MIM: High-Definition Maps Incorporated Multi-View 3D Object Detection

Jinsheng Xiao, *Senior Member, IEEE*, Shurui Wang, Jian Zhou, *Member, IEEE*, Ziyue Tian, Hongping Zhang, and Yuan-Fang Wang

***Abstract*— 3D object detection has aroused increasing interest as a crucial component of autonomous driving systems. While recent works have explored various multi-modal fusion methods to enhance accuracy and robustness, fusing multi-view images and high-definition (HD) maps remains uncharted. Inspired by our previous work, we endeavor to introduce HD maps to camera-based detection, prompting the design of a new framework. To address this, we first analyze the function of HD maps in object detection to understand their benefits and the rationale for their fusion. From this analysis, we identify key disparities in view, semantics, and scale, leading to the development of MIM, a framework for HD Maps Incorporated Multi-view 3D object detection. HD maps are enriched in semantics by sampling unlabeled areas and encoding them into map features. Simultaneously, multi-view images are transformed into features in bird's-eye view (BEV) using the adopted baseline. These features are then fused using attention mechanisms to align scales. Experiments conducted on the nuScenes dataset demonstrate that MIM outperforms camera-based methods. Moreover, an in-depth analysis investigates how HD maps impact object detection regarding each semantic layer. The results underscore the operational intricacies of HD maps in perception, setting the stage for future research. Code is available at https://github.com/WHU-xjs/MIM-3D-Det.**

***Index Terms*— 3D object detection, autonomous driving, multi-modal fusion, multi-view images, high-definition maps.**

I. Introduction

PRECISE 3D object detection is crucial for autonomous driving systems [1], as it predicts the 3D geometric and semantic information of objects surrounding the vehicle. Multi-modal fusion approaches have attracted increasing interest because they capture complementary signals from diverse data sources [2], such as cameras, LiDAR, and offline HD maps, leading to more reliable detection results. However, the fusion approach remains an open question given the inherent and substantial disparities in view, semantics, and scale between different modalities, as shown in Tab. I.

TABLE I
CHARACTERISTICS OF DIFFERENT MODALITIES. PERS.:PERSPECTIVE

Sensors	Geometry	View	Information	Intensity	Object
Camera	vague	pers.	low-level	informative	included
LiDAR	precise	3D	low-level	ignored	included
HD Map	precise	BEV	high-level	missing	excluded

Extensive studies have been conducted on camera-LiDAR fusion given the complementary nature of images and point clouds, exploring various fusion stages, representations, operators, etc. Early works [3] based on LiDAR crop the input point clouds within the region of interest, i.e., camera frustum. Most methods fuse intermediate features [4] or detection results [5] thereafter, building upon sophisticated backbones. Some approaches [6] propose a potentially more generalized approach that unifies representation before fusion. Nevertheless, cameras and LiDAR both provide low-level data with object information, which differentiates them from HD maps and renders the above frameworks not ideal.

Some recent works devise LiDAR-map fusion methods to take advantage of HD maps in perception, incorporating their rich geometric and semantic information of the environment. HDNet [7] pioneers its utilization through rasterizing vector maps and concatenating with projected 3D point clouds in BEV, setting up a common workflow. Subsequent studies [8], [9] delve into representing, encoding, and fusing strategies. Despite the efforts made, existing approaches for HD maps are still under development. Moreover, the restricted view and vague geometry of images could lead to challenges in transplanting LiDAR-map fusion.

Camera-only detection approaches have gained favor within the research community. Recent advancements either leverage 2D paradigms [10] or advance towards 3D modeling [11], [12], [13], making breakthroughs thanks to the development of 2D image understanding. While they benefit from an exceptional cost advantage, they face inherent limitations by relying solely on cameras, specifically the lack of geometries. HD maps have been overlooked among vision based researches, mainly due to insufficient recognition and the challenges associated with their utilization.

Received 9 June 2024; revised 16 September 2024 and 13 December 2024; accepted 17 December 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42471490 and in part by the Major Program (JD) of Hubei Province under Grant 2023BAA026. The Associate Editor for this article was Y. Yu. *(Corresponding author: Jian Zhou.)*

Jinsheng Xiao and Shurui Wang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: xiaojs@whu.edu.cn; shuruiwang@whu.edu.cn).

Jian Zhou and Ziyue Tian are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: jianzhou@whu.edu.cn; ziyue_tian@whu.edu.cn).

Hongping Zhang is with the GNSS Research Center, Wuhan University, Wuhan 430072, China (e-mail: hpzhang@whu.edu.cn).

Yuan-Fang Wang is with the Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: yfwang@cs.ucsb.edu).

Digital Object Identifier 10.1109/TITS.2024.3520814

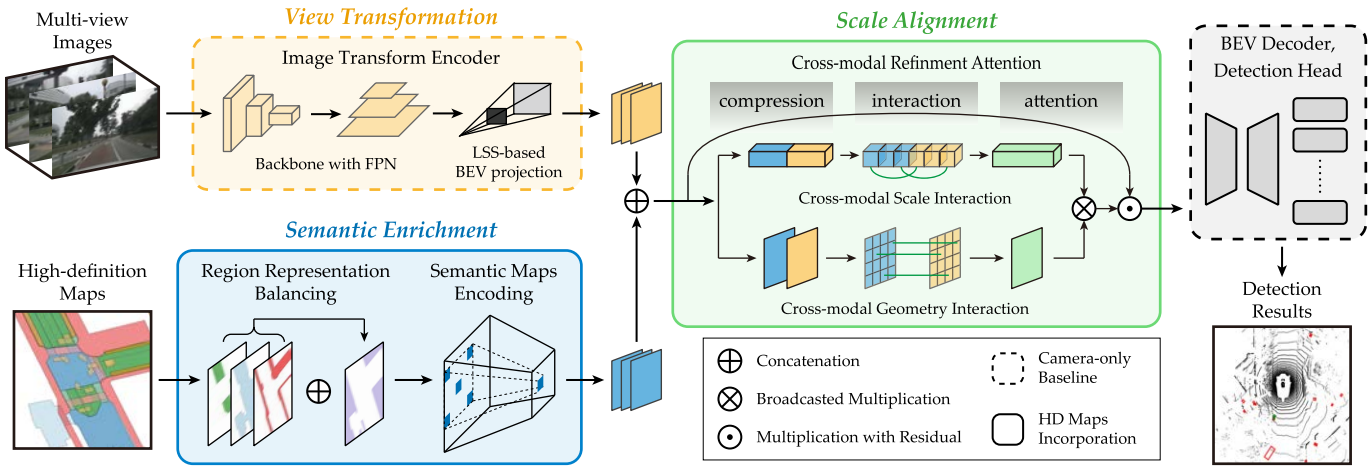


Fig. 1. The proposed framework MIM. The dotted lines mark the parts adopted from the camera-only baseline, which are responsible for transforming views and predicting objects. The solid lines mark the proposed modules, which are responsible for enriching semantics and aligning scales, thereby incorporating HD maps. RRB appends an out-of-map layer to HD maps, followed by SME to encode them using dilated convolutions. The concatenated cross-modal features are handled by CRA to adjust scales and highlight geometries in HD maps, through a series of compression, interaction, and attention in parallel branches. These modules fuse HD maps into the camera-only baseline that depends on BEV. Multiplications are element-wise. Point clouds are for visualization only.

In summary, integrating HD maps can address the limitations of camera-based methods while keeping deployment costs low, but significant discrepancies in view, semantics, and scale must be tackled. To this end, we propose HD Maps Incorporated Multi-view detection (MIM), which first explores camera-map fusion to the best of our knowledge. Similar to our previous work [14] on LiDAR-map fusion, MIM is built upon camera-based methods, as illustrated in Fig. 1.

MIM adopts a decoupled baseline, in which the encoder transforms image features into BEV, aligning them with HD maps in view. Three modules are introduced to incorporate HD maps. First, Region Representation Balancing (RRB) enriches the background of HD maps, balancing the representation of labeled and unlabeled environments. Next, Semantic Maps Encoding (SME) encodes HD maps into high-dimensional features, aligning them closely with image features in semantics. Finally, Cross-modal Refinement Attention (CRA) fuses the features, adjusting scales across modalities and emphasizing geometries. These modules integrate HD maps with image features, while the decoder and detection head from the baseline generate the final predictions.

