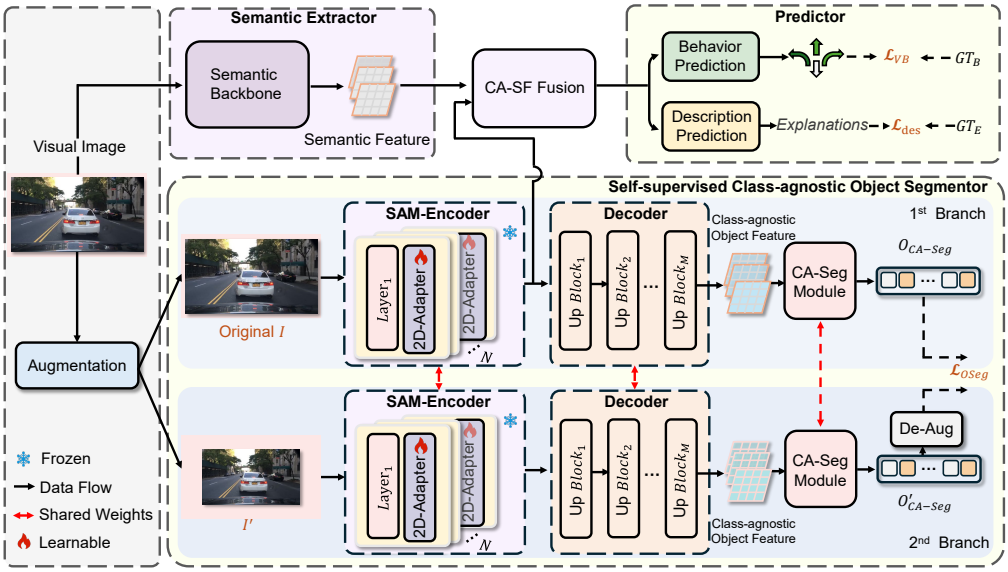

Figure 5.2: Overall structure of our VB-CASeg.

5.2.2 Semantic Extractor

Semantic-aware features have been widely utilized in strengthening feature representations and improving the performance of visual models in diverse tasks [83]. To effectively extract semantic cues, we modify the ResNet [31] architecture as our semantic backbone. This variant consists of multiple residual blocks and down-sampling operations, omitting the final global pooling and fully connected layers to preserve spatial information.

Given an input image I, we reduce the input resolution to $1/32$ of its original size using stacked convolutional and down-sampling layers to extract hierarchical features. Specifically, we extract multi-scale semantic representations from Layer1 to Layer4 of the backbone and the final semantic feature map has a channel dimension of 2048.

5.2.3 Self-supervised Class-agnostic Segmentor

SAM Encoder

To leverage its segmentation capabilities while minimizing training overhead, SAM is commonly used in a frozen state and adapted to downstream tasks without fine-tuning its internal parameters. In our framework, we adopt Fast-SAM [117], a lightweight and efficient variant of SAM, as the backbone of our class-agnostic object segmentor, referred to as the SAM-Encoder. Considering both inference efficiency and memory constraints, we adopt Fast-SAM [117] as the backbone of our class-agnostic segmentation module, referred to as SAM-Encoder. It is used as the encoder in both the first and second branches, as illustrated in Fig. 5.2. We do not train the SAM component from scratch. Instead, we directly utilize the pre-trained encoder and freeze its parameters throughout training. Furthermore, no prompt information is introduced into the SAM module.

Adapter Module

Although SAM exhibits strong generalization ability, its direct deployment in downstream tasks, such as ego-car behavior prediction, is constrained by domain-specific challenges and limited task-specific data. To address these limitations and effectively leverage the rich knowledge embedded in SAM, we introduce a lightweight adapter module. This adapter serves as a bridge, enabling the transfer of the segmentation capabilities of SAM to our Self-Supervised Class-Agnostic Segmentor (S2CASeg). By incorporating this adapter, our model benefits from the general-purpose object-level understanding of SAM while adapting to the ego-car behavior decision-making for autonomous driving scenarios. Drawing inspiration from Convpass [38], we introduce a lightweight 2D adapter module, integrated into

every block of the SAM framework. Structurally, the adapter component includes a depth-wise convolution, an activation function, and an additional depth-wise convolutional operation. The initial convolutional layer serves to project and reduce feature dimensionality, introducing a bottleneck structure for efficient representation learning. This configuration ensures a balance between model complexity and adaptability, allowing the adapter to incorporate non-linear transformations critical for domain-specific learning. During training, only the parameters of the adapter are updated, while the original SAM remains frozen. The resulting feature representations through the adapter are formulated as:

$$\mathcal{X}^{\cdot\cdot} \leftarrow \mathcal{X} + \boldsymbol{C}\left(\sigma\left(\boldsymbol{C}\left(\sigma\left(\boldsymbol{C}\left(\mathcal{X}\right)\right)\right)\right)\right), \tag{5.1}$$

where σ denotes the activation function, $\boldsymbol{C}$ denotes convolutional layer.

Specifically, the designed lightweight adapter module is integrated into each stage of the SAM encoder to effectively learn hierarchical representations. The output channel dimensions of the adapter modules are set to 256, 256, 128, and 64, respectively, across the four stages. It balances model capacity and computational efficiency while ensuring effective multi-scale feature integration.

Decoder Module

To achieve more detailed feature representations, a specialized decoder is employed for the class-agnostic object segmentor. It comprises four stacked decoder blocks, each responsible for successive refinement and resolution enhancement of the feature maps. Each decoder block comprises a combination of convolutional layers, a transposed convolutional layer for upsampling, normalization, an activation function, and a residual connection. To improve computational efficiency and

reduce memory usage, the segmentation predictions are generated at half the resolution of the original input. The final output of decoder has a channel dimension of 256. It is important to note that all parameters in the decoder module are trainable.

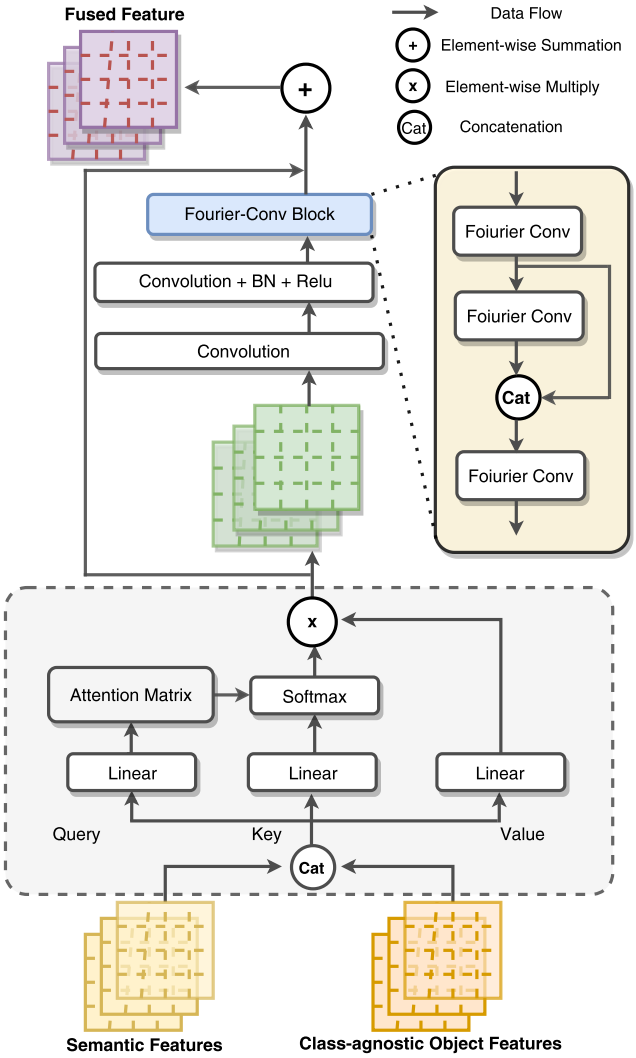

Figure 5.3: Architecture of the class-agnostic & semantic feature fusion component.

Self-supervised Segmentation Mechanism

Given the absence of annotations on class-agnostic object segmentation and

to reduce manual labeling efforts, we adopt a self-supervised learning strategy to extract class-agnostic object cues. To this end, our proposed module incorporates a dual-branch architecture, as illustrated in Fig. 5.2. In particular, the original visual data is fed into the first branch, while its augmented counterpart is handled by the second branch. This bifurcation enables consistency learning between the two branches, allowing the model to effectively learn object-level features without relying on additional annotations. All the branches utilize the same encoder, decoder, and shared parameter weights. For effective training and coverage, S2CASeg utilizes scale equivariance by conducting similarity assessments directly on feature representations.

5.2.4 Class-agnostic & Semantic Feature Fusion

To enhance feature representations for ego-car behavior and interpretability prediction, we propose a Class-Agnostic & Semantic Feature Fusion module (CA-SF). An overview of the architecture is shown in Fig. 5.3. The CA-SF module is designed to produce reliable fused feature representations that serve as global contextual cues. It comprises a self-attention block, standard convolutional layers, a Fourier block, and residual connections. To integrate complementary visual cues, the self-attention block receives a concatenation of class-agnostic features F_C and semantic features F_S as input. The intermediate output of the attention mechanism is given by:

$$F^{\cdot\cdot} = \text{Softmax}\left(\text{Norm}\left(QK^T\right)\right) \cdot V \tag{5.2}$$

where Q, K and V represent the feature concatenation of (F_S, F_C). The Fourier block is composed of a fast Fourier convolution layer [18] and a non-linear activa-

tion function. It is designed to effectively capture both local and global contextual features. By encoding input features in the frequency space, the Fourier block produces robust fused feature maps, denoted as F_g. It is important to note that the class-agnostic object feature maps used as input to the CA-SF module are derived from the last layer of the first branch in the S2CASeg module.

5.2.5 Vehicle Behavior and Explanation Prediction

The ego-car behavior decision and the related language-based explanation predictions are generated based on the fused feature representations F_g. The ego-car behavior decision head is composed of sequential convolutional layers, with the output dimension corresponding to the predefined set of behavior categories. This module is designed to support both single-label (one action) and multi-label (multiple actions) prediction scenarios. Similarly, the interpretation prediction head consists of multiple convolutional layers, with the output dimension matching the number of predefined language-based interpretation classes.

5.2.6 Loss Function

We introduce a multi-task loss function to supervise the training of the VB-CASeg framework, comprising three key components. As discussed, VB-CASeg jointly outputs the ego-car behavior prediction with its aligned language-based explanations. In addition to these two supervised tasks, we introduce a self-supervised learning strategy to optimize the class-agnostic object segmentor. The first component is ego-car behavior prediction loss, $\mathcal{L}_{VB}$. The second component is the behavior language-based explanation loss $\mathcal{L}_{des}$. The third component is class-agnostic

object segmentation loss, $\mathcal{L}_{OSeg}$. $\mathcal{L}_{VB}$ and $\mathcal{L}_{des}$ are both the binary cross-entropy losses for ego-car behavior decisions and aligned explanation prediction. The related loss functions are given by:

$$\mathcal{L}_{VB} = \sum_{i=1}^{N} \mathcal{L}\left(\hat{B}_i, B_i\right), \tag{5.3}$$

$$\mathcal{L}_{des} = \sum_{i=1}^{C} \mathcal{L}\left(\hat{E}_i, E_i\right), \tag{5.4}$$

where $\mathrm{B_i}$, $\mathrm{E_i}$ are the annotations of ego-car action and associated explanations. $\widehat{\mathrm{B}}_\mathrm{i}$ and $\widehat{\mathrm{E}}_\mathrm{i}$ denote the predictions for ego-car action and respective explanations. N and C denote the numbers of ego-car behavior and explanation categories, respectively, while $\mathcal{L}$ indicates the binary cross-entropy loss.

Besides, it is noteworthy that the segmentation loss $\mathcal{L}_{OSeg}$ is to measure and constrain the class-agnostic object segmentor from the two branches in S2CASeg, which is given by:

$$\mathcal{L}_{OSeg} = \frac{1}{2}\left(\mathcal{L}_M(\sigma(O), \sigma(\ddot{A}(\ddot{O}))) + \mathcal{L}_M(\sigma(\ddot{A}(\ddot{O})), \sigma(O))\right), \tag{5.5}$$

where $\ddot{A}$ corresponds to the inverse augmentation operation, O and $\ddot{O}$ represent the class-agnostic object segmentation outputs of the first and second branches, respectively, σ is activation function, and $\mathcal{L}_M$ is mean squared error (MSE) function.

Thus, the final loss function of our framework is formulated by the following expression:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{VB} + \beta \cdot \mathcal{L}_{des} + \lambda \cdot \mathcal{L}_{OSeg}, \tag{5.6}$$

where α, β, and λ denote the weight hyper-parameters of ego-car behavior, language-based explanation predictions, and self-supervised class-agnostic object segmentor.

5.3 Training Details and Evaluation Metrics

5.3.1 Datasets

We evaluate our method on publicly available autonomous driving datasets, specifically BDD Object Induced Actions (BDD-OIA) [104] and BDD Actions and Descriptions (BDD-AD) [26], to demonstrate its effectiveness.