Experiments were conducted on the large-scale public autonomous driving dataset nuScenes [15]. MIM achieved an mAP of 40.9% and an NDS of 46.3%, surpassing the baseline by 1.7% and 1.9%, respectively, on the validation set, and outperforming camera-only methods on both subsets. The comparison and ablation results demonstrate the effectiveness of HD maps and the proposed modules. Further experiments evaluate the contribution of individual HD map layers, providing insights for future studies. The key contributions of this paper are summarized as follows:

- To explore camera-map fusion, we investigated the rationale behind integrating HD maps with images, identified key challenges to resolve in view, semantics, and scale, and proposed MIM as the first approach of its kind.
- To achieve camera-map fusion, we developed three modules that enrich semantics and align scales on top of a camera-only baseline, which handles view transformation, leading to improved detection performance.
- To deepen the understanding of camera-map fusion, we conducted an in-depth analysis of the operational intricacies of each HD map layer, paving the way for future research in HD map-integrated perception.

II. Related Work

In this section, we first introduce camera-LiDAR fusion detection methods as references for multi-modal architectures. Then we introduce camera-only 3D object detection as the baseline of camera-map fusion, and LiDAR-map fusion perception that explores the utilization of HD maps.

A. Camera-LiDAR Fusion 3D Object Detection

Camera-LiDAR fusion has been extensively studied thanks to the complementary nature of images and point clouds. PointPainting [16] assigns the semantic segmentation labels from images to the 3D points. FS-Net [4] matches image features to key points according to the scale and receptive field. MENet [17] expands the mapping range and levels. Instead of input point clouds, a few works [18], [19] devise to fuse image pixels at voxel-level after point cloud feature extraction. Except for minor differences in image feature extraction, these approaches can also be classified as feature-level fusion, which usually share a common architecture. The middle-fusion or feature-level fusion paradigm [6], [20], [21] fuses output features from backbone networks of each modality and propose various ways to build correspondence between features, including view projection, graph modeling, attention mechanism, etc. In addition, research also explores late-fusion paradigm [22], [23], [24] that fuses detected instances from both modalities, like bounding boxes and feature tokens. It often brings computational advantages for the network since objects are sparse compared to dense sensor data. There are also many methods [5], [25] that combine different level fusion

to further improve detection performance, or propose fusion strategies that are hard to categorize [26], [27].

The tremendous success achieved by camera-LiDAR fusion makes it so representative in the field of multi-modal fusion that other approaches are paid less attention. Extra computation and deployment overhead are inevitable trade for combining camera and LiDAR.

B. Camera-Only 3D Object Detection

1) Following 2D Detection Paradigms: The success of 2D object detection encourages numerous 3D derivatives. FCOS3D [10] extends the advanced 2D anchor-free detector FCOS to predict transformed 3D object geometries directly in the camera view. PGD [28] further constructs geometric relation graphs between objects, facilitating depth estimation. Much work has been done on introducing auxiliary tasks to 2D paradigms. In [29], a 2D segmentation mask is produced using annotated point clouds, which help the model distinguish objects and occlusions. MonoPixel [30] develops a way to utilize point clouds as depth supervision, but in an object-centric manner instead of generating dense depth map [31]. Without point clouds or external data, [32] designs some parallel heads to predict 2D key points projected by 3D bboxes during training, which eases model learning.

2) Developing Paradigms With 3D Modeling: DETR3D [33] follows DETR [34] but projects learnable 3D queries in 2D images, and then samples the corresponding features for end-to-end 3D detection without post-processing. The subsequent study [35] focuses on improving the detection of objects across views via graph modeling. Another concurrent work PETR [12] proposes 3D position embedding to elevate 2D features, an alternative to 3D projection. Other methods propose to generate BEV features from multi-view images. BEVDet [11] adopts LSS [36] and feeds BEV feature into LiDAR-based 3D detection head, demonstrating the feasibility of 3D detection in BEV. Meanwhile, BEVFormer [13] argues for the necessity of LSS and proposes to learn BEV feature adaptively with Transformer blocks. Recent advances enlarge the input window to include more temporal cues, developing hybrid temporal fusion [37] with parallel and recurrent flow.

These methods strongly believe in the irreplaceable cost advantage of camera-only solutions. HD maps, which could serve as an alternative to LiDAR, have been overlooked due to discrepancies between modalities.

C. LiDAR-Map Fusion 3D Perception

HD maps, though providing rich environment information [38], have not gained much attention due to their special properties. To make use of continuous vector maps in discrete network, an early work HDNet [7] proposes to rasterize HD map and concatenate drivable area mask to LiDAR BEV grids, telling the potential of HD maps in 3D perception and put forward the idea of raster representation. MapFusion [8] attempts to extract map features with 2D modules before fusion and introduces a subtask to reproduce the map input from voxel features. LaneFusion [9] further provides a detailed discussion on map representation, fusion type, and map backbone structure, which evidences the advantage of parallel backbone architecture. MENet [14] tests a few attention mechanisms as fusion plugins, finding that dilated convolution is potentially a better choice for HD maps encoding.

LiDAR is the best sensor for capturing precise environment geometries, but LiDAR-map fusion does not address its lack of detailed texture. Moreover, the fusion of HD maps has not been extensively studied.

III. Methods

In this section, we illustrate the proposed camera-map fusion framework MIM. We first introduce the overview and the camera-only baseline, followed by discussions on the ideology of our framework design. We then show the implementation of HD maps incorporation, including three modules in the network and a customized data augmentation strategy.

A. Overview

As displayed in Fig. 1, MIM encodes multi-view images and HD maps in parallel and fuses features in BEV, then feeds the fused features into a decoder and detection head, predicting 3D objects. In terms of resolving discrepancies, the whole network can be divided into four parts: an encoder of multi-view images that aligns views, an encoder of HD maps that aligns semantics, a multi-modal feature fuser that aligns scales, and the remainder that produces detection results. The image transform encoder obtains multi-view image features and transforms them into BEV. RRB simulates the sampling of HD maps with a non-trivial twist in the representations, SME then encodes the sampled local HD maps into features for fusion. CRA further refines the concatenated multi-modal features through adjusting the scales and highlighting the geometries. At last, a BEV decoder and detection head from the baseline predict objects using the fused features.

B. Camera-Only Detection Baseline

To leverage HD maps, a decoupled and scalable multi-view 3D detection baseline is required, therefore BEVDet [11] is adopted. It includes an encoder for images, and a BEV decoder and detection head.

1) Image Transform Encoder: The image transform encoder first extracts multi-scale high-level image features, then converts features from the perspective view to BEV. The feature extraction network constructs a feature pyramid network (FPN) employing the attention-based backbone swin-transformer [39]. The view transformation network is primarily adopted from LSS, involving depth distribution estimation, inverse projection, and BEV pooling.

2) BEV Decoder and Detection Head: The BEV decoder utilizes ResNet blocks to construct FPN, capturing critical cues defined in the BEV space such as scale, orientation, and velocity. The detection head from CenterPoint [40] is adopted, which employs the decoded BEV features to generate 3D bounding boxes, describing their geometric attributes.

C. The Role of Background in Detection

The emphasis on detection accuracy often leads to a lack of interest in HD maps, whose rationality we argue for.

This section aims to analyze the pros and cons of different detection strategies, thereby elucidating the role of background in them.

Let a frame of scene denoted as $\mathcal{U} = \mathcal{S} \cup \mathcal{D}$, in which $\mathcal{S}$ is the set of background and $\mathcal{D}$ is the set of foreground. And let sensor data acquired be $U = \varphi(\mathcal{U})$, where φ represents sampling procedure of the sensor, then a vanilla object detection method following strategy (a) is described as:

$$O = \omega[\mathcal{F}(U)] \tag{1}$$

Here, O is the set of objects detected, and the method is divided into a general feature extraction network $\mathcal{F}$ and a detection head ω. Based on the previous consensus that background is nothing but a distraction to the representation of objects, strategy (b) which expects to remove background can be ideally formulated as:

$$O = \omega[\mathcal{F}(D)] \tag{2}$$

This strategy assumes that the optimal result O is obtained when the input contains only D. The tremendous success achieved in most image recognition tasks demonstrates its superiority over the vanilla strategy (a). However, recent advance [41] holds a different opinion that detection relies on background context when D proves insufficient, detecting tiny objects for example. The proposed strategy (c) is ideally expressed as:

$$O = \omega\{f[\mathcal{F}(D), \mathcal{F}(S)]\} \tag{3}$$

where f represents fusion. The conflict between (b) and (c) motivates us to seek a unified ideology for detection. Comparing Eq. 2 and Eq. 3 tells that (c) should behave no worse than (b), as $\mathcal{F}(S)$ may be discarded by f through learning if it appears noisy to detection. In real-world situations, however, the question lies in the mixed sampled data $U = \varphi(\mathcal{S}) \cup \varphi(\mathcal{D})$. In order to obtain $\mathcal{F}(D)$, research has focused on developing $\mathcal{F}_{sep}$, which ideally should meet Eq. 4:

$$\mathcal{F}_{\text{sep}}(S \cup D) = (\mathcal{F}(S), \mathcal{F}(D)) \tag{4}$$

$\mathcal{F}_{\text{sep}}$ aims to separate objects from background at feature-level, particularly leveraging attention mechanism. Obviously, such an estimation cannot be completely accurate since the question is ill-posed. Noticing the difference between Eq. 1 and Eq. 3, we draw a novel conclusion: the success of (b) comes from *separation* instead of *suppression*. When it comes to the role of S in detection, we claim that *background is noise when mixed with objects, but is helpful when separated.*

Nevertheless, (c) has not gained popularity as (b) in studies. We contribute it to not only the misunderstanding of the role of S but also the difficulty of learning an accurate $\mathcal{F}_{\text{sep}}$. The next subsection discusses the benefits brought by HD maps and how they make (c) a viable option.

D. Involving HD Maps as Background

Before delving into the function of HD maps, we offer at-a-glance design of multi-modal fusion approaches. Denoting inputs from two modalities are $U_1 = \varphi_1(\mathcal{U}), U_2 = \varphi_2(\mathcal{U})$, a widely accepted paradigm is formulated as:

$$O = \omega\{f[\mathcal{F}_1(U_1), \mathcal{F}_2(U_2)]\} \tag{5}$$

where $\mathcal{F}_1, \mathcal{F}_2$ are common feature extraction networks for U_1, U_2, respectively, usually adopted from single-modal approaches. There are other frameworks for camera-LiDAR fusion but Eq. 5 has been the choice for LiDAR-map fusion.

In datasets, HD maps can be regarded as $\mathcal{S}_M \subset \mathcal{S}$ and require procession to be utilized. HDNet [7] proposes to acquire the drivable area as grids and concatenate it with BEV features, which can be described as inputting $\varphi_{da}(\mathcal{S}_M)$ to f. Followers [8], [14] then improve the sampling method φ to represent more information, and introduce simple $\mathcal{F}_C$ from image processing for $\varphi(\mathcal{S}_M)$. Let point clouds be $U_L = \varphi_L(\mathcal{U})$, backbone networks be $\mathcal{F}_L$, modern LiDAR-map fusion approaches obey the following form:

$$O = \omega\{f[\mathcal{F}_L(U_L), \mathcal{F}_C(\varphi(\mathcal{S}_M))]\} \tag{6}$$

Equation 6 shares the same form as Eq. 5, except for the explicit expression of the sampling procedure φ. Perception methods do not get involved with online data acquisition, but need to figure out their own way in using offline HD maps, e.g. various representation methods experimented in [9]. Note that φ decides the content and structure of input map data, which is beyond $\mathcal{F}$ and cannot be merged.

Interestingly, Eq. 6 is similar to Eq. 3 as well, thanks to the static nature of HD maps. We believe this similarity is not coincidental, but rather the underlying reason for the effectiveness of HD map-fused methods, which should also extend to camera-map fusion. We break down this reasoning into three key factors: discretionary sampling, precise annotation, and the natural separation inherent in HD maps, as discussed below.

- **Discretionary sampling.** HD maps deliver descriptions of environment $\mathcal{S}_M \subset \mathcal{S}$, allowing customized sampling methods to meet the need of approaches, alleviating the burden of modality alignment. It is possible to acquire lane directions, define 2D or polygonal data structure, and scale to desired resolutions, to name a few.
- **Precise annotation.** Stronger backbones for extracting higher-level semantics are always the pursuit of researchers, while HD maps deliver high-level precise annotations, relieving the burden of feature extraction and eliminating noise (excluding annotation errors).
- **Natural separation.** The most vital characteristic of HD maps lies in their static nature. We argue that containing no direct cues about objects turns out to benefit strategy (c). The aforementioned analyses are based on the fact that D and S are estimated. However, as it is easy to derive $D = U - S$, learning f satisfying Eq. 7 would be much easier than learning $\mathcal{F}_{\text{sep}}$ provided $\mathcal{S}_M \subset \mathcal{S}$.

$$f[\mathcal{F}(U), \mathcal{F}(\varphi(\mathcal{S}_M))] = \mathcal{F}(D) \tag{7}$$

These characteristics make HD maps an abstract background that fits seamlessly into (c). Besides, it is noteworthy that all discussions do not restrict a certain modality, therefore the success achieved by LiDAR-map fusion would be granted to

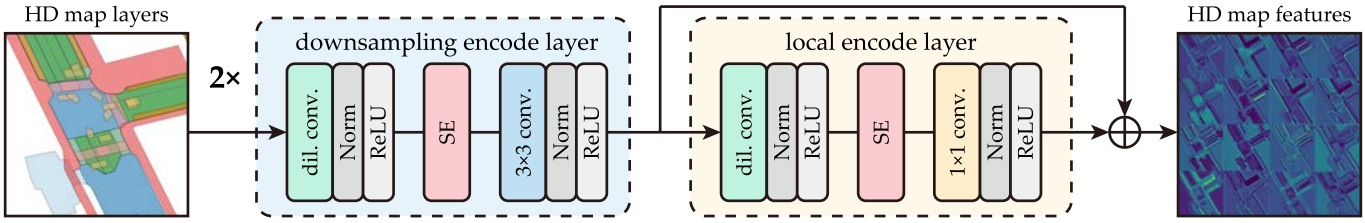


Fig. 2. Illustration of semantic maps encoding. SME comprises three major layers including two downsampling encode layers (DEL) and one local encode layer (LEL). The two types of layers have a very similar structure except for the second convolution. DEL halves the length and height of features and expands the channels. LEL keeps features in the same size, with a residual connection across itself.

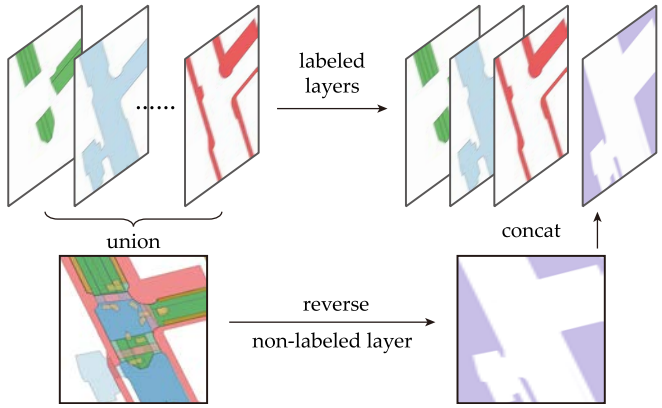


Fig. 3. The process of RRB. Each color indicates a semantic label in the HD maps, except purple for the non-labeled layer. We use 6 labeled semantic layers, only 3 of them are drawn for a simplified view.

camera-map fusion. Summarizing all above, we propose to fuse HD maps with images to enhance 3D object detection. To begin with, we formulate the camera-only baseline as:

$$O = \omega[\mathcal{F}_C(U_C)] \tag{8}$$

where $\mathcal{F}_C$ is the image transform encoder and ω is the BEV decoder and detection head, U_C is multi-view images sampled by cameras. Then our MIM, following Eq. 3 in (c) and Eq. 6 in LiDAR-map fusion, is formulated as:

$$O = \omega\{f_{\text{CRA}}[\mathcal{F}_C(U_C), \mathcal{F}_{\text{SME}}(\varphi_{\text{RRB}}(\mathcal{S}_M))]\} \tag{9}$$

where φ_{RRB} is a simulated sampling module, or representation module, $\mathcal{F}_{\text{SME}}$ is a simple feature extraction module, and f_{CRA} is a multi-modal fusion module.