BDD-OIA: This public dataset can be viewed as a curated subset of BDD100K [109], comprising $22,924$ ego-view video clips manually annotated. Each clip contains at least five pedestrians or bicyclists and more than five vehicles to ensure complex traffic scenarios. The videos are captured under varying weather and lighting conditions. The dataset provides annotations for four types of driving actions alongside twenty-one categories of natural-language, human-defined explanations.

BDD-AD: Derived from BDD100K [26], this dataset contains $100,000$ videos annotated with image-level labels, tailored for autonomous driving research. It includes 4 types of behavior decisions and 6 categories of environmental descriptions, such as front area is free of obstruction and left/left-turn area is clear. The dataset is split into $5,000$ training samples, $2,500$ validation samples, and $2,500$ test samples.

5.3.2 Implementation Details

Our framework is developed using PyTorch [74]. The network is trained using the stochastic gradient descent (SGD) optimizer to accelerate convergence. The fused feature maps generated by the CA-SF module have 256 channels. Notably, our framework is an end-to-end solution that does not require any pre-training or additional segmentation labels.

5.3.3 Evaluation Metrics

The F1 score, a widely adopted metric in behavior decision-making and explanation prediction tasks [104, 116], is used to quantitatively evaluate the performance of our method. We compute two types of F1 scores: the overall F1 score and the mean F1 score. The overall F1 scores for ego-car behavior prediction and possible associated language-based explanation generation are computed as:

$$\mathrm{F}_{\text{oval}}^{\text{behavior}} = \frac{1}{N} \sum^{N} \mathrm{F1}\left(\hat{\boldsymbol{B}}_i, \boldsymbol{B}_i\right), \tag{5.7}$$

$$\mathrm{F}_{\text{oval}}^{\text{des}} = \frac{1}{C} \sum^{C} \mathrm{F1}\left(\hat{\boldsymbol{E}}_i, \boldsymbol{E}_i\right), \tag{5.8}$$

where $\mathrm{B_i}$ and $\mathrm{E_i}$ denote the ground truth labels for ego-car behavior and the respective explanations, respectively, while $\widehat{\mathrm{B}}\mathrm{i}$ and $\widehat{\mathrm{E}}\mathrm{i}$ represent the predicted behavior and explanation outputs.

The mean F1 scores for both ego-car behavior decisions and their associated natural language explanations are calculated as:

$$\mathrm{F1}_{\mathrm{m}}^{\mathrm{behavior}} = \frac{1}{N}\left(\mathrm{F1}_{\mathrm{F}} + \mathrm{F1}_{\mathrm{S}} + \mathrm{F1}_{\mathrm{L}} + \mathrm{F1}_{\mathrm{R}}\right), \tag{5.9}$$

$$\mathrm{F1}_{\mathrm{m}}^{\mathrm{des}} = \frac{1}{C}\sum_{j=1}^{C} \mathrm{F1}_{\mathrm{j}}^{\mathrm{des}}, \tag{5.10}$$

where $\mathrm{F1}_{\mathrm{F}}$, $\mathrm{F1}_{\mathrm{S}}$, $\mathrm{F1}_{\mathrm{L}}$, and $\mathrm{F1}_{\mathrm{R}}$ represent F1 scores related to the ego-car behavior decisions: move forward, stop/slow down, turn left/change to left lane and turn right/change to left lane. $\mathrm{F1}_{\mathrm{j}}^{\mathrm{des}}$ represents the prediction score for individual language-based explanation category. Here, *N* corresponds to the number of ego-car behavior categories, and *C* corresponds to the number of possible explanation categories associated with ego-car behavior.

Table 5.1: Comparative results on the BDD-OIA dataset.

	$\mathrm{F1}_{\mathrm{m}}^{\mathrm{behavior}}$	$\mathrm{F1}_{\mathrm{oval}}^{\mathrm{behavior}}$	$\mathrm{F1}_{\mathrm{m}}^{\mathrm{des}}$	$\mathrm{F1}_{\mathrm{oval}}^{\mathrm{des}}$	F	S	L	R
Local selector [99]	0.699	0.711	0.196	0.406	0.810	0.762	0.600	0.624
OIA [104]	0.718	0.734	0.208	0.422	0.829	0.781	0.630	0.634
Inaction [39]	0.694	0.714	0.347	0.565	0.800	0.747	0.612	0.619
CBM [47]	0.610	0.661	0.292	0.412	0.795	0.732	0.483	0.431
CBM-AUC [86]	0.658	0.704	0.342	0.522	0.803	0.751	0.551	0.525
C-SENN [85]	0.618	-	-	-	0.772	0.744	0.469	0.486
NLE-DM [26]	0.723	0.733	0.312	0.517	0.827	0.760	**0.651**	0.653
Interrelation [116]	0.701	0.722	0.335	0.537	0.802	0.753	0.619	0.625
VB-CASeg (Ours)	**0.734**	**0.753**	**0.384**	**0.573**	**0.829**	**0.798**	0.649	**0.660**

Table 5.2: Results of the ablation study evaluating the impact of different semantic encoders.

Variant No.	Backbone	$\mathrm{F1}_{\mathrm{m}}^{\mathrm{behavior}}$	$\mathrm{F1}_{\mathrm{oval}}^{\mathrm{behavior}}$	$\mathrm{F1}_{\mathrm{m}}^{\mathrm{des}}$	$\mathrm{F1}_{\mathrm{oval}}^{\mathrm{des}}$	F	S	L	R
Variant 1	ResNet-18	0.724	0.745	0.368	0.553	0.830	0.791	0.634	0.641
Variant 2	ResNet-34	0.727	0.746	0.385	0.563	0.825	0.789	0.647	0.645
Variant 3	ResNet-50	**0.734**	**0.753**	**0.384**	**0.573**	**0.829**	**0.798**	**0.649**	**0.660**
Variant 4	RegNet-x-400mf	0.716	0.733	0.371	0.549	0.820	0.778	0.649	0.617
Variant 5	RegNet-x-800mf	0.722	0.741	0.372	0.556	0.829	0.784	0.639	0.637
Variant 6	RegNet-y-400mf	0.720	0.740	0.376	0.552	0.825	0.784	0.630	0.641
Variant 7	RegNet-y-800mf	0.723	0.747	0.371	0.560	0.826	0.794	0.635	0.636
Variant 8	EfficientNet-B0	0.721	0.739	0.368	0.558	0.825	0.791	0.630	0.637
Variant 9	EfficientNet-B1	0.724	0.741	0.363	0.554	0.828	0.788	0.640	0.641
Variant 10	MobileNetV3-small	0.702	0.720	0.340	0.520	0.810	0.764	0.616	0.620
Variant 11	MobileNetV3-large	0.718	0.736	0.370	0.545	0.826	0.779	0.629	0.639

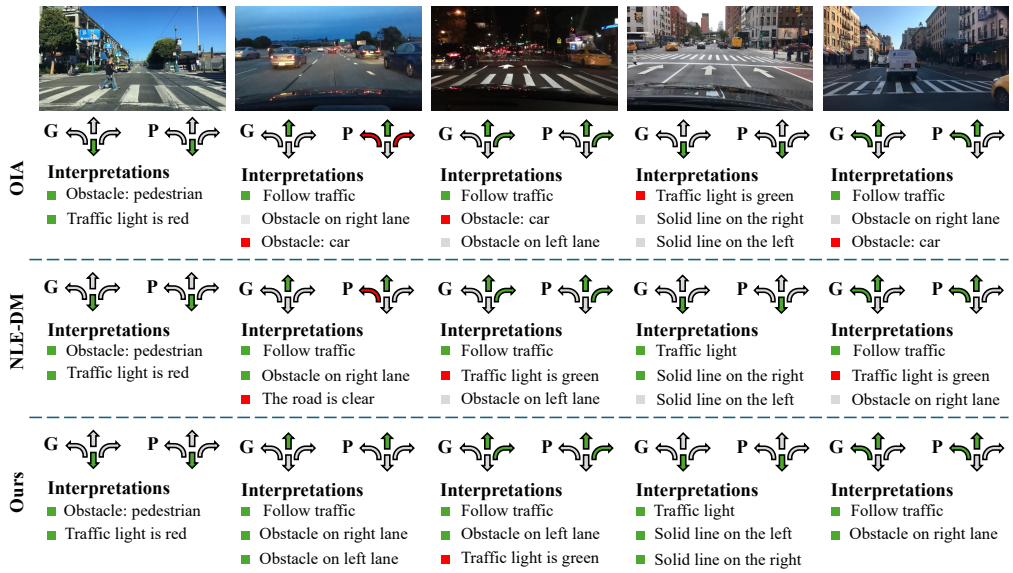

Figure 5.4: Some representative examples of ego-car behavior decisions and associated language-based explanations under the BDD-OIA dataset.

5.4 Comparative Study

5.4.1 Performance

We compare the VB-CASeg framework with several state-of-the-art approaches to further validate the effectiveness. Given the same visual inputs, we evaluate vehicle behavior decisions and respective explanation predictions across multiple methods. Specifically, we compare against the following baselines: Local Selector [99], OIA [104], Inaction [39], CBM [47], CBM-AUC [86], C-SENN [85], NLE-DM [26], and Interrelation [116]. The OIA network [104] utilizes object proposals generated by a backbone for local feature extraction combined with global scene reasoning. This approach relies on a pre-trained model, Faster R-CNN. NLE-DM [26] incorporates semantic segmentation and produces both behavior and language-based explanations, requiring pre-training of the segmenta-

tion module. Similarly, Interrelation [116] leverages a pre-trained object detection module to guide the generation of both decisions and associated explanations. Table 5.1 presents the comparative results of the experimental evaluation. It is apparent that VB-CASeg framework attains the best overall performance in the context of multi-task learning. Specifically, for vehicle behavior decision-making, VB-CASeg consistently outperforms existing methods in terms of both mean F1 score and overall F1 score. The results highlight the effectiveness of our proposed approach in accurately predicting driving behaviors while maintaining high reliability across different behavior categories. For the behavior interpretation predictions, we observe that our method achieves superior performance compared to other state-of-the-art approaches, such as OIA, NLE-DM, Interrelation, and Inactionparticularly in terms of the $\mathrm{F1}_{\mathrm{m}}^{\mathrm{des}}$ and $\mathrm{F1}_{\mathrm{oval}}^{\mathrm{behavior}}$ metrics. We hypothesize that it arises from the architectural differences in feature utilization. While the comparative methods tend to rely directly on the final feature representations of their respective models, our approach benefits from the integration of class-agnostic and semantic-aware cues through the fused global features, enabling more robust and informative decision-making.

Figure 5.4 illustrates the predicted ego-car behavior decisions alongside their associated explanations. In the second column, both OIA and NLE-DM incorrectly predict the driving actions and their associated explanations. The fourth column illustrates a scenario at a crossroads, where both NLE-DM and OIA also fail to produce accurate explanations. Specifically, the OIA method generates the description the traffic light is green, whereas the actual traffic light is yellow. In contrast, our method successfully predicts both the correct driving action and accurate descriptions, From the fifth column, our method accurately predicts both the

Table 5.3: Ablation study evaluating the impact of self-supervised class-agnostic object segmentation module.

	Segmentation	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$
Variant 1	✗	0.708	0.730	0.311	0.497
Variant 2	CASeg	0.716	0.738	0.358	0.542
Ours	S2CASeg	**0.734**	**0.753**	**0.384**	**0.573**
	Segmentation	F	S	L	R
Variant 1	✗	0.814	0.779	0.620	0.620
Variant 2	CASeg	0.819	0.788	0.629	0.627
Ours	S2CASeg	**0.829**	**0.798**	**0.649**	**0.660**

vehicle behavior and the respective explanation. In contrast, the OIA and NLE-DM methods can not percept the obstacles present in the right lane. As shown in Fig. 5.4, even without incorporating auxiliary tasks, our model successfully perceives lane markings. We hypothesize that this improvement is largely due to the integration of the SAM module. In the final column, our method continues to provide accurate predictions of ego-car behavior and respective explanations. Furthermore, Fig. 5.5 presents visualizations of the segmentation outputs obtained from our self-supervised class-agnostic object segmentor. Notably, our segmentor is capable of detecting obstacles surrounding the ego vehicle, not only those in close proximity even under varying weather and lighting conditions.

5.5 Ablation Study

5.5.1 Analysis of Different Semantic Extractors

An ablation study is performed to determine the suitable semantic backbone for the semantic extractor module in VB-CASeg. Several variants are constructed

Table 5.4: Ablation study assessing the impact of various fusion strategies on the semantic and class-agnostic object feature representations.

Variant No.	Fusion Type	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$	F	S	L	R
Variant 1	Concat	0.725	0.738	0.367	0.550	0.822	0.779	0.647	0.654
Variant 2	CA-SConv	0.724	0.746	0.367	0.555	0.827	0.794	0.644	0.631
Variant 3	Addition	0.709	0.730	0.362	0.553	0.817	0.778	0.631	0.611
Variant 4 (Ours)	CA-SF	**0.734**	**0.753**	**0.384**	**0.573**	**0.829**	**0.798**	**0.649**	**0.660**

Table 5.5: Results of the ablation study assessing the contribution of the adapter module.