E. HD Maps Incorporation

1) Region Representation Balancing: As previously discussed, we view the presentation of HD maps as a sampling procedure, which differs from the conventional approach of acquiring supportive information for objects. Therefore, we argue that unlabeled areas in HD maps have been inadvertently overlooked, rendering off-road environments and objects non-existent rather than unknown.

To address this bias to model recognition, the balancing layer is derived as illustrated in Fig. 3. RRB modifies the representation of HD maps to ensure the completeness of the static environment, which is conducive to (c). For a layer $\mathbf{L}$ that only has binary values (labeled or not), we first define layer-wise boolean addition as Eq. 10:

$$\mathbf{L} = \mathbf{L}_i \oplus \mathbf{L}_j \iff \mathbf{L}(x, y) = \mathbf{L}_i(x, y) \oplus \mathbf{L}_j(x, y) \tag{10}$$

where $\oplus$ means boolean addition, (x, y) is a location within the map, $(x, y) \in [0, H) \times [0, W)$. Given the input HD map M is a set of n binary map layers $\mathrm{M} = \{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_n\}$, we obtain the out-of-map layer $\mathbf{L}_b$ that marks all regions without any initial label by Eq. 11:

$$\mathbf{L}_b = \mathbf{1}^{H \times W} - (\mathbf{L}_1 \oplus \mathbf{L}_2 \oplus \cdots \oplus \mathbf{L}_n) \tag{11}$$

where $\mathbf{1}^{H \times W}$ refers to a matrix of shape (H, W) in which all elements are 1. Then the balanced HD maps M_b can be obtained by Eq. 12:

$$\mathrm{M}_b = \mathrm{M} \cup \{\mathbf{L}_b\} \tag{12}$$

In RRB, layers of vector maps are obtained by dividing labeled areas into grids, following common practices. Although $\mathbf{L}_b$ is simple, it puts the unlabeled areas, which take up around 85% of the entire perception space for nuScenes, on equal footing with the other layers. It enhances the contextual information around objects as well as aids in their separation.

2) Semantic Maps Encoding: It has been empirically found and evaluated [9] that "ground-truth" HD maps require encoding. SME aligns local HD maps with image features in terms of semantics and adjusts them to the same resolution. This module is primarily adapted from LiDAR-map fusion [14] due to the similar functions involved.

SME employs multiple convolutional layers for encoding, with Squeeze-Excitation (SE) layers added to enhance its modeling capability. The module consists of three major layers, with the detailed structure shown in Fig.2. The overall procedure can be formulated as Eq.13:

$$\mathcal{F}_{\text{SME}}(M) = LEL[DEL^2(M)] + DEL^2(M) \tag{13}$$

where M represents the input HD map layers processed by RRB, DEL and LEL are the encoding layers. The simplified form of M, containing only class labels, is used, which has been proven sufficient [9].

For input, M is sampled at a higher resolution to preserve the precise structure of vector maps. Two DEL layers downsample it to match the resolution of BEV grids. LEL then encodes it further with a residual connection to enrich semantics. Both DEL and LEL follow Eq. 14:

$$layer(M) = conv_{\text{head}}\{SE[conv_{\text{tail}}(M)]\} \tag{14}$$

where $conv$ refers to convolutional layer, $conv_{\text{head}}$ always employs dilated convolution, as indicated by conclusions from [14], and $conv_{\text{tail}}$ differentiates DEL from LEL. These layers work together to represent HD maps at a lower resolution through expanded channels and continuous values. With map features aligned to BEV grids, simple operations can take place as the basis for multi-modal feature fusion.

3) Cross-Modal Refinement Attention: CRA plays the role of multi-modal fusion, taking effect in BEV. The encoded image and map features are concatenated along channel dimension and fed to parallel channel and spatial attention to reduce scale discrepancy and emphasize geometries, respectively. The refined features are obtained by combining both branches.

The natural discrepancy between images and HD maps can lead to a prominent scale gap between their features. Unfortunately, we cannot determine their scales after encoding, nor the optimal scales for the detection head. The lack of prior forces us to develop a learnable approach for adjustment, which becomes the channel attention branch. Let F denotes the concatenated features, it can be formulated as Eq. 15:

$$\text{channel}(F) = MLP[AvgPool(F)] \tag{15}$$

where $AvgPool$ is global average pooling, and MLP comprises a linear layer, batch norm, and ReLU.

Additionally, the introduction of SME necessitates a tailored design. Drawing insights from position embedding [42], we attempt to compensate for the loss of ground-truth information during the encoding. Since features have already been aligned to BEV grids, this can be addressed by emphasizing geometries through spatial attention, re-weighting the multi-modal features in the spatial domain, formulated as Eq. 16:

$$\text{spatial}(F) = conv\{dil^2[reduce(F)]\} \tag{16}$$

where $conv$ is a standard convolution, $reduce$ is a convolutional layer with a stride of 2, and dil is a dilated convolutional layer with a stride of 1. The $AvgPool$ and $reduce$ shrink the shape of F to save computations, followed by linear or convolutional layers to perform cross-modal interactions. Finally, the refined features F' is obtained with Eq. 17:

$$F' = F \odot \{\mathbf{1} + \text{sigmoid}[\text{channel}(F) \odot \text{spatial}(F)]\} \tag{17}$$

where $\odot$ represents element-wise multiplication, operands are expanded to the shape of F. Adding $\mathbf{1}$ introduces a residual connection between F and F'.

4) HD Maps Constrained Data Augmentation: Data augmentation [43] plays a crucial role in accelerating learning, with some methods [44], [45] achieving notable success on the nuScenes dataset under the default limited learning schedule. However, the widely used CBGS [45] is not feasible in terms of computational cost for us. Therefore, we introduce map-constrained data augmentation (MCDA) based on GT-Paste [44], anticipating that it will offer competitive acceleration while maintaining efficiency.

MCDA begins by constructing a ground-truth database for multi-view images. For each instance in the dataset, we first retrieve their annotations and crop corresponding image patches by projecting 3D bounding boxes onto the images. Low-quality objects that are too large or too small are removed. The image patches for all instances are then stored in the database, along with metadata such as the scene index, instance index, and image view, which are saved in a separate file.

The key to pasting objects lies in the consistency of annotations with respect to autonomous vehicles, regardless of the frame. For a random object sampled from the database, we locate it by projecting its annotated 3D bounding box onto local HD maps at the current frame. Vehicle classes are restricted to the drivable_area, while other classes are confined to any labeled map layer to avoid pasting objects into buildings. Objects that do not meet these criteria are discarded. This sampling process repeats until a desired number of objects have been pasted for each category.

IV. Experiments

In this section, we first show the details of conducting experiments. To validate our method, we then discuss quantitative and qualitative comparison results, and the ablations of proposed modules. Additionally, we analyze the effects of HD maps on each object class and the associations of map layers with spatial attention, exploring their operational dynamics.

A. Experimental Details

1) Data for Camera-Based Detection: The widely used dataset for monocular detection in autonomous driving, KITTI, does not contain HD maps as it was found early. Therefore we use nuScenes [15], another popular large-scale dataset consisting of 1000 scenes of roughly 20s duration each. The key frames contain six RGB images, annotated at 2Hz. These sum up to 40,157 key frames and 1.4 million annotated 3D bounding boxes from 10 categories.

2) Data for HD Maps: In nuScenes, four HD maps from different regions are provided. Each HD map consists of 3 geometry layers representing elements as points and 10 semantic layers labeling areas, all merged into 6 layers for use. The dense and large-scale point cloud used to generate these HD maps is not publicly available in the dataset. Additionally, localization data (including heading) for the autonomous vehicle is provided at every key frame.

3) Evaluation Metrics: The detection metrics given by nuScenes are adopted, where a match is defined by thresholding the 2D center distance d on the ground plane. The mean Average Precision (mAP) is then calculated as the normalized area under the precision-recall curve, and averaged over matching thresholds of $\mathbb{D} = \{0.5, 1, 2, 4\}$ meters and the set of classes $\mathbb{C}$:

$$\text{mAP} = \frac{1}{|\mathbb{C}||\mathbb{D}|} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} \text{AP}_{c,d} \tag{18}$$

Under $d = 2$m matching threshold, five True Positive (TP) metrics are defined to evaluate translation (ATE), scale (ASE), orientation (AOE), velocity (AVE), and attribute (AAE). The mean TP (mTP) over all classes are computed as Eq.19:

$$\text{mTP} = \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} \text{TP}_c \tag{19}$$

TABLE II

QUANTITATIVE COMPARISON ON NUSCENES *validation* SET. †: TRAINED WITH CBGS THAT ENLONGATES ONE EPOCH INTO 4.5 EPOCHES

Method	Resolution	Epoch	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [10]	1600×900	24	34.3	41.5	72.5	26.3	42.2	129.2	15.3
PGD [28]	1600×900	24	36.9	42.8	68.3	26.0	43.9	126.8	18.5
DETR3D [33]	1600×900	90†	34.9	43.4	71.6	26.8	37.9	84.2	20.0
G-DETR3D [35]	1600×900	-	35.1	43.3	-	-	-	-	-
PETR [12]	1600×900	90†	37.0	44.2	71.1	26.7	38.3	86.5	20.1
MonoPixel [30]	800×448	140	33.2	-	69.2	27.1	42.5	-	-
BEVFormer-S [13]	1600×900	24	37.5	44.8	72.5	27.2	39.1	80.2	20.0
BEVDet [11]	1408×512	24	39.2	44.4	63.6	26.9	43.2	96.5	21.5
MIM(Ours)	1408×512	24	**40.9** +1.7	**46.3** +1.9	60.3	26.8	44.2	86.5	23.9

TABLE III

QUANTITATIVE COMPARISON ON NUSCENES *Test* SET

Method	Resolution	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [10]	1600×900	Res101	35.8	42.8	69.0	24.9	45.2	143.4	12.4
PGD [28]	1600×900	Res101	38.6	44.8	62.6	24.5	45.1	150.9	12.7
PETR [12]	-	Res101	39.1	45.5	64.7	25.1	43.3	93.3	14.3
BEVFormer-S [13]	1600×900	Res101-DCN	40.9	46.2	65.0	26.1	43.9	92.5	14.7
BEVDet [11]	1408×512	Swin-B	40.0	45.6	57.8	25.7	44.2	116.9	16.9
MIM(Ours)	1408×512	Swin-B	**42.0** +2.0	**46.6** +1.0	54.1	25.4	48.0	103.3	17.1

TABLE IV

PER-CLASS COMPARISON WITH THE BASELINE ON TWO NUSCENES SUBSETS, EVALUATED WITH AVERAGE PRECISION. *Val.*: *validation*, Δ: DIFFERENCE OVER THE BASELINE, C.V.: CONSTRUCTION VEHICLE, PED.: PEDESTRIAN, T.C.: TRAFFIC CONE

subset	Method	mAP	car	truck	bus	trailer	C.V.	ped.	motor	bicycle	T.C.	barrier
val.	baseline	39.2	58.3	31.8	49.2	17.3	9.0	46.1	38.2	28.7	60.9	51.6
	MIM	40.9	62.5	35.5	48.3	22.8	10.5	48.0	39.2	28.7	60.9	52.3
	Δ	1.7	4.2	3.7	-0.9	5.5	1.5	1.9	1.0	-1.1	0.3	0.7
test	baseline	40.0	56.6	31.9	34.1	31.0	17.2	42.3	41.9	28.6	63.1	53.7
	MIM	42.0	59.9	32.7	40.5	33.6	15.7	44.3	43.1	31.0	63.8	55.6
	Δ	2.0	3.3	0.8	6.4	2.6	-1.5	2.0	1.2	2.4	0.7	2.0

Finally, the nuScenes Detection Score (NDS) is defined based half on the Average Precision metric mAP, half on the set of five TP metrics $\mathbb{TP}$, formulated as Eq. 20:

$$\text{NDS} = \frac{1}{10}\left[5 \cdot \text{mAP} + \sum_{\text{mTP}\in\mathbb{TP}} (1 - \min(1, \text{mTP}))\right] \quad (20)$$

4) Model Configurations: In our modules, a convolutional layer comprises convolution with a 3×3 kernel, batch norm, and ReLU. Specifically, the $conv_{\text{tail}}$ in LEL, SME, the first and the last convolution in the spatial(·), CRA, use a 1×1 kernel. Inputs and BEV features are augmented mainly following BEVDet, with the exception of using MCDA instead of CBGS. MCDA is disabled in the second half schedule to help the model learn data distribution in real scenes.

5) Optimization: Models are trained and tested on a single NVIDIA GeForce RTX 3090 GPU. Optimization is carried out using AdamW with a weight decay of 10^{-2}. Under a resolution of 256×704, the batch size is 16 and learning rate is 5×10^{-5}. Under a resolution of 512×1408, the batch size is 1 and learning rate is 3×10^{-6}, and *batch norms* are replaced with *layer norms* since the former fails at batch size 1. The cyclic policy is adopted as common, and the total schedule is terminated within 24 epochs by default.

B. Comparison Results

1) Quantitative Results: We compare MIM with recent camera-only methods on the nuScenes *validation* set. Competitors include monocular methods FCOS3D, PGD, and MonoPixel, DETR based methods DETR3D, G-DETR3D, and PETR, BEV based methods BEVDet and BEVFormer-S. BEVDet is MIM retrained under the same settings, but without HD maps and incorporation modules, ensuring a fair comparison. Other methods are loyal to their original papers. As illustrated in Tab. II, MIM boosted BEVDet by 1.7% in mAP and 1.9% in NDS, achieving a remarkable performance with an mAP of 40.9% and NDS of 46.3%.

Moreover, the increase is achieved at minimal computational cost. As shown in Tab. VI, the three modules introduced add only 0.3M parameters and 8.3G FLOPs compared to BEVDet. Notably, RRB incurs almost no FLOPs on its own, with the slight increase attributed to the appended map layer to encode in SME. This negligible computational cost results in a comparable running speed. The high overall FLOPs is primarily due to the Swin-B backbone. PETR's FLOPs is significantly lower, but this is for a single view, whereas MIM and other methods process six views.