	Adapter	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$
Variant 1	Onelayer	0.710	0.734	0.357	0.542
Variant 2	Adapter-1*1	0.717	0.734	0.361	0.539
Variant 3	Adapter-Separable	0.730	0.747	0.371	0.543
Ours	Adapter	**0.734**	**0.753**	**0.384**	**0.573**
	Adapter	F	S	L	R
Variant 1	One-layer	0.818	0.788	0.623	0.612
Variant 2	Adapter-1*1	0.816	0.781	0.638	0.631
Variant 3	Adapter-Separable	0.824	0.794	0.641	0.660
Ours	Adapter	**0.829**	**0.798**	**0.649**	**0.660**

based on different backbone families, including the ResNet family, RegNet [78], EfficientNet [94], and MobileNet [33]. The experimental results, summarized in Table 5.2, demonstrate that the modified ResNet-50 achieves the best overall performance across all evaluation metrics, striking a favorable balance between memory consumption, the number of training parameters, and recognition accuracy.

Table 5.6: Ablation study evaluation the effects of various segmentation and adapter strategies.

Variant No.	Segmentation	Adapter	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$	F	S	L	R
Variant 1	CASeg	Onelayer	0.701	0.723	0.359	0.542	0.808	0.773	0.608	0.614
Variant 2	CASeg	Adapter-1x1	0.712	0.738	0.360	0.542	0.826	0.788	0.616	0.618
Variant 3	CASeg	Adapter-Separable	0.717	0.739	0.364	0.543	0.823	0.786	0.633	0.626
Variant 4 (Ours)	S2CASeg	Adapter	**0.734**	**0.753**	**0.384**	**0.573**	**0.829**	**0.798**	**0.649**	**0.660**

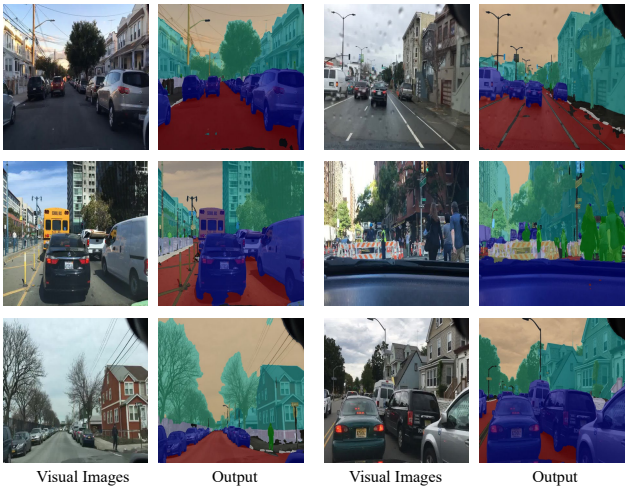

Figure 5.5: Example outputs from the self-supervised class-agnostic object segmentor module.

Therefore, unless otherwise specified, ResNet-50 is employed as the default semantic backbone in our implementation.

5.5.2 Effect on Self-supervised Class-agnostic Segmentor

To evaluate the effectiveness of self-supervised class-agnostic object segmentor module, we construct several ablation variants for the ablation study. In the first variant, the entire self-supervised class-agnostic object segmentor module is removed, leaving only the semantic extractor branch for decision-making. In the second variant, the second branch of the S2CASeg module is to be omitted, effectively disabling the self-supervised learning mechanism. This variant is referred to as CASeg.

The experimental results, presented in Table 5.3, indicate that the full S2CASeg module yields superior performance in both vehicle behavior prediction and interpretability generation tasks. Moreover, as illustrated in Fig. 5.4, our model

demonstrates the ability to perceive not only nearby but also distant obstacles. This improved perceptual capability can be attributed to the integration of the SAM module, which provides a large receptive field and enables a more comprehensive understanding of the driving scene.

Table 5.7: Ablation study results evaluating the relationship between behavior decisions and their respective descriptions.

Variant No.	Weight	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$	F	S	L	R
Variant 1	$\alpha:1, \beta:0.75, \lambda:0.25$	0.719	0.743	0.373	0.555	0.831	0.790	0.627	0.627
Variant 2	$\alpha:1, \beta:0.5, \lambda:0.5$	**0.734**	**0.753**	**0.384**	**0.573**	**0.829**	**0.798**	**0.649**	**0.660**
Variant 3	$\alpha:1, \beta:0.25, \lambda:0.75$	0.723	0.747	0.371	0.560	0.826	0.794	0.635	0.636
Variant 4	$\alpha:1, \beta:0, \lambda:1$	0.728	0.741	-	-	0.822	0.787	0.643	0.659
Variant 5	$\alpha:0, \beta:1, \lambda:1$	-	-	0.383	0.552	-	-	-	-

5.5.3 Analysis of Fusion Module

Here, we explore different fusion configurations to generate joint feature representations. Variant 1 employs a straightforward concatenation operation. Variant 2, denoted as CA-SConv, replaces the Fourier convolutional block with conventional convolutional layers to fuse the class-agnostic and semantic information. Variant 3 performs element-wise addition for feature fusion.

The results, summarized in Table 5.4, demonstrate that our CA-SF fusion module outperforms the other variants across the metrics. We attribute this performance gain to the integration of a Fourier-based convolution and self-attention mechanism, which enables the module to effectively capture global contextual features via frequency-domain representations while concurrently incorporating local information to model fine-grained spatial details.

5.5.4 Analysis on Adapter Module

An ablation study is carried out to assess the contribution of our adapter component, with three different variants designed for comparison. The first variant, referred to as Onelayer, adopts a conventional adapter composed of a single convolutional layer. Variant 2, named Adapter-1*1, replaces the 3*3 convolutional kernels with 1*1 kernels. Variant 3, Adapter-Separable, substitutes the 3*3 kernels with separable convolutions using 1*3 and 3*1 kernels. The results are summarized in Table 5.5. As shown, our adapter structure outperforms all other variants, particularly in explanation prediction. The results indicate that using 3*3 convolutional kernels yields superior performance compared to separable or smaller kernel configurations, as evidenced by the overall F1 score.

Table 5.8: Robustness demonstration over the BDD-AD dataset.

Method	$F1_{m}^{behavior}$	$F1_{oval}^{behavior}$	$F1_{m}^{des}$	$F1_{oval}^{des}$	EM Acc
OIA [104]	0.865	0.862	0.879	0.868	0.584
NLE-DM [26]	0.876	0.877	0.907	0.880	0.618
VB-CASeg (Ours)	**0.905**	**0.906**	**0.927**	**0.905**	**0.685**

Furthermore, we conduct trials to investigate the interplay among the segmentation module and the adapter strategy. Specifically, three model variants are built by jointly modifying various segmentation and adapter configurations. Table 5.6 shows the results. As observed, simultaneously altering both components results in a noticeable decline in performance for both vehicle behavior decision-making and interpretability prediction. This outcome supports our hypothesis that the segmentation module and adapter strategy serve complementary roles, each contributing critically to the overall effectiveness of our framework.

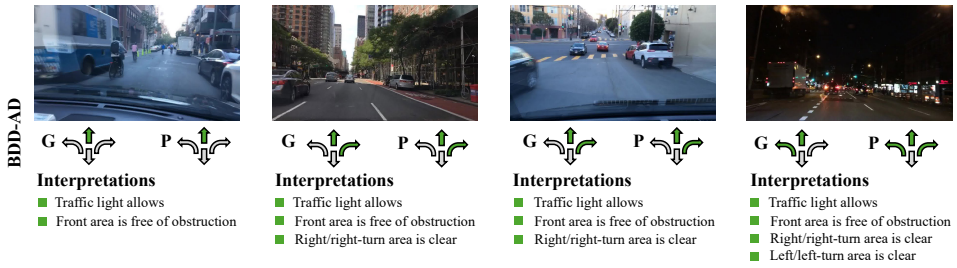

Figure 5.6: Some qualitative results over BDD-AD dataset under diverse conditions.

5.5.5 Effect over Vehicle Behaviors and respective explanations

Several variants are developed to examine the effects of hyperparameters and to analyze the correlation between vehicle behavior decisions and their associated explanations. We build multiple variants by varying the values of the loss function weights α, β, and λ. We fix the α at 1, while β and λ are constrained such that their sum equals 1, with increments of 0.25. Table 5.7 provided the results. Based on Table 5.7, we can observe that the combination of $(\alpha, \beta, \lambda) = (1.0, 0.5, 0.5)$ for ego-car behavior prediction and possible interpretation prediction, respectively, yields the best performance across the metrics $\mathrm{F1}_{\mathrm{m}}^{\mathrm{behavior}}$, $\mathrm{F1}_{\mathrm{oval}}^{\mathrm{behavior}}$, $\mathrm{F1}_{\mathrm{oval}}^{\mathrm{des}}$, and $\mathrm{F1}_{\mathrm{m}}^{\mathrm{des}}$.

To further evaluate the relationship between human-centric interpretation predictions and vehicle behavior decision-making, we design several variants that selectively omit specific tasks. The results are summarized in Table 5.7. In particular, Variant 4 evaluates the performance of the VB-CASeg approach when interpretation prediction is excluded, i.e., the model predicts only vehicle behaviors by setting the interpretation loss weight (β) to 0. Variant 5 examines the behavior explanation performance when the behavior prediction task is disabled, causing VB-CASeg to focus exclusively on predicting behavior explanations by setting the

weight parameter α to 0. The results show that our model maintains relatively robust performance in both ego-car behavior decision-making and associated explanation. Notably, when the behavior decision-making task is removed, the performance of behavior interpretation remains largely unaffected. However, excluding the behavior explanation prediction leads to a slight decline in the performance of behavior prediction. Based on Table 5.7, it shows that incorporating behavior interpretation prediction provides beneficial context for improving behavior decision accuracy. Accurate behavior explanations enhance the understanding of vehicle behaviors, thereby improving the reliability and interpretability of behavior predictions. This, in turn, fosters greater user trust and confidence in the system's decisions.

5.6 Robustness

To further assess the generalizability and effectiveness of VB-CASeg, we evaluate it on an additional publicly available autonomous driving dataset, BDD-AD. This dataset provides six categories of human-annotated rationale that serve as explanations for vehicle behavior decisions. Accordingly, we modify the output channel of the interpretation head to match the label format. For consistency, we adopt the same experimental settings used for the BDD-OIA dataset. Additionally, we measure the inference time and model size for all methods. The results are presented in Table 5.8. The results demonstrate that our method maintains strong robustness and generalizability across multiple publicly datasets. Additionally, qualitative results on the BDD-AD dataset are presented in Fig. 5.6. As illustrated, our model delivers consistent and reliable predictions under diverse scenarios and varying

lighting conditions, further validating its robustness and adaptability.

To better quantify the accuracy of explanation outputs, we utilize the Exact Match Accuracy (EM Acc) metric, which evaluates whether the predicted set of explanations exactly matches the human-annotated ground truth. This metric reflects real-world expectations, where precise and complete interpretability is crucial for ensuring user trust and system reliability. In this context, EM Acc serves as a user-centered correctness metric. As shown in Table 5.8, our VB-CASeg demonstrates superior performance, indicating its effectiveness in generating accurate and trustworthy explanations for the ego-car behavior.

5.7 Conclusion

In this chapter, we present a novel framework that adopts a self-supervised class-agnostic object segmentor with a feature fusion strategy to perform driving behavior decision-making and enrich the explainability. The proposed segmentor is built upon a vision foundation model augmented with lightweight adapter modules and is capable of generating class-agnostic object representations of the surrounding environment without the need for additional annotations. Simultaneously, semantic information is extracted through a semantic extractor. To further enhance explainability, we introduce a class-agnostic and semantic feature fusion module that effectively combines object-level and semantic cues, enabling human-centric explanations of the ego-car actions. By leveraging both class-agnostic object features and semantic context, our model achieves more accurate behavior predictions and provides clear insights for the decision-making process.

Chapter 6

Contrastive Learning-based Place Descriptor Representation for Cross-modality Place Recognition

6.1 Introduction

As a fundamental perception capability, place recognition enables essential functionalities in embodied agents, including loop closure detection [11, 105] and 3D reconstruction [24]. Place recognition addresses the challenge of re-identifying visited locations by measuring similarity between a query and a reference LiDAR map, under diverse viewpoints and environmental changes. Such a process facilitates accurate position estimation in dynamic environments and supports the construction of drift-free maps, which is an essential requirement for robust navigation and mapping in real-world scenarios.

In recent years, much research has been devoted to LiDAR-based place recog-

nition [67, 51]. Despite their effectiveness, LiDAR-only approaches encounter notable challenges, including high deployment costs and inherent limitations of the sensor. To more cost-efficient and scalable, the localization of visual images within LiDAR point-cloud maps has attracted growing interest. This task is typically formulated as an image-to-point-cloud retrieval problem, where the objective is to retrieve the most relevant LiDAR point-cloud map corresponding to a given visual image query.