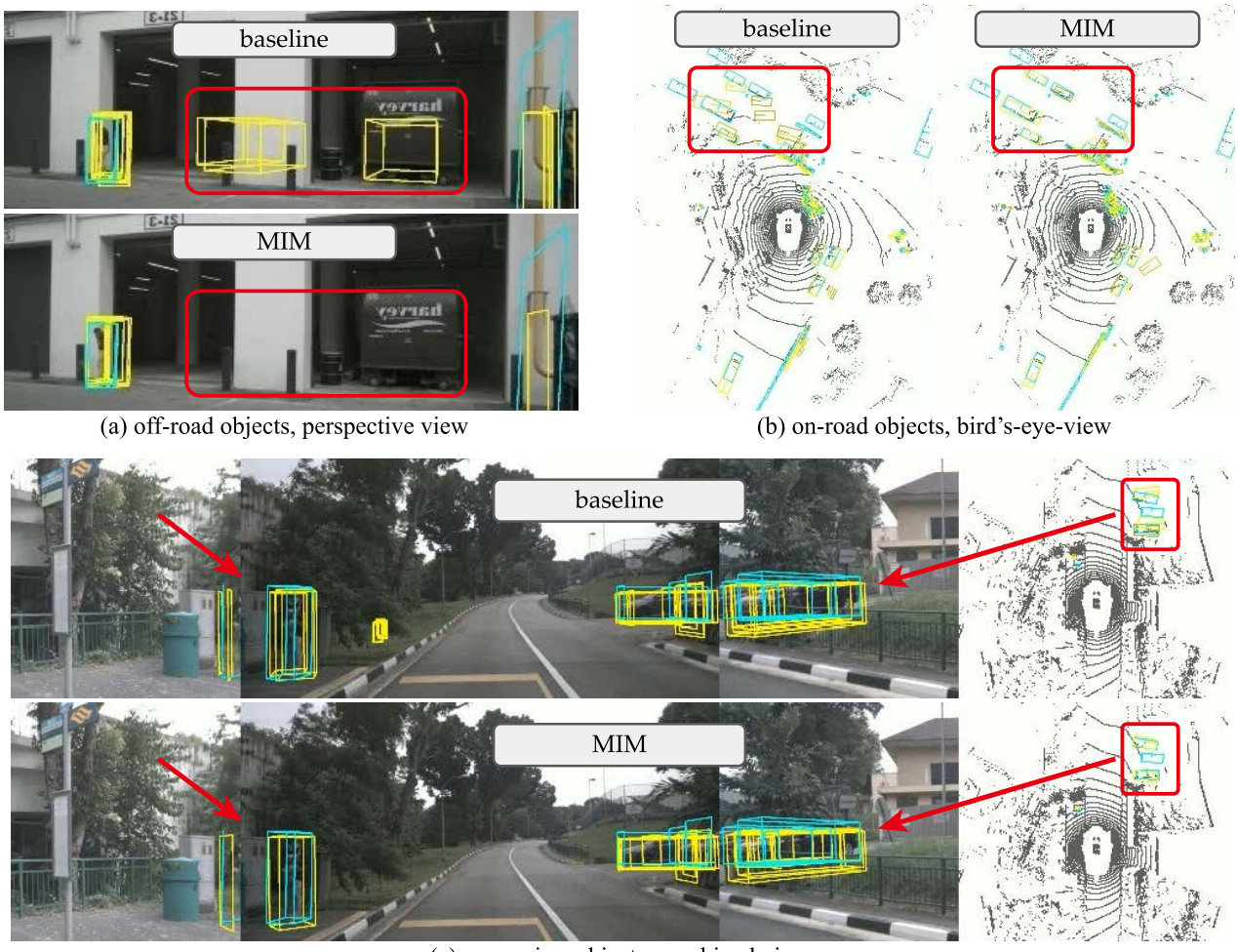
We then compare MIM with other available methods on the nuScenes *test* set in Tab. III. Considering the results across both subsets, our method demonstrates a substantial improvement over the baseline. It excels in predicting objects' locations with the assistance of HD maps, albeit sometimes at the expense of predicting orientations and attributes.

To validate HD maps' effect on different classes, we present a detailed per-class comparison in Tab. IV. The detection on vehicles enjoys a significantly higher AP increase

TABLE V

INCORPORATING HD MAPS TO THE LIDAR-BASED APPROACHES, EVALUATED ON NUSCENES *validation* SET. C.V.: CONSTRUCTION VEHICLE, PED.: PEDESTRIAN, T.C.: TRAFFIC CONE

Method	HD maps	mAP	NDS	car	truck	bus	trailer	C.V.	ped.	motor	bicycle	T.C.	barrier
CenterPoint [40]		52.7	61.2	82.9	52.0	62.6	35.7	12.5	79.8	44.7	32.8	59.8	64.4
MENet [14]	✓	55.7	62.6	81.9	52.6	65.6	38.9	17.9	78.8	60.0	45.0	59.3	57.4
MIM(Ours)	✓	56.4	63.4	82.0	53.7	63.8	42.0	17.3	78.3	61.3	43.5	59.9	61.1



(a) off-road objects, perspective view

(b) on-road objects, bird's-eye-view

(c) cross-view objects, combined view

Fig. 4. Visualizations of detection from the baseline and MIM. Cyan boxes are ground-truths and yellow boxes are predicted targets, red rectangles and circles mark the major differences. Images are cropped and rearranged for a better view, point clouds are for visualization only.

TABLE VI

COMPARISON AND DECOMPOSITION OF MODEL COMPLEXITY

Method	Param.	FLOPs	FPS	Module	FLOPs
PETR	56.5M	*162.0G	2.1	RRB	0.15G
BEVFormer-S	68.7M	1303.5G	-	SME	7.95G
BEVDet	125.1M	1893.9G	2.5	CRA	0.23G
MIM	125.4M	1906.2G	2.4	Swin-B	1407G

(3.2% in average), which aligns with expectations since HD maps provide detailed annotations for on-road elements. C.V. is the only vehicle class that shows no overall increase. We attribute this to its arbitrary location on HD maps and shared detection sub-head with the truck. The increase observed in other classes (1.1% in average) indicates that HD maps also contribute to the detection of non-vehicle objects.

We further extended our evaluation by applying MIM to LiDAR-based approaches, as shown in Tab. V. This transplantation shows a clear improvement over the CenterPoint baseline [40] and a competitive performance compared to the state-of-the-art [14]. MIM achieves the highest mAP and NDS scores, offering a general performance increase.

2) Qualitative Results: We compared MIM with the baseline on the nuScenes *validation* set. Fig. 4(a) shows that MIM distinguishes off-road objects from cars where the baseline suffers from their similar appearance. Fig. 4(b) displays a heavy-loaded intersection with numbers of targets. MIM keeps the high detection accuracy learning from lanes in HD maps when the baseline becomes incapable. Fig. 4(c) illustrates an easy environment with hard samples, including a person and partly occluded parked cars appearing across views.

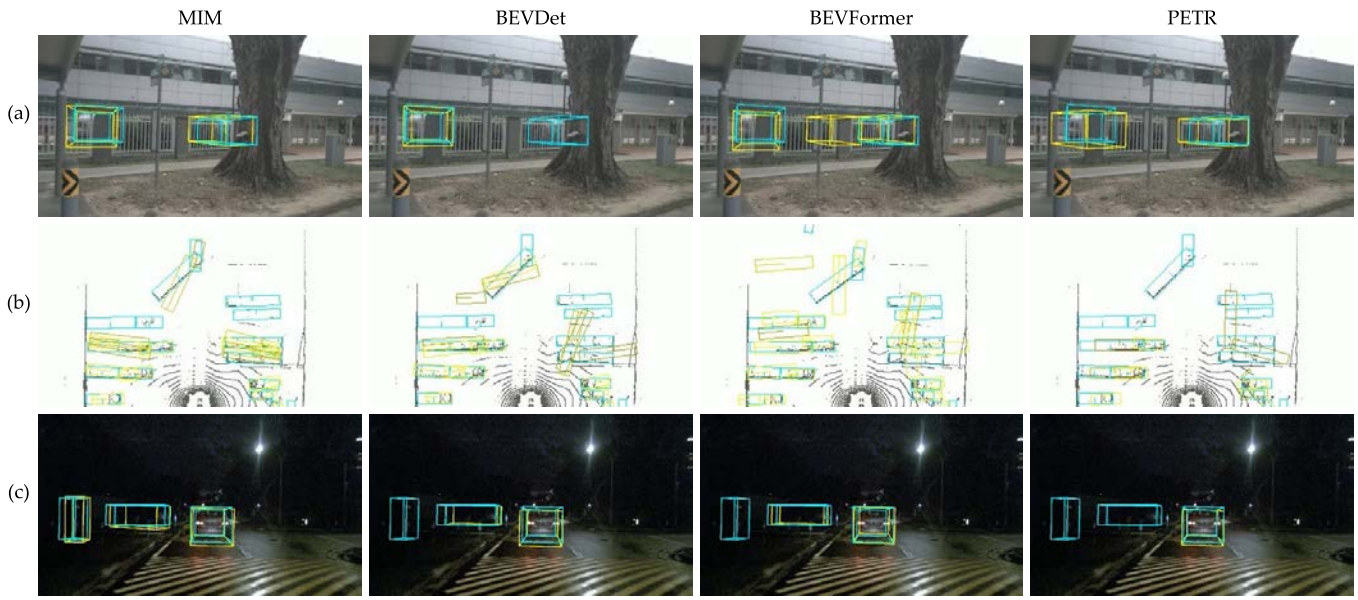


Fig. 5. Visualization comparisons with the state-of-the-arts. The BEV view is synthetic and point clouds are for visualization only.

MIM has a smaller prediction error and one less false detection compared to the baseline.

More visualization comparisons with the state-of-the-arts are displayed as Fig. 5. Models with similar performance are tested if reported ones are not public. In Fig.5(a), MIM detects two off-road vehicles without false positive or false negative at a higher precision than PETR. The long trucks in Fig.5(b) are hard to distinguish since they occlude each other, and MIM delivers a better result regarding those in the front and to the right, especially in their heading. For the comprehensiveness of evaluation, Fig.5(c) includes a scene in extreme lighting with non-vehicle objects, where MIM is the only method that detects the person on the roadside.