The image-to-point-cloud matching paradigm is more economical, yet remains difficult due to the substantial modality discrepancy, which hinders the construction of effective and consistent place representations. Owing to the distinct sensing characteristics of LiDAR and cameras, LiDAR primarily captures sparse geometric and depth information, whereas cameras provide dense, appearance-based data that is sensitive to lighting conditions. Furthermore, the unordered nature of point clouds contrasts with the structured grid of image pixels, further hindering direct feature alignment. To bridge this modality gap, recent works [103, 119] have explored transforming visual images into point-cloud-like representations via depth estimation. However, these methods typically involve multi-stage pipelines, making them susceptible to cumulative errors introduced at each processing stage.

Recent advances in place recognition have been largely driven by deep learning, particularly CNNs [13]. CNNs offer several advantages, such as local connectivity and weight sharing, which make them well-suited for capturing spatially localized features with high computational efficiency. However, they often fall short in modeling long-range dependencies. In contrast, Transformers excel at capturing global contextual information and long-range interactions, albeit at the cost of significantly increased computational complexity. More recently, Mamba

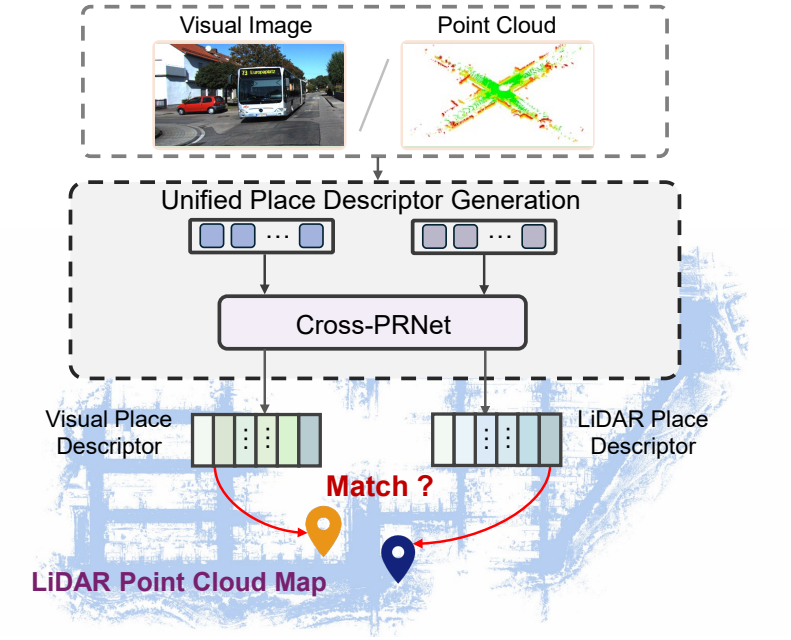

Figure 6.1: The paradigm of the framework, Cross-PRNet.

[114] has been introduced as a promising alternative, offering linear computational complexity. Nonetheless, its ability to fully exploit large receptive fields for visual understanding remains limited. These observations motivate the exploration of a unified paradigm that combines the strengths of Transformer and Mamba architecturesaiming to enhance cross-modal feature representation while maintaining an optimal balance between performance and computational efficiency.

To this end, we propose Cross-PRNet, a novel cross-modality place recognition network that integrates Transformer and Mamba architectures to learn unified feature representations for cross-modality localization via contrastive learning. The overall framework is depicted in Fig. 6.1. To address the challenge of generating a modality-invariant place descriptor, we design an end-to-end Siamese architecture that harnesses the complementary strengths of Transformer and Mamba modules. To reduce the modality gap, LiDAR point clouds are first projected into

range images, enabling a more structured representation. A dual-branch feature extraction module is utilized to effectively learn hierarchical representations from visual and point cloud inputs. Capturing both intra and inter-modal contextual dependencies, we introduce a Transformer-Mamba Mixer, which transforms these hierarchical features into compact place descriptors. Furthermore, a semantic-promoted descriptor enhancer is incorporated to refine these descriptors by leveraging estimated semantic distributions. Last, the generated global discriminative descriptors are optimized using a contrastive learning objective, enabling effective cross-modality place recognition through descriptor similarity.

6.2 Methodology

6.2.1 Overall Architecture

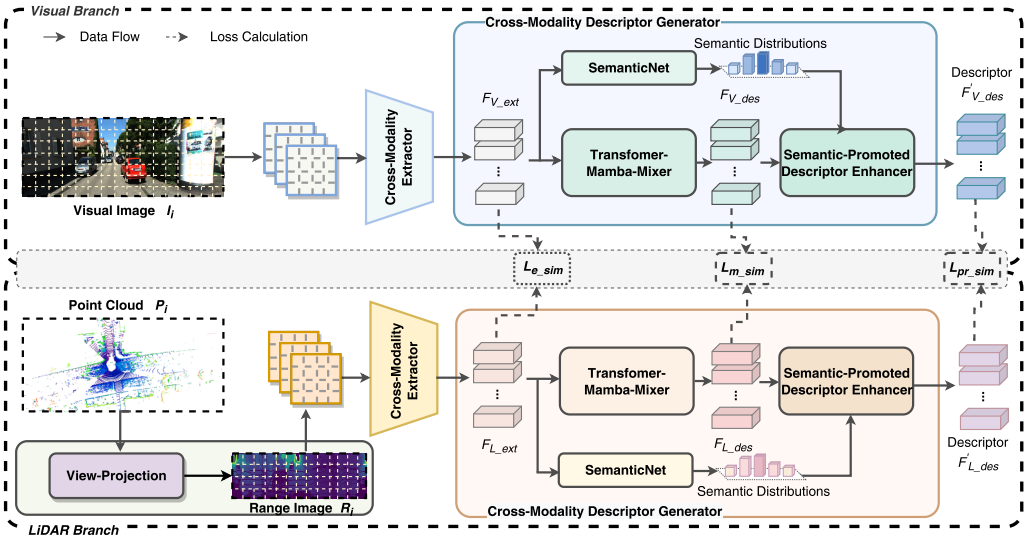

Figure 6.2: Overview of the proposed framework for cross-modality place recognition.

We propose a unified Siamese descriptor generation framework designed to capture cross-modality information. The architectural details are illustrated in Fig. 6.2. As shown, the framework comprises two parallel branches, each consisting of a Cross-Modality Extractor, Transformer-Mamba Mixer, and Semantic-Promoted Descriptor Enhancer. For the visual branch, hierarchical features are extracted from images, followed by semantic distribution estimation via the SemanticNet module, which provides high-level semantic guidance. These features are subsequently refined by the Transformer-Mamba Mixer and Semantic-Promoted Descriptor Enhancer to generate global descriptors. For the LiDAR branch, corresponding descriptors are generated using the same Siamese architecture applied to range images derived from point clouds. The generated descriptors are stored in database and used to retrieve the most similar information for a visual query, thereby enabling accurate image-to-point-cloud place recognition.

6.2.2 Data Process

To mitigate modality-specific discrepancies and enrich the quality of feature representations, modality-specific preprocessing steps are applied. We utilize spherical projection to convert the raw LiDAR scans P into structured range images, which serve as the input to the point cloud branch. Specifically, the 3D point $P = (x, y, z)$ is projected onto the 2D image coordinates (m, n) according to the following formulation:

$$\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left(1\text{-}\arctan(y,x)\pi^{\text{-}1}\right) w \\ \left(1\text{-}\left(\arcsin\left(zr^{\text{-}1}\right) + \mathrm{f}_{\mathrm{up}}\right)\mathrm{f}^{\text{-}1}\right) h \end{pmatrix}, \tag{6.1}$$

where (m, n) denote the 2D visual coordinates, $r = \left\|P_i\right\|_2$ represents the range value of the point cloud, h and w correspond to the height and width of the pro-

jected range image, respectively. Therefore, structured range images are generated from raw point clouds via spherical projection. For the visual branch, the input query consists of monocular images captured online from a front-view camera. To ensure computational efficiency and maintain consistency across modalities, both the projected range images and visual images are resized and cropped.

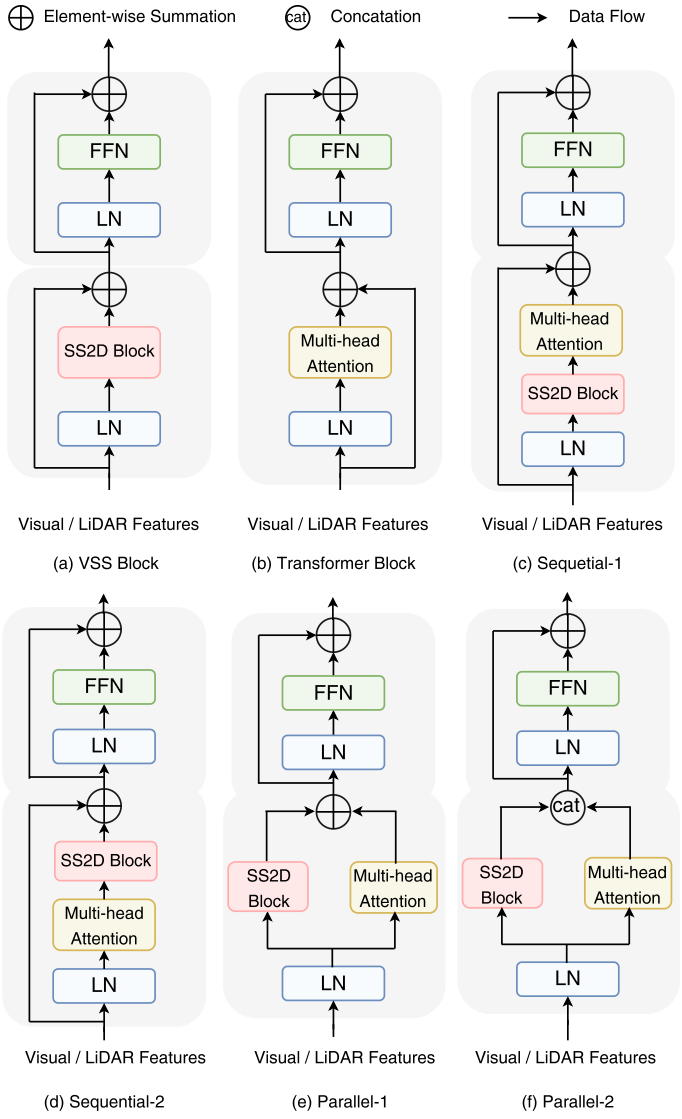

Figure 6.3: Paradigm of the Transformer-Mamba Mixer.

6.2.3 Cross-Modality Extractor

To capture multi-level information, a feature extraction backbone is applied to both the visual image I and the range image R, producing hierarchical feature representations. Each input image is first partitioned into patches, which are then processed through a series of stacked down-sampling blocks within the extractor to capture multi-scale contextual information. The final output features have a spatial resolution of $1/32$ of the original one. We utilize the features from the final layer of the extractor. Specifically, the visual feature representation is denoted as F_{V_ext}, while the corresponding spatial features extracted from the LiDAR point cloud are denoted as F_{L_ext}. The unified feature extractor facilitates consistent representations across both modalities, enabling effective cross-modal alignment.

6.2.4 Siamese Cross-Modality Descriptor Generator

Transformer-Mamba Mixer

To fully exploit the complementary strengths of Transformer and Mamba architectures, we introduce Transformer-Mamba Mixer for effective cross-modality place descriptor representation. We first provide a brief overview of the Mamba architecture, followed by detailed explanations of our Transformer-Mamba Mixer.

Formulation of SSM: Mamba can be interpreted as an enhanced solution that integrates a selection mechanism into the State Space Model framework. At its core, an SSM captures the relationship between input and output sequences through a latent dynamic state [114]. The continuous-time formulation of an SSM

can be expressed as follows:

$$\ddot{h}(t) = \mathrm{A}h(t) + \mathrm{B}x(t), \tag{6.2}$$

$$y(t) = \mathrm{C}h(t) + \mathrm{D}x(t), \tag{6.3}$$

This formulation defines a sequence-to-sequence mapping, wherein an input sequence $x(t)$ is transformed into an N-dimensional latent state $h(t)$, which is subsequently projected to produce output response $y(t)$. he term $\mathrm{D}x(t)$ functions as a skip connection. To integrate the state-space model (SSM) as a black-box module within deep learning frameworks, the matrices A, B, C, and D are optimized through gradient-based learning.

Mixer Strategy: To identify an effective integration strategy between the Transformer and Mamba blocks, we investigate four configurations based on either sequential or parallel fusion mechanisms. The detailed architectures of these combination strategies are illustrated in Fig. 6.3(c) to 6.3(f). We refer to F_V and F_L collectively as F. The first combination type, illustrated in Fig. 6.3(c), adopts a sequential configuration, where the Mamba block is applied first, followed by the Transformer block. In the second configuration (Fig. 6.3d), a sequential arrangement is employed, placing the Transformer block before the Mamba block. In the third combination scheme (Fig. 6.3e), a parallel design is employed, with the final representation computed as the element-wise summation of the Transformer and Mamba block outputs. The final configuration (Fig. 6.3f) employs a parallel structure that concatenates features from the Transformer and Mamba blocks to generate the output.

Semantic-promoted Descriptor Enhancer

To incorporate semantic context and enhance the generation of global descriptors, we introduce the Semantic-Promoted Descriptor Enhancer for the integration of cross-Mamba features and semantic estimation.

Semantic Prior Generation: Semantic cues plays a critical role in the perception systems. To extract semantic distributions from the features F_{V_ext} and F_{L_ext}, a SemanticNet estimates the cross-modality semantic priors, F_{V_s}, and F_{L_s}. The C_N presents the number of semantic bins. F_{V_ext} or F_{L_ext} is the input of SemanticNet. We produce the semantic distributions through a lightweight structure composed of residual blocks, including convolutional layers, batch normalization, and activation functions.

Descriptor Enhancement: We propose a semantic-promoted descriptor enhancer to further enrich and generate unified global descriptors. The structure is shown in Fig. 6.4. We denote F_V and F_L collectively as F for simplicity. From Eq. 6.3, the output representation is generated by projecting the hidden state through the system matrix C. Motivated by this, we introduce a cross-matrix alignment mechanism based on the system matrices derived from the semantic distributions F_s and the descriptor features F_d. The semantic bins is further refined into $\ddot{F_s}$ by incorporating both the Mamba block and cross-modality descriptor interaction. The formulation is presented as follows:

$$y_s^t = \mathrm{C}_d h_s^{t*1} + \mathrm{D}_s \chi_s^t \tag{6.4}$$

where t denotes the discrete index, χ_s^t denotes the input sequence and C_d is the system matrix used to decode the hidden state h_s and reconstruct the output representation y_s^t. The place descriptor is further refined as $\ddot{F_d}$ by leveraging the Mamba

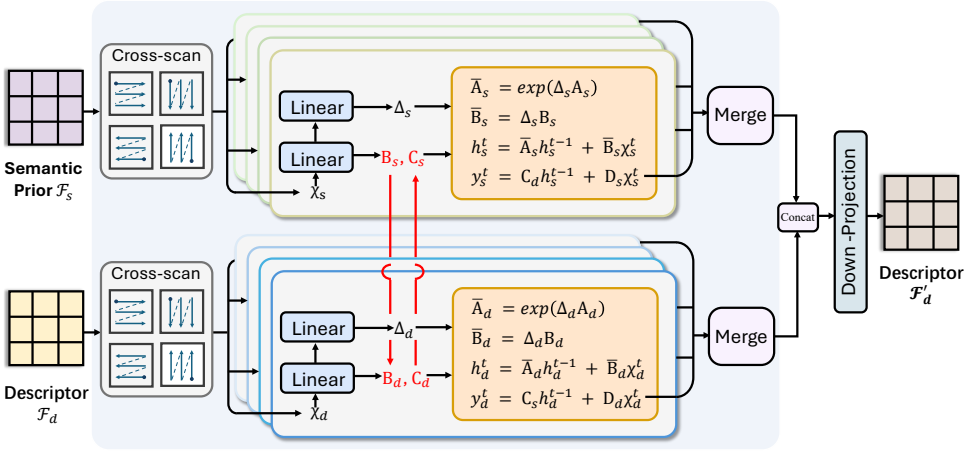

Figure 6.4: An illustration of the proposed Semantic-Promoted Descriptor Enhancer.

block in conjunction with the cross-modality semantic information.

$$y_d^t = \mathrm{C}_s h_d^{t*1} + \mathrm{D}_d \chi_d^t \tag{6.5}$$

where t denotes the discrete index, χ_d^t corresponds to the input sequence and C_s is the system matrix to decode the h_d and recover the output y_d^t. $\ddot{F_s}$ and $\ddot{F_d}$ are generated. As illustrated in Fig. 6.4, a concatenation operation is applied to $\ddot{F_s}$ and $\ddot{F_d}$ and followed by a down-projection neural network to further enrich the learned representations. This process is given by:

$$\ddot{F_{des}} = \mathrm{Project}(\mathrm{Concat}(\ddot{F_s}, \ddot{F_d})) \tag{6.6}$$

Through the semantic-promoted descriptor enhancer module, the final place descriptor $\ddot{F_{des}}$ is generated for the cross-modality retrieval, where N refers to batch size, and D represents embedding dimension.

6.2.5 Cross-Modality Loss

The proposed loss function promotes the alignment of paired visual and point cloud features in the latent space by increasing cosine similarity for N matched pairs and decreasing it for the N^2-N unmatched pairs within a batch. We adopt a contrastive loss, denoted as $\mathcal{L}_{cl}$, to measure the similarity between the generated place descriptors. The loss function is formulated as follows:

$$\mathcal{L}_{pr\_sim} = \frac{1}{N}\sum_{i=1}^{N} * \log\left(\frac{\exp\left(\mathrm{sim}\left(F_V^i, F_L^i\right)/t\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}\left(F_V^i, F_L^j\right)/t\right)}\right), \tag{6.7}$$

where t denotes the temperature hyperparameter, and N denotes the batch size.

To make the learned place descriptor more similar, we impose a feature constraint on the prototypes extracted from the feature extractor for different modalities.

$$\mathcal{L}_{e\_sim} = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F_V^{ext}(i,j) - F_L^{ext}(i,j)\right)^2, \tag{6.8}$$

where F_V^{ext} and F_L^{ext} have the identical shape.

To further enforce representation consistency, we introduce an additional feature representation constraint on the outputs learned by Transformer-Mamba Mixer. This constraint promotes alignment in the feature space and enhances the robustness of the learned place descriptors.

$$\mathcal{L}_{m\_sim} = \frac{1}{n}\sum_{i=1}^{n}\left(F_i^{V\_mix} - F_i^{L\_mix}\right)^2, \tag{6.9}$$

where $F_i^{V\_mix}$ and $F_i^{L\_mix}$ have the same feature shape. So, the final loss consists

of the above-mentioned losses:

$$\mathcal{L}_{total} = \mathcal{L}_{pr\_sim} + \mathcal{L}_{e\_sim} + \mathcal{L}_{m\_sim}, \tag{6.10}$$

Algorithm 1 Inference Phase of Our Proposed Cross-PRNet

Input: Inference sequence S, Threshold distance τ
Output: $Metric$

filenames ← get_filenames(S)
LiDAR_descriptors ← Cross-PRNet(filenames)
visual_descriptors ← Cross-PRNet(filenames)
$match$ = 0
for filename f in filenames **do**
    query q ← visual_descriptors(f)
    $pred$ ← retrieval($LiDAR\_descriptors$, q)
    dis ← distance_compute(q, $pred$)
    if dis $< \tau$ **then**
        $match \leftarrow match + 1$
Compute Metrics
Metric = compute_metric()
return $Metric$

6.3 Training Details

The main focus of our work is an approach to conduct place recognition across the visual and LiDAR modalities through contrastive learning. We begin by describing the datasets and the experimental implementation details. Then, we present the comparisons with other methods to show the capabilities of our method. Moreover, we evaluate the robustness and generalization ability of our network. Finally, we design the ablation study to demonstrate the effectiveness of each proposed component.

6.3.1 Datasets

To demonstrate the superiority and performance of our solution, we adopt two public driving datasets for evaluation.

KITTI: The KITTI dataset [28] offers synchronized image and LiDAR data, collected using a Velodyne HDL-64E and multiple onboard cameras. It also provides accurate ground-truth pose annotations, facilitating rigorous performance evaluation. The sequences were recorded across various urban scenes with differing traffic densities and weather conditions. Sequences 00, 02, 05, and 08 are selected for evaluation.

KITTI-360: KITTI-360 [56] is a large-scale outdoor dataset acquired in urban environments distinct from those in the original KITTI dataset. It comprises over $300,000$ images and $80,000$ LiDAR scans, covering a total driving distance of 73.7 kilometers. The dataset includes 64-beam LiDAR point clouds, high-resolution visual images, and comprehensive calibration data. Importantly, it provides distinct training and testing splits with no overlap with the original KITTI dataset, ensuring fair evaluation in cross-dataset scenarios.

6.3.2 Implementation Details

All components of the proposed framework and the associated experiments are implemented using the PyTorch library [74]. During training, a batch of size N yields N^2 possible visual imagepoint cloud pairs, consisting of N positive pairs and $(N^2\text{-}N)$ negative pairs. The final dimensionality of the generated global place descriptor is set to 256. We adopt the AdamW optimizer with default parameters to optimize the model. The initial learning rate is set to $1e\text{-}5$, and the batch size during

training is fixed at 32. The proposed Cross-PRNet is trained under a contrastive learning framework for cross-modality place recognition. The inference procedure of our model is detailed in Algorithm 1, and all performance metrics are evaluated based on this inference paradigm.

6.3.3 Evaluation Metrics

We use $Recall@N$ as the evaluation metric to assess the performance.

$$\text{Recall}@N = \frac{1}{M}\sum_{m=1}^{M}\Lambda\left(\text{Ç}d_{mn} < \tau, n = 1, \ldots, N\right), \tag{6.11}$$

where τ denotes the recognition threshold. This metric quantifies the ability of the system to correctly retrieve relevant matches within the top-k candidates. A higher metric value indicates better recognition performance.

6.4 Comparative Study

6.4.1 Baseline Approaches

To evaluate the performance of our method, we implement several baseline approaches for comparative analysis:

- NetVLAD: This baseline employs NetVLAD [96], a well-established method originally designed for visual place recognition.
- MixVPR: The baseline leverages MixVPR [1], a visual place recognition framework. Depending on the modality, either RGB images or point cloud-derived range images are used as input.

- AnyLoc: The AnyLoc [40] is a visual place recognition framework built on a pre-trained backbone combined with VLAD aggregation, requiring no fine-tuning. We evaluate this baseline using both visual images and LiDAR-generated range images as input.

Table 6.1: Comparative performance of the proposed Cross-PRNet on the KITTI dataset.

Method	S-00		S-02		S-07		S-08	
	Recall@1	Recall@5	Recall@1	Recall@5	Recall@1	Recall@5	Recall@1	Recall@5
NetVLAD [96]	9.65	21.49	3.48	6.03	34.60	53.59	9.31	20.07
MixVPR [1]	51.86	77.19	18.13	39.82	82.20	95.00	47.70	74.01
AnyLoc [40]	24.73	56.51	11.03	30.12	37.78	72.75	23.31	56.79
PlainEBD [10]	19.93	31.71	16.50	27.48	39.95	59.30	20.49	35.96
VXP [54]	24.22	38.16	17.72	30.83	43.69	61.31	24.01	37.68
I2P-Rec [119]	44.84	59.55	28.00	42.33	63.03	73.39	45.91	62.44
AECMLoc [118]	43.87	62.94	14.80	27.68	78.20	87.74	32.35	51.56
LIP-Loc [87]	72.63	91.43	36.47	57.37	91.37	96.91	74.33	91.08
UniLoc [101]	66.48	88.24	43.00	65.37	93.46	98.09	73.13	89.81
Ours	**90.40**	**98.15**	**70.29**	**87.34**	**98.91**	**100.0**	**90.86**	**98.08**

We conduct evaluations on four sequences of the KITTI dataset: 00, 02, 07, and 08. Table 6.1 summarizes the comparative experimental results against the baseline methods. The results demonstrate that our approach consistently outperforms all baselines. We attribute this improvement to the fact that, while baseline methods are primarily designed to learn from visual data representations, they exhibit limitations in effectively adapting to and extracting discriminative features from depth images generated by LiDAR projections.

6.4.2 Comparison to State-of-the-art Methods

A comprehensive comparison with several state-of-the-art cross-modality place recognition methods is conducted, including PlainEBD [10], VXP [54], AECM-Loc [118], LIP-Loc [87], and UniLoc [101]. For evaluation, we adopt localiza-

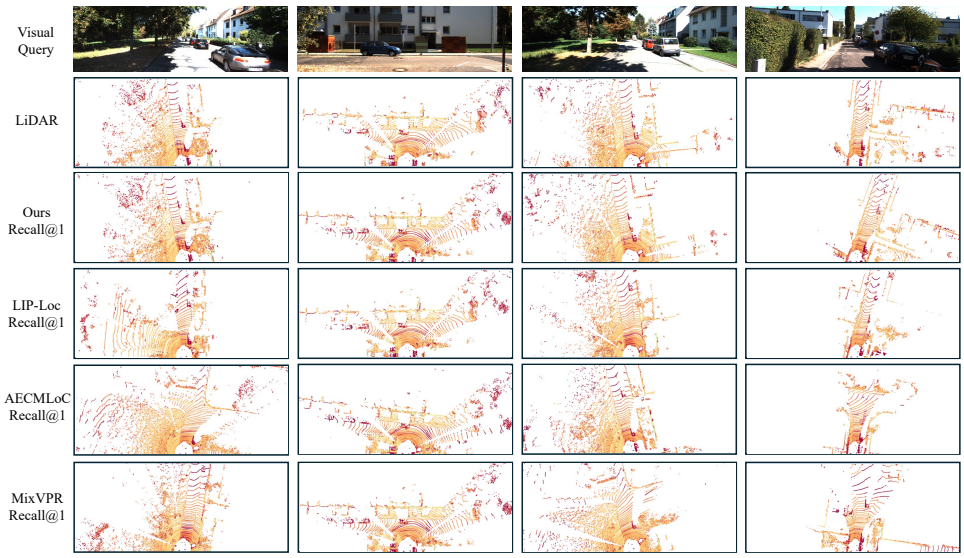

Figure 6.5: Performance comparison between the proposed Cross-PRNet and baseline approaches.

tion thresholds of 20, 10, and 5 meters. The experiments are conducted on four representative KITTI sequences (00, 02, 07, and 08). As shown in Table 6.1, our Cross-PRNet consistently outperforms the baseline methods across all distance thresholds, demonstrating its superior accuracy and robustness in large-scale cross-modality localization. We attribute this performance gain to the synergistic integration of Transformer and Mamba structures, which enables our model to effectively capture contextual features, thereby enhancing retrieval precision. In addition, qualitative comparisons presented in Fig. 6.5 further highlight the effectiveness of Cross-PRNet. It can be seen that our method consistently retrieves the correct locations.