C. Ablation Studies

1) Overall: By default, ablation experiments were carried out on nuScenes validation set under 704×256 resolution with MCDA. Note that RRB refers to appending the out-of-map layer, as HD maps cannot be directly used without sampling. The results are presented in Tab. VII. A1 proves that introducing HD maps without further handling does not improve perception. A2 to A4 show that each proposed module contributes to a certain performance increase, with SME being the most significant. This finding is reasonable as SME is essential for fusion, while RRB and CRA play auxiliary roles. Collectively, the proposed modules improve the fusion effect by 3.3% in both mAP and NDS.

2) Region Representation Balancing: Experiments are conducted on various compositions of representing and fusing HD maps, listed in Tab. VIII. It shows that models with RRB achieved higher mAP and NDS than models without it. An average increase of 0.3% mAP and 0.6% NDS under three different fusion approaches is non-trivial, supporting the idea of sampling and the out-of-map layer.

3) Semantic Maps Encoding: Experiments are conducted to investigate the proper map feature dimension in SME.

TABLE VII

ABLATIONS OF MIM. THE BASELINE DOES NOT USE HD MAPS. EXPERIMENTS ARE CONDUCTED ON NUSCENES *validation* SET UNDER 704 × 256 RESOLUTION WITH MCDA BY DEFAULT

	RRB	SME	CRA	mAP	NDS
baseline				32.0	38.6
A1				31.3	39.3
A2	✓			32.0	39.5
A3		✓		34.1	41.8
A4			✓	32.1	39.8
A5	✓	✓	✓	**34.6**	**42.6**

TABLE VIII

ABLATIONS OF RRB AND CRA USING DIFFERENT FUSION METHODS. Δ: CHANGE FROM APPENDING OUT-OF-MAP LAYER

		mAP			NDS		
with RRB			✓	Δ		✓	Δ
B1	concat	34.1	34.5	0.4	41.8	42.2	0.4
B2	mult	33.9	34.2	0.3	41.0	42.1	1.1
B3	CRA	34.4	**34.6**	0.2	42.2	**42.6**	0.4

TABLE IX

ABLATIONS ON THE MAP FEATURE DIMENSION IN SME, NO MCDA

channels	4	8	16	32	64
mAP	31.5	32.1	32.2	32.0	32.0
NDS	38.7	40.0	40.1	40.1	39.8

The number is set to 64 in other low-resolution experiments to test different fusion strategies. Intuitively, the alignment in dimension may benefit alignment in semantics. However, Tab. IX shows that 16 is a better option, although the influence is minor in most cases. The improvement from expanding channels is substantial when the number is less than 8, but saturates quickly as it becomes larger.

4) Cross-modal Refinement Attention: Tab. VIII illustrates that the model with a combination of RRB and CRA achieved the best mAP of 34.6% and NDS of 42.6%. Comparing

TABLE X
ABLATIONS OF AUGMENTATION TECHNIQUES

augmentation	None	GT-Paste	MCDA
mAP	32.0	34.2	34.6
NDS	39.8	42.5	42.6

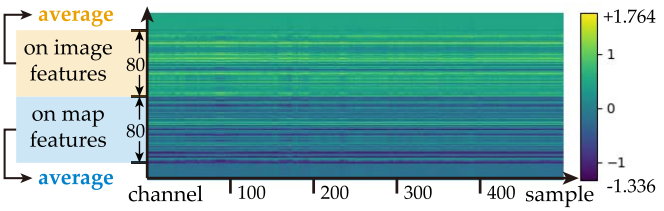


Fig. 6. Visualization of channel attentions, the light colors represent higher attention than the dark colors. The attention on image features (top half) is generally higher than that on map features (bottom half).

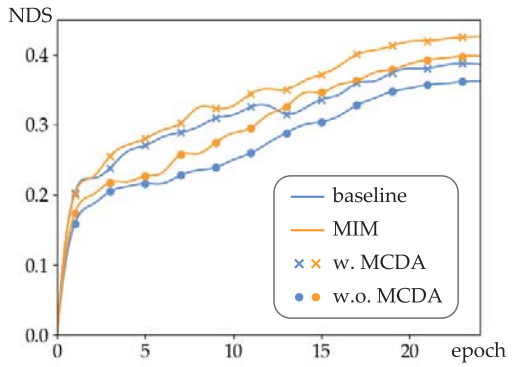


Fig. 7. The evolution of NDS with or without MCDA, experimented on the MIM and camera-only baseline BEVDet.

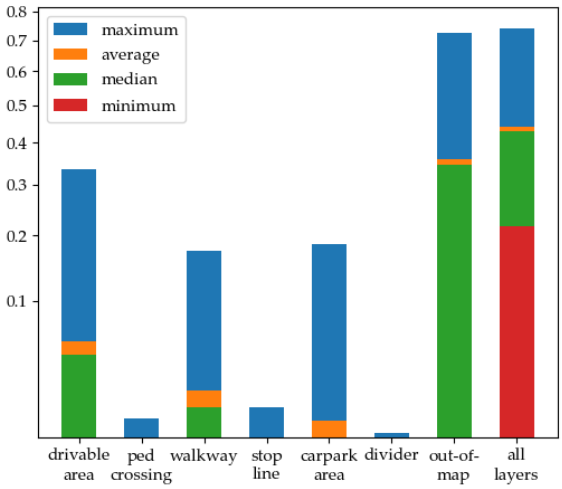


Fig. 8. SSIM between the HD map layers and spatial attention, including the out-of-map layer and the sum of all layers. Note that "all layers" is indicative rather than quantitative, as the sum of SSIM is not meaningful. The axis is stretched to focus on the lower range where most values lie in.

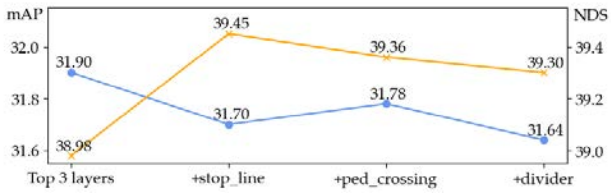


Fig. 9. Ablations on the HD map layers used, incrementally appending layers. Top 3 layers include *drivable_area*, *walkway*, and *carpark_area*.

different rows, B3 takes a steady lead in mAP and NDS, demonstrating the necessity of CRA. Using the best model under 1408×512 resolution, the channel attentions learned by CRA are visualized as Fig. 6, which remain highly consistent across different samples. This indicates that the natural scale gap is the major disparity beyond sample-wise differences, and CRA learns a stable ratio for adjustment.

5) Map Constrained Data Augmentation: A comparison of MIM using different augmentations is shown in Tab. X, where GT-Paste refers to our image-based MCDA without HD maps for filtering samples. GT-Paste improves MIM by a substantial 2.2% in mAP and 2.7% in NDS, taking only about 8% overall extra training time. These gains are further enhanced by an additional 0.4% in mAP and 0.1% in NDS when HD maps filtering is applied at negligible time cost. Additionally, Fig. 7 shows that MCDA provides a comparable performance boost to the baseline, making it easily transferable.

D. HD Maps Analysis

1) Evaluation Methodology: We conducted experiments on the *validation* set to delve into the contributions of different layers of HD maps, aiming to unveil their underlying functionality. The local HD map layers $\varphi_{\text{RRB}}(\mathcal{S}_M)$ are compared to the spatial attention f_{CRA}, which is more straightforward than compared to multi-channel map features and helps identify the role of CRA. We utilized Structural Similarity (SSIM) as the evaluation metric due to its descriptive name, and the fact that shapes directly reflect geometries in BEV. Layers were resized to 1/4 in height and width using bilinear interpolation to match the dimensions of spatial attention.

2) Contribution Layout w.r.t. Layers: We initiated a straightforward experiment to assess the statistics of SSIM between layers and attention, as depicted in Fig. 8. It shows a clear hierarchical pattern in the contributions, with the majority from *out-of-map*. Then follows *drivable_area* with the only non-zero minimum SSIM of 0.002, underscoring its significance. The *carpark_area* exhibits a maximum SSIM comparable to *walkway*, yet its median value hovers near 0 due to its infrequent appearance. The SSIM values for the others are notably low. We posit that these semantic layers heavily rely on temporal information and traffic signs for contextual understanding. Absent such cues, discerning navigation cues for vehicles or pedestrians becomes challenging.