Furthermore, we utilize t-SNE [97] to project the learned place descriptor embeddings of our method and comparative baselines into a two-dimensional space, as shown in Fig. 6.6. The visualizations highlight the distributions of descriptors retrieved at $Recall@1$ through $Recall@5$, in comparison with the ground truth lo-

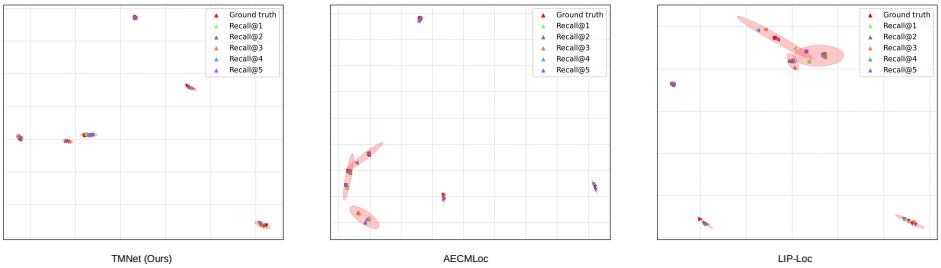

Figure 6.6: Visualization of place descriptor distributions for our Cross-PRNet model compared to other approaches. Six randomly selected retrieval-prediction samples are shown using distinct symbolic shapes.

Table 6.2: Robustness demonstration of our Cross-PRNet on the KITTI-360 dataset. The symbol £ indicates results obtained using our method without any additional training.

Method	$\tau = 20$		$\tau = 5$		Runtime (ms)
	Recall@1	Recall@5	Recall@1	Recall@5	
NetVLAD [96]	8.63	18.15	3.87	8.57	7.01
MixVPR [1]	50.73	75.17	45.44	69.60	9.09
AnyLoc [40]	44.12	68.70	37.28	60.39	67.75
AECMLoc [118]	46.23	66.04	-	-	22.02
LIP-Loc [87]	68.65	86.85	64.00	83.04	19.83
Ours£	63.33	83.96	57.35	78.95	-
Ours	**89.58**	**96.79**	**87.57**	**95.88**	15.09

cations. The solid red symbols denote the ground truth locations. As observed, the descriptor distributions generated by our method exhibit more compact clustering and are more closely aligned with the corresponding LiDAR feature distributions compared to those of the baseline methods. The pink area denotes a retrieval case, alongside the corresponding predicted descriptor distributions, which are visualized using computed covariance ellipses. A relatively compact covariance ellipse indicates low variance in the predicted descriptors, suggesting that the model produces stable and consistent representations. This compactness serves as evidence

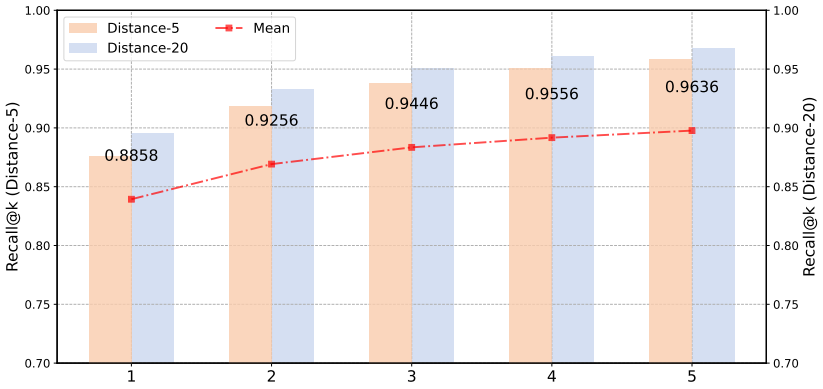

Figure 6.7: Robustness evaluation of the Cross-PRNet model on the KITTI-360 dataset.

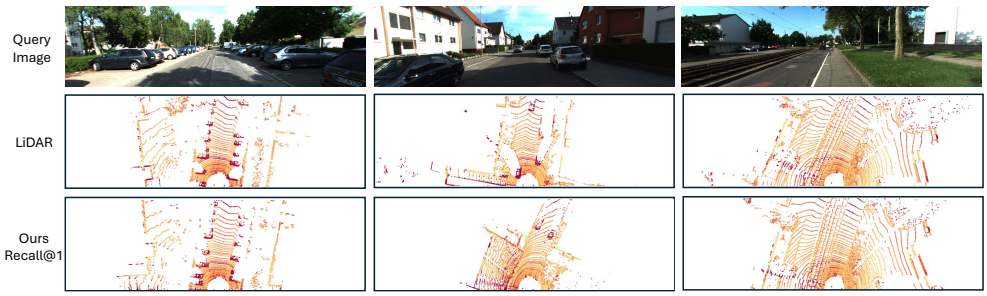

Figure 6.8: Place recognition robustness performance of our Cross-PRNet on the KITTI-360 dataset.

of the reliability and robustness of our approach. Furthermore, the consistent clustering around the ground truth highlights the superiority of the proposed method in generating discriminative and modality-invariant place descriptors.

6.5 Robustness Study

We evaluate the robustness and generalization capability of the proposed Cross-PRNet model. First, to assess robustness, we conduct cross-dataset evaluation on the KITTI-360 dataset. The corresponding performance results are presented in Table 6.2. As shown, our Cross-PRNet consistently outperforms other state-of-

Table 6.3: Ablation study on the analysis of different feature extractors on the KITTI dataset. SwinT denotes the Swin Transformer.

Variant	Extractor	Params	$\tau = 20$			$\tau = 10$			$\tau = 5$		
			Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Variant 1	ResNet-18	11.1M	75.36	91.72	96.12	72.00	89.36	94.13	67.87	86.61	91.80
Variant 2	ResNet-34	21.2M	78.02	92.90	96.54	75.58	91.18	95.58	70.97	89.17	94.03
Variant 3	EfficientNet-b0	4.3M	74.80	90.57	95.26	71.55	88.26	93.32	66.74	84.45	90.74
Variant 4	EfficientNet-b1	6.1M	78.51	93.69	97.25	75.66	91.82	95.58	71.46	89.44	93.88
Variant 5	RegNet-y-800mf	5.7M	71.80	90.37	95.31	69.22	88.18	93.37	63.57	84.97	91.16
Variant 6	RegNet-x-1.6G	52.2M	79.98	93.88	97.27	77.38	92.19	96.07	72.83	90.27	94.72
Variant 7	SwinT-small	86.7M	90.15	98.11	99.21	88.68	97.52	98.99	85.46	96.56	98.48
Variant 8	ConvNext-Tiny	27.8M	87.87	97.00	98.58	85.58	96.24	98.06	82.95	95.23	97.42
Variant 9 (Ours)	ConvNext-Small	49.4M	**92.29**	**98.75**	**99.66**	**90.86**	**98.08**	**99.39**	**89.00**	**97.32**	**98.70**

the-art methods across various evaluation metrics, demonstrating its generalization ability and robustness in cross-modal place recognition scenarios. In addition, we investigate the inference efficiency by measuring the time required to generate place descriptors for each comparative method. To ensure a fair comparison, all models are tested under identical hardware settings using the same GPU device. Our Cross-PRNet requires approximately 15 ms per query to compute the place descriptor. The Transformer-Mamba Mixer and Semantic-Promoted Descriptor Enhancer modules contribute approximately 4.48 ms and 3.55 ms to the total inference time, respectively. This demonstrates that our Cross-PRNet framework is efficient enough for online cross-modal localization tasks. Furthermore, detailed quantitative robustness evaluations are illustrated in Fig. 6.7, where it can be observed that our model maintains stable and robust performance across different cities and diverse scenarios. Additionally, we provide qualitative visualization results of the recognition performance of our model on the KITTI-360 dataset in Fig. 6.8, further validating the effectiveness of our approach.

Furthermore, we evaluate the generalization capability of our model by conducting extensive tests on unseen sequences from various scenes. Specifically, we utilize the well-trained model weights on the KITTI dataset and directly ap-

Table 6.4: Ablation study of our fusion design on the KITTI dataset.

Variant	Strategy	Recall@1	Recall@5	Recall@10
Variant 1	✗	91.05	97.42	98.80
Variant 2	Summation	91.97	98.60	99.58
Variant 3	✗ Semantic	91.21	97.64	99.12
Variant 4 (Ours)	Enhancer	**92.29**	**98.75**	**99.66**

ply them to distinct test environments, without any further fine-tuning. These test scenarios include diverse traffic conditions, such as rural areas, which differ significantly from the training data. The experimental results under different recognition thresholds τ are presented in Fig. 6.9. As shown, our model maintains robust performance across unseen environments, highlighting its strong generalization ability.

6.6 Ablation Studies

We design several ablation studies to demonstrate the performance and effectiveness of every proposed component in our Cross-PRNet.

6.6.1 Analysis of Feature Extractor

To determine the suitable feature extractor, we perform an ablation study within the framework of our network. We constructed several network variants by employing different backbone architectures, including the ResNet [31] family, EfficientNet [94] family, ConvNeXt [62] family, RegNet [78] family, and Swin Transformer [60] family, as feature extractors. The results, shown in Table 6.3, demonstrate that various backbone types yield relatively competitive performance, highlighting

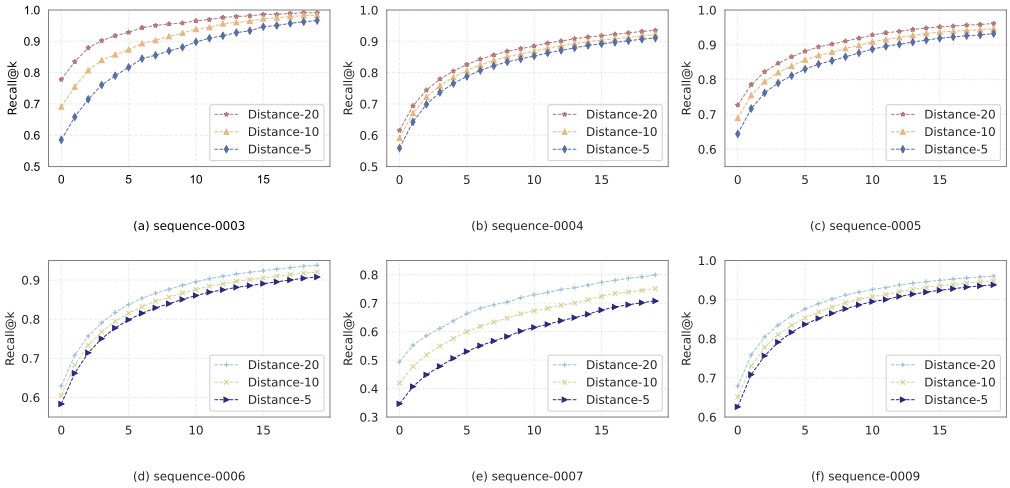

Figure 6.9: Generalization performance of the proposed Cross-PRNet model across different sequences under varying thresholds.

the generality and adaptability of our method. Notably, variant 9 achieves the highest performance in metric learning evaluation. Overall, the ConvNeXt-Small backbone provides a favorable balance between model complexity and recognition accuracy. Therefore, unless otherwise specified, ConvNeXt-Small is adopted as the default feature extractor.

6.6.2 Effect on Transformer-Mamba Mixer

Table 6.5: Ablation study of the Transformer-Mamba Mixer module on the KITTI dataset.

Variant	Module	Recall@1	Recall@5	Recall@10
Variant 1	Sequential-1	89.20	97.06	98.80
Variant 2	Sequential-2	90.13	98.75	99.63
Variant 3	Parallel-1	90.86	98.23	99.29
Variant 4	w/o Transformer	84.83	95.36	97.91
Variant 5	w/o Mamba	86.81	96.46	98.45
Variant 6 (Ours)	Parallel-2	**92.29**	**98.75**	**99.66**

We conduct an ablation study to evaluate the effectiveness of our Transformer-Mamba Mixer component. Several model variants are constructed, as illustrated in the Fig.6.3.

The corresponding results are summarized in Table 6.5. Among these, the combination type *f* achieves the best overall performance, demonstrating the benefits of integrating both Transformer and Mamba blocks. In contrast, using only the VSS block (variant 4) or only the Transformer block (variant 5) results in noticeably lower place recognition accuracy. We attribute this to the limited global-context modeling capacity of the VSS block when used in isolation, which restricts its ability to capture long-range dependencies necessary for effective cross-modal representation alignment. The paradigm employing both parallel structure and concatenation operation proves to be the optimal choice in terms of place recognition performance, as supported by both recognition accuracy and statistical significance tests. This design enables the model to simultaneously leverage the complementary strengths of the Mamba and Transformer blocks. The concatenation operation plays a crucial role by facilitating effective integration of their respective feature representations, thereby enhancing the capacity to learn robust and discriminative descriptors across modalities.

6.6.3 Analysis of Semantic-Promoted Descriptor Enhancer

We conduct a comparative analysis between our full model and several variants to evaluate the effectiveness of the semantic-promoted descriptor enhancer strategy. The first variant excludes the semantic-promoted enhancer module entirely. The second variant replaces the concatenation operation with an element-wise sum-

mation strategy. Variant 3 removes the semantic distribution information from the enhancement process.

The ablation results, as presented in Table 6.4, demonstrate that our full model consistently outperforms all variants in terms of recognition accuracy. These findings validate the importance of the semantic-promoted enhancer in improving the discriminative capability of the learned descriptors. Moreover, the incorporation of semantic distribution estimation proves to be beneficial, further enhancing the performance on cross-modal place recognition.

Table 6.6: Ablation study of various loss functions on the KITTI dataset.

Variant	$\mathcal{L}_{e\_sim}$	$\mathcal{L}_{m\_sim}$	$\mathcal{L}_{pr\_sim}$	Recall@1	Recall@5
Variant 1			✓	90.81	98.50
Variant 2		✓	✓	91.71	97.75
Variant 3 (Ours)	✓	✓	✓	**92.29**	**98.75**

6.6.4 Analysis of Loss Function

Here, we conduct the ablation study on the loss function to demonstrate the effectiveness of our loss function design. We construct several variants by systematically altering the integration of feature learning constraints at different positions within the network.

The corresponding results are summarized in Table 6.6. As observed, our model consistently achieves superior recognition performance compared to all other variants. This suggests that the three proposed feature-level constraints play a vital role in reinforcing the discriminability and alignment of cross-modal representations, thereby enhancing overall recognition accuracy.

6.7 Conclusion

We propose Cross-PRNet, a unified and efficient framework for cross-modal image-to-point-cloud place recognition. Our method employs an end-to-end Siamese network to generate robust place descriptors across modalities without relying on point cloud cropping or hard negative mining. To address the heterogeneity between modalities, we project LiDAR point clouds into range image representations, facilitating more effective feature extraction. We further introduce a Transformer-Mamba Mixer paradigm that implicitly captures contextual information to produce coarse place descriptors. Moreover, a semantic-promoted descriptor enhancer module is proposed to refine these descriptors by incorporating semantic distribution estimation, resulting in enhanced global representations. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods in recognition accuracy, efficiency, and robustness.

Chapter 7

Conclusion and Future Work

This thesis presents an investigation into end-to-end semantic understanding for autonomous driving by addressing four critical research tasks: BEV dynamic scene perception, multi-modal fusion, interpretable behavior prediction, and cross-modal place recognition.

First, to perceive dynamic environments for vision-based autonomous vehicles, we proposed a novel vision-based framework for end-to-end moving-obstacle segmentation in the BEV domain. The framework explicitly incorporates camera intrinsic/extrinsic parameters and temporal semantic priors by projecting multi-view image features into a unified BEV space. An auxiliary supervision mechanism based on movable-obstacle segmentation is further introduced to enhance performance. Extensive experiments on real-world benchmarks validate the effectiveness of our method in generating dense and accurate BEV representations without relying on LiDAR sensors.

Second, we introduced DPMoSeg, a depth-aware segmentation network that integrates 3D LiDAR point clouds into the BEV representation via a sparse-dense

attention mechanism. This design effectively compensates for visual limitations under challenging conditions by leveraging geometric information, and further incorporates drivable area segmentation as an auxiliary task to facilitate spatial reasoning in the BEV plane.

Third, to tackle the challenge of explainable behavior prediction, we proposed a framework that integrates a self-supervised, class-agnostic object segmentor with semantic reasoning modules. This design leverages foundation models and adaptive fusion strategies to produce accurate behavior predictions alongside human-understandable explanations, thereby improving the transparency and trustworthiness of autonomous decision-making systems.

Lastly, we presented Cross-PRNet, a unified and efficient framework for cross-modal place recognition between RGB images and 3D point clouds. By projecting LiDAR data into range image space and employing a Transformer-Mamba Mixer architecture along with a semantic-enhanced descriptor refinement module, our approach achieves accurate and robust cross-modality localization, even in the absence of dense LiDAR data.

Although our study has achieved satisfactory results across multiple dimensions of semantic understanding for autonomous driving, several limitations remain that warrant further exploration in future work. For the proposed Semantic-MoSeg framework, the current design relies solely on visual inputs from multi-view cameras, which limits robustness in adverse scenarios such as low illumination. To address this, we would like to explore the integration of additional sensor modalities, such as LiDAR point clouds or thermal information, to provide complementary geometric and spectral cues. In addition, the DPMoSeg model, which currently operates over short temporal intervals, will be extended to longer-

term temporal horizons, enabling a more comprehensive understanding of scene dynamics and broader applicability. We aim to investigate the utility of BEV moving-obstacle segmentation as a foundational layer for other downstream tasks, such as panoptic segmentation, trajectory forecasting. For the interpretable vehicle decision-making framework, we would like to incorporate multi-modal inputs, such as LiDAR point clouds, to further improve prediction robustness and contextual awareness. We also plan to explore model compression techniques, such as pruning and quantization, to facilitate real-time deployment on resource-constrained platforms. Moreover, we plan to conduct user studies to quantitatively assess how natural language explanations influence user trust and confidence in autonomous systems. Finally, for the Cross-PRNet framework, future enhancements will focus on integrating richer modalities, including depth maps, scene semantics, and language descriptions, to further improve cross-modal recognition performance, particularly in visually degraded or GPS-denied environments. These extensions are expected to enhance the generalization and robustness of place recognition under real-world conditions.

Bibliography

[1] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. “Mixvpr: Feature mixing for visual place recognition”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023, pp. 2998–3007.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, et al. “NetVLAD: CNN architecture for weakly supervised place recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5297–5307.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[4] Stefan Andreas Baur, David Josef Emmerichs, Frank Moosmann, et al. “SLIM: Self-supervised LiDAR scene flow and motion segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13126–13136.

[5] Hédi Ben-Younes, Éloi Zablocki, Patrick Pérez, et al. “Driving behavior explanation with multi-level fusion”. In: *Pattern Recognition* 123 (2022), p. 108421.

[6] Lukas Bernreiter, Lionel Ott, Juan Nieto, et al. “Spherical multi-modal place recognition for heterogeneous sensor systems”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 1743–1750.

[7] Gabriele Berton, Carlo Masone, and Barbara Caputo. “Rethinking visual geo-localization for large-scale applications”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4878–4888.

[8] Holger Caesar, Varun Bankiti, Alex H Lang, et al. “nuscenes: A multimodal dataset for autonomous driving”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11621–11631.

[9] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, et al. “Structured Bird’s-Eye-View Traffic Scene Understanding From Onboard Images”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15661–15670.

[10] Daniele Cattaneo, Matteo Vaghi, Simone Fontana, et al. “Global visual localization in LiDAR-maps through shared 2D-3D embedding space”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 4365–4371.

[11] Daniele Cattaneo, Matteo Vaghi, and Abhinav Valada. “Lcdnet: Deep loop closure detection and point cloud registration for lidar slam”. In: *IEEE Transactions on Robotics* 38.4 (2022), pp. 2074–2093.

[12] Dian Chen and Philipp Krähenbühl. “Learning from all vehicles”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 17222–17231.

[13] Jiaxuan Chen, Xiaoxian Chen, Shuang Chen, et al. “Shape-Former: Bridging CNN and Transformer via ShapeConv for multimodal image matching”. In: *Information Fusion* 91 (2023), pp. 445–457.

[14] Li Chen, Penghao Wu, Kashyap Chitta, et al. “End-to-end autonomous driving: Challenges and frontiers”. In: *arXiv preprint arXiv:2306.16927* (2023).

[15] Xieyuanli Chen, Shijie Li, Benedikt Mersch, et al. “Moving object segmentation in 3D LiDAR data: A learning-based approach exploiting sequential data”. In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 6529–6536.

[16] Xieyuanli Chen, Benedikt Mersch, Lucas Nunes, et al. “Automatic Labeling to Generate Training Data for Online LiDAR-Based Moving Object Segmentation”. In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 6107–6114.

[17] Bowen Cheng, Alex Schwing, and Alexander Kirillov. “Per-pixel classification is not all you need for semantic segmentation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17864–17875.

[18] Lu Chi, Borui Jiang, and Yadong Mu. “Fast fourier convolution”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4479–4488.

[19] Pranav Singh Chib and Pravendra Singh. “Recent advancements in end-to-end autonomous driving using deep learning: A survey”. In: *IEEE Transactions on Intelligent Vehicles* (2023).

[20] Sungha Choi, Joanne T Kim, and Jaegul Choo. “Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9373–9383.

[21] Laurene Claussmann, Marc Revilloud, Dominique Gruyer, et al. “A review of motion planning for highway autonomous driving”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.5 (2019), pp. 1826–1848.

[22] Isht Dwivedi, Srikanth Malla, Yi-Ting Chen, et al. “Birds eye view segmentation using lifted 2D semantic features”. In: *British Machine Vision Conference (BMVC)*. 2021, pp. 6985–6994.

[23] Mark Everingham, Luc Van Gool, Christopher KI Williams, et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[24] George Fahim, Khalid Amin, and Sameh Zarif. “Single-View 3D reconstruction: A Survey of deep learning methods”. In: *Computers & Graphics* 94 (2021), pp. 164–190.

[25] Jingjing Fan, Lili Fan, Qinghua Ni, et al. “Perception and Planning of Intelligent Vehicles Based on BEV in Extreme Off-road Scenarios”. In: *IEEE Transactions on Intelligent Vehicles* (2024).

[26] Yuchao Feng, Wei Hua, and Yuxiang Sun. “Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding”. In: *IEEE Transactions on Intelligent Transportation Systems* (2023).

[27] Roland Gao. “Rethinking dilated convolution for real-time semantic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 4675–4684.

[28] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361.

[29] Jhony H Giraldo, Sajid Javed, Maryam Sultana, et al. “The emerging field of graph signal processing for moving object segmentation”. In: *International Workshop on Frontiers of Computer Vision*. Springer. 2021, pp. 31–45.

[30] Peng Hang, Chen Lv, Yang Xing, et al. “Human-like decision making for autonomous driving: A noncooperative game theoretic approach”. In: *IEEE Transactions on Intelligent Transportation Systems* 22.4 (2020), pp. 2076–2087.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[32] Carl-Johan Hoel, Katherine Driggs-Campbell, Krister Wolff, et al. “Combining planning and deep reinforcement learning in tactical decision making for autonomous driving”. In: *IEEE transactions on intelligent vehicles* 5.2 (2019), pp. 294–305.

[33] Andrew Howard, Mark Sandler, Grace Chu, et al. “Searching for mobilenetv3”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[34] Anthony Hu, Zak Murez, Nikhil Mohan, et al. “FIERY: Future Instance Prediction in Bird’s-Eye View From Surround Monocular Cameras”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15273–15282.

[35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[36] Le Hui, Hang Yang, Mingmei Cheng, et al. “Pyramid point cloud transformer for large-scale place recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6098–6107.

[37] Hongxiang Jiang, Wenming Meng, Hongmei Zhu, et al. “Multi-Camera Calibration Free BEV Representation for 3D Object Detection”. In: *arXiv preprint arXiv:2210.17252* (2022).

[38] Shibo Jie and Zhi-Hong Deng. “Convolutional bypasses are better vision transformer adapters”. In: *arXiv preprint arXiv:2207.07039* (2022).

[39] Taotao Jing, Haifeng Xia, Renran Tian, et al. “Inaction: Interpretable action decision making for autonomous driving”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 370–387.

[40] Nikhil Keetha, Avneesh Mishra, Jay Karhade, et al. “Anyloc: Towards universal visual place recognition”. In: *IEEE Robotics and Automation Letters* (2023).

[41] R Kesten, M Usman, J Houston, et al. “Lyft level 5 av dataset 2019”. In: *https://level5.lyft.com/dataset* 1 (2019), p. 3.

[42] Gwangbin Kim, Dohyeon Yeo, Taewoo Jo, et al. “What and When to Explain? On-road Evaluation of Explanations in Highly Automated Vehicles”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.3 (2023), pp. 1–26.

[43] Jaeyeul Kim, Jungwan Woo, and Sunghoon Im. “RVMOS: Range-View Moving Object Segmentation Leveraged by Semantic and Motion Features”. In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 8044–8051.

[44] Jinkyu Kim and John Canny. “Interpretable learning for self-driving cars by visualizing causal attention”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2942–2950.

[45] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, et al. “Textual explanations for self-driving vehicles”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 563–578.

[46] Nuri Kim, Jeongho Park, Mineui Hong, et al. “Semantic Environment Atlas for Object-Goal Navigation”. In: *Knowledge-Based Systems* 304 (2024), p. 112446.

[47] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, et al. “Concept bottleneck models”. In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.

[48] Varun Ravi Kumar, Senthil Yogamani, Hazem Rashed, et al. “Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving”. In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2830–2837.

[49] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, et al. “Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review”. In: *arXiv preprint arXiv:2402.10086* (2024).

[50] Alex H Lang, Sourabh Vora, Holger Caesar, et al. “Pointpillars: Fast encoders for object detection from point clouds”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 12697–12705.

[51] Ji Li, Qingxiao Liu, Boyang Wang, et al. “RangePlace: A Hierarchical Range Image Transformer for LiDAR-Based Place Recognition”. In: *IEEE Transactions on Intelligent Vehicles* (2024).

[52] Lin Li, Xin Kong, Xiangrui Zhao, et al. “RINet: Efficient 3D lidar-based place recognition using rotation invariant neural network”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 4321–4328.

[53] Yinhao Li, Zheng Ge, Guanyi Yu, et al. “Bevdepth: Acquisition of reliable depth for multi-view 3d object detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2. 2023, pp. 1477–1485.

[54] Yun-Jin Li, Mariia Gladkova, Yan Xia, et al. “VXP: Voxel-Cross-Pixel Large-scale Image-LiDAR Place Recognition”. In: *arXiv preprint arXiv:2403.14594* (2024).

[55] Zhiqi Li, Wenhai Wang, Hongyang Li, et al. “Bevformer: Learning birds-eye-view representation from multi-camera images via spatiotemporal transformers”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer. 2022, pp. 1–18.

[56] Yiyi Liao, Jun Xie, and Andreas Geiger. “Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022), pp. 3292–3310.

[57] Chien-Chuan Lin and Ming-Shi Wang. “A vision based top-view transformation model for a vehicle parking assistant”. In: *Sensors* 12.4 (2012), pp. 4431–4446.

[58] Wenyu Liu, Wentong Li, Jianke Zhu, et al. “Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[59] Yiling Liu and Hesheng Wang. “MotionRFCN: Motion Segmentation Using Consecutive Dense Depth Maps”. In: *Pacific Rim International Conference on Artificial Intelligence*. Springer. 2019, pp. 510–522.

[60] Ze Liu, Yutong Lin, Yue Cao, et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[61] Zhi Liu, Shaoyu Chen, Xiaojie Guo, et al. “Vision-based uneven bev representation learning with polar rasterization and surface estimation”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 437–446.

[62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[63] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, et al. “Efficient dense modules of asymmetric convolution for real-time semantic segmentation”. In: *Proceedings of the ACM Multimedia Asia*. 2019, pp. 1–6.

[64] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[65] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. “Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks”. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 445–452.

[66] Lei Lu, Yun Xiao, Xiaojun Chang, et al. “Deformable attention-oriented feature pyramid network for semantic segmentation”. In: *Knowledge-Based Systems* 254 (2022), p. 109623.

[67] Lun Luo, Shuhang Zheng, Yixuan Li, et al. “BEVPlace: Learning LiDAR-based place recognition using bird’s eye view images”. In: *Proceedings*

of the IEEE/CVF International Conference on Computer Vision. 2023, pp. 8700–8709.

[68] Junyi Ma, Jun Zhang, Jintao Xu, et al. "OverlapTransformer: An efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition". In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 6958–6965.

[69] AV Shreyas Madhav and Amit Kumar Tyagi. "Explainable Artificial Intelligence (XAI): connecting artificial decision-making and human trust in autonomous vehicles". In: *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*. Springer. 2022, pp. 123–136.

[70] Benedikt Mersch, Xieyuanli Chen, Ignacio Vizzo, et al. "Receding Moving Object Segmentation in 3D LiDAR Data Using Sparse 4D Convolutions". In: *arXiv preprint arXiv:2206.04129* (2022).

[71] Keisuke Mori, Hiroshi Fukui, Takuya Murase, et al. "Visual explanation by attention branch network for end-to-end learning-based self-driving". In: *2019 IEEE intelligent vehicles symposium (IV)*. IEEE. 2019, pp. 1577–1582.

[72] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, et al. "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review". In: *IEEE Transactions on Intelligent Transportation Systems* 23.1 (2020), pp. 33–47.

[73] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, et al. “Cross-view semantic segmentation for sensing surroundings”. In: *IEEE Robotics and Automation Letters* 5.3 (2020), pp. 4867–4873.

[74] Adam Paszke, Sam Gross, Francisco Massa, et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[75] Prashant W Patil, Kuldeep M Biradar, Akshay Dudhane, et al. “An end-to-end edge aggregation network for moving object segmentation”. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8149–8158.

[76] Jonah Philion and Sanja Fidler. “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer. 2020, pp. 194–210.

[77] Charles R Qi, Hao Su, Kaichun Mo, et al. “Pointnet: Deep learning on point sets for 3d classification and segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[78] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, et al. “Designing network design spaces”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10428–10436.

[79] Prabhu Ramanathan and Kartik. “Autonomous Driving Cars: Decision-Making”. In: *Internet of Vehicles and its Applications in Autonomous Driving* (2021), pp. 31–39.

[80] Hazem Rashed, Mariam Essam, Maha Mohamed, et al. “Bev-modnet: Monocular camera based bird’s eye view moving object detection for autonomous driving”. In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE. 2021, pp. 1503–1508.

[81] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, et al. “Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

[82] Thomas Roddick and Roberto Cipolla. “Predicting semantic map representations from images using pyramid occupancy networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 11138–11147.

[83] Giulio Rossolini, Federico Nesti, Gianluca DAmico, et al. “On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[84] Abbas Sadat, Sergio Casas, Mengye Ren, et al. “Perceive, predict, and plan: Safe motion planning through interpretable semantic representations”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer. 2020, pp. 414–430.

[85] Yoshihide Sawada and Keigo Nakamura. “C-SENN: Contrastive self-explaining neural network”. In: *arXiv preprint arXiv:2206.09575* (2022).

[86] Sawada, Yoshihide and Nakamura, Keigo. “Concept bottleneck model with additional unsupervised concepts”. In: *IEEE Access* 10 (2022), pp. 41758–41765.

[87] Sai Shubodh, Mohammad Omama, Husain Zaidi, et al. “Lip-loc: Lidar image pretraining for cross-modal localization”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 948–957.

[88] Mennatullah Siam, Sara Eikerdawy, Mostafa Gamal, et al. “Real-time segmentation with appearance, motion and geometry”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 5793–5800.

[89] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, et al. “Modnet: Motion and appearance based moving object detection network for autonomous driving”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 2859–2864.

[90] Farhana Sultana, Abu Sufian, and Paramartha Dutta. “Evolution of image segmentation using deep convolutional neural network: A survey”. In: *Knowledge-Based Systems* 201 (2020), p. 106062.

[91] Yuxiang Sun, Ming Liu, and Max Q-H Meng. “Active perception for foreground segmentation: An RGB-D data-based background modeling method”. In: *IEEE Transactions on Automation Science and Engineering* 16.4 (2019), pp. 1596–1609.

[92] Yuxiang Sun, Weixun Zuo, Huaiyang Huang, et al. “PointMoSeg: Sparse tensor-based end-to-end moving-obstacle segmentation in 3-D lidar point

clouds for autonomous driving". In: *IEEE Robotics and Automation Letters* 6.2 (2020), pp. 510–517.

[93] Ardi Tampuu, Tambet Matiisen, Maksym Semikin, et al. "A survey of end-to-end driving: Architectures and training methods". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.4 (2020), pp. 1364–1384.

[94] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[95] Mingxing Tan, Ruoming Pang, and Quoc V Le. "Efficientdet: Scalable and efficient object detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.

[96] Mikaela Angelina Uy and Gim Hee Lee. "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4470–4479.

[97] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[98] Johan Vertens, Abhinav Valada, and Wolfram Burgard. "Smsnet: Semantic motion segmentation using deep convolutional neural networks". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 582–589.

[99] Dequan Wang, Coline Devin, Qi-Zhi Cai, et al. “Deep object-centric policies for autonomous driving”. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 8853–8859.

[100] Sijie Wang, Qiyu Kang, Rui She, et al. “PRFusion: Toward Effective and Robust Multi-Modal Place Recognition With Image and Point Cloud Fusion”. In: *IEEE Transactions on Intelligent Transportation Systems* (2024).

[101] Yan Xia, Zhendong Li, Yun-Jin Li, et al. “UniLoc: Towards Universal Place Recognition Using Any Single Modality”. In: *arXiv preprint arXiv:2412.12079* (2024).

[102] Xiaoyang Xiao, Yuqian Zhao, Fan Zhang, et al. “BASeg: Boundary aware semantic segmentation for autonomous driving”. In: *Neural Networks* 157 (2023), pp. 460–470.

[103] Huaiyuan Xu, Huaping Liu, Shoudong Huang, et al. “C2l-pr: Cross-modal camera-to-lidar place recognition via modality alignment and orientation voting”. In: *IEEE Transactions on Intelligent Vehicles* (2024).

[104] Yiran Xu, Xiaoyin Yang, Lihang Gong, et al. “Explainable object-induced action decision for autonomous vehicles”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9523–9532.

[105] Haozhi Yang, Jing Yuan, Yuanxi Gao, et al. “UPLP-SLAM: Unified point-line-plane feature fusion for RGB-D visual SLAM”. In: *Information Fusion* 96 (2023), pp. 51–65.

[106] Hesheng Yin, Shaomiao Li, Yu Tao, et al. “Dynam-SLAM: An Accurate, Robust Stereo Visual-Inertial SLAM Method in Dynamic Environments”. In: *IEEE Transactions on Robotics* 39.1 (2023), pp. 289–308.

[107] Changqian Yu, Jingbo Wang, Changxin Gao, et al. “Context prior for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 12416–12425.

[108] Changqian Yu, Jingbo Wang, Chao Peng, et al. “Bisenet: Bilateral segmentation network for real-time semantic segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 325–341.

[109] Fisher Yu, Haofeng Chen, Xin Wang, et al. “Bdd100k: A diverse driving dataset for heterogeneous multitask learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2636–2645.

[110] Jun Yu, Chaoyang Zhu, Jian Zhang, et al. “Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition”. In: *IEEE transactions on neural networks and learning systems* 31.2 (2019), pp. 661–674.

[111] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, et al. “A survey of autonomous driving: Common practices and emerging technologies”. In: *IEEE access* 8 (2020), pp. 58443–58469.

[112] Wenyuan Zeng, Wenjie Luo, Simon Suo, et al. “End-to-end interpretable neural motion planner”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8660–8669.

[113] Dingwen Zhang, Junwei Han, Gong Cheng, et al. “Weakly supervised object localization and detection: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 5866–5885.

[114] Hanwei Zhang, Ying Zhu, Dan Wang, et al. “A survey on visual mamba”. In: *Applied Sciences* 14.13 (2024), p. 5683.

[115] Xuetao Zhang, Zhenxue Chen, QM Jonathan Wu, et al. “Fast semantic segmentation for scene perception”. In: *IEEE Transactions on Industrial Informatics* 15.2 (2018), pp. 1183–1192.

[116] Zhengming Zhang, Renran Tian, Rini Sherony, et al. “Attention-based interrelation modeling for explainable automated driving”. In: *IEEE Transactions on Intelligent Vehicles* 8.2 (2022), pp. 1564–1573.

[117] Xu Zhao, Wenchao Ding, Yongqi An, et al. “Fast segment anything”. In: *arXiv preprint arXiv:2306.12156* (2023).

[118] Zhipeng Zhao, Huai Yu, Chenwei Lyu, et al. “Attention-enhanced cross-modal localization between spherical images and point clouds”. In: *IEEE Sensors Journal* (2023).

[119] Shuhang Zheng, Yixuan Li, Zhu Yu, et al. “I2P-Rec: Recognizing Images on Large-Scale Point Cloud Maps Through Bird’s Eye View Projections”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 1395–1400.

[120] Brady Zhou and Philipp Krähenbühl. “Cross-view Transformers for real-time Map-view Semantic Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13760–13769.

[121] Wujie Zhou, Han Zhang, Weiqing Yan, et al. “MMSMCNet: Modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[122] Yao Zhou, Guowei Wan, Shenhua Hou, et al. “Da4ad: End-to-end deep attention-based visual localization for autonomous driving”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer. 2020, pp. 271–289.

[123] Shuangquan Zuo, Yun Xiao, Xiaojun Chang, et al. “Vision transformers for dense prediction: A survey”. In: *Knowledge-Based Systems* 253 (2022), p. 109552.