From a result-oriented perspective, the significance of each layer is evaluated by discarding it, as shown in Fig. 9. Discarding the *stop_line* layer results in a change of -0.47% in NDS and $+0.20\%$ in mAP, a small overall degradation. Accounting for noise, the performance remains nearly unchanged with or without the other two less important layers.

Besides, Fig. 8 highlights that while individual layers may register minimum SSIM values of 0, their collective impact exceeds 0.2. In other words, it is the combined effect of multiple layers rather than any specific layer that consistently contributes to the attention. To further understand this

Fig. 10. Composition of spatial attention w.r.t. layers and Top-k. Top-k refers to the layer with the k-th largest average SSIM.

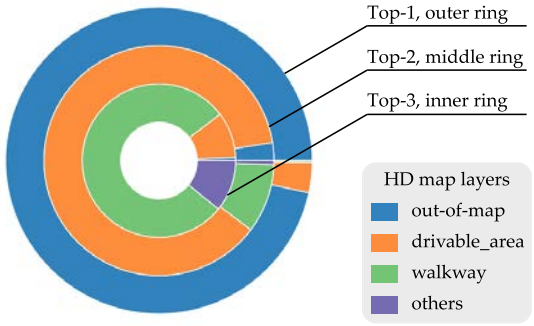


Fig. 11. Ratios of HD map layers at the Top-k position.

dynamic, we analyze the composition of attention in a new view of Top-k order, as discussed below.

3) Contribution Layout w.r.t. Order: In Fig. 10, the top row illustrates the layer-wise decomposition, providing a direct comparison between layers. The areas for *ped_crossing*, *stop_line*, and *divider* are very narrow. Notable traits emerge from the bottom row, Top-k decomposition: the Top-1 ratio is surprisingly high, and the Top-1 to Top-3 determines the major influence of HD maps on attention. These findings suggest that attention can be selective in utilizing different layers. Furthermore, the similar composition between layer-wise and Top-k motivates us to investigate their associations. Consequently, we conducted an experiment whose results are presented in Fig. 11.

Fig. 11 reveals that in typical scenarios, the ranking of layers based on SSIM remains stable. The Top-1 to Top-3 layers follow the same order as discussed earlier, with little variation. The likelihood of variation increases as the importance of the layer decreases. For instance, out-of-map consistently occupies over 95% of the Top-1 position, while this figure decreases to around 80% for *walkway* at the Top-3 position. These high ratios indicate that the median case represents the most common occurrence.

To comprehensively investigate the impact factors of contribution, we present the standard deviation (std) of SSIM concerning layer labels and Top-k in Tab.XI. Interestingly, except for Top-3, the stds at Top-k are smaller compared to those at labels, indicating that the substituted layers closely match the average SSIM at their respective positions. Furthermore, we note that the sole increase in std results from including *carpark_area* in Top-3, which lowers the SSIM for Top-4 below 0.01. These observations indicate that the model prioritizes 3 key layers from the HD maps when constructing

TABLE XI

SORTED STD OF SSIM REGARDING LAYER AND TOP-K ON NUSCENES *validation* SET UNDER 1408 × 512 RESOLUTION. UNIT: 10^{-2}

order	1	2	3	4	5	6	7
layer	11.9	4.7	2.2	2.1	0.2	0.1	<0.1
Top-k	11.3	3.9	2.3	0.3	0.1	<0.1	<0.1

TABLE XII

INFLUENCE OF HD MAPS SAMPLING RESOLUTION ON NUSCENES *validation* SET UNDER 704 × 256 RESOLUTION

Resolution	0.2m	0.4m	0.8m	None
mAP	34.6	34.2	34.0	32.0
NDS	42.6	42.0	41.9	38.6

the attention, and their contributions are more influenced by their selection order than their labels.

4) Resolution for Sampling HD Maps: Experiments were conducted to evaluate the model at different HD map resolutions. The default sampling resolution is 0.2m×0.2m per grid. We increased the grid size to 0.4m and 0.8m and compared the results with the baseline. Convolution strides in SME were adjusted accordingly to maintain spatial alignment of the features. As shown in Tab. XII, although a lower resolution results in a performance drop, the model sampling at 0.8m resolution still achieves an mAP of 34.0% and NDS of 41.9%, which are 2.0% and 3.3% higher than the baseline, respectively. These results demonstrate that the performance decrease at lower sampling resolutions is acceptable, suggesting potential applications with more accessible standard-definition maps.

In conclusion, the addition of the *out-of-map* layer from RRB significantly influences attention. Among the six semantic layers, the *drivable_area*, *walkway*, and *carpark_area* are favored by the model, potentially due to the absence of temporal cues. Further empirical experiments shed light on the operational dynamics of HD maps.

V. CONCLUSION

In this paper, the absence of research on fusing HD maps and multi-view images and the associated challenges are discussed. Therefore, a unified framework MIM is proposed to explore camera-map fusion. To start with, the reason *why* to fuse HD maps in object detection is analyzed, consequently a solution on *how* to fuse is introduced. Recognizing disparities in view, semantics, and scale, MIM employs a baseline and develops three modules to align HD maps and images. The baseline transforms image features to align views, RRB and SME enrich and encode the HD maps to align semantics, and CRA uses attention to align scales. Extensive experiments conducted on the nuScenes dataset demonstrate the efficacy of MIM and its modules, and importantly, identify *what* to fuse in the HD maps. An in-depth analysis reveals the operational dynamics of HD maps in object detection. We envision that MIM could serve as inspiration for further investigations into HD maps-integrated 3D perception.

ACKNOWLEDGMENT

The numerical calculations were done at the Supercomputing Center of Wuhan University.

REFERENCES

[1] J. Mao, S. Shi, X. Wang, and H. Li, "3D object detection for autonomous driving: A comprehensive survey," *Int. J. Comput. Vis.*, vol. 131, pp. 1–55, Apr. 2023.

[2] X. Wang, K. Li, and A. Chehri, "Multi-sensor fusion technology for 3D object detection in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1148–1165, Feb. 2024.

[3] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927.

[4] L. Zhang et al., "FS-Net: LiDAR-camera fusion with matched scale for 3D object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12154–12165, Nov. 2023.

[5] Y. Xie et al., "SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17591–17602.

[6] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2774–2781.

[7] B. Yang, M. Liang, and R. Urtasun, "HDNet: Exploiting hd maps for 3D object detection," in *Proc. Conf. Robot Learn.*, 2018, pp. 146–155.

[8] J. Fang, D. Zhou, X. Song, and L. Zhang, "MapFusion: A general framework for 3D object detection with HDMaps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 3406–3413.

[9] T. Fujimoto, S. Tanaka, and S. Kato, "LaneFusion: 3D object detection with rasterized lane map," in *Proc. 4th IEEE Intell. Vehicles Symp.*, Jun. 2022, pp. 396–403.

[10] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 913–922.

[11] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[12] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 531–548.

[13] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 1–18.

[14] Y. Huang et al., "MENet: Map-enhanced 3D object detection in bird's-eye view for LiDAR point clouds," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, Jun. 2023, Art. no. 103337.

[15] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[16] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4603–4611.

[17] M. Liu et al., "MENet: Multi-modal mapping enhancement network for 3D object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9397–9410, Aug. 2024.

[18] H. Liu, J. Du, Y. Zhang, H. Zhang, and J. Zeng, "PVConvNet: Pixel-voxel sparse convolution for multimodal 3D object detection," *Pattern Recognit.*, vol. 149, May 2024, Art. no. 110284.

[19] C.-H. Wang, H.-W. Chen, Y. Chen, P.-Y. Hsiao, and L.-C. Fu, "VoPiFNet: Voxel-pixel fusion network for multi-class 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8527–8537, Aug. 2024.

[20] Z. Song, C. Jia, L. Yang, H. Wei, and L. Liu, "GraphAlign++: An accurate feature alignment by graph matching for multi-modal 3D object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2619–2632, Apr. 2024.

[21] B. Ding, J. Xie, J. Nie, Y. Wu, and J. Cao, "C^2BG-Net: Cross-modality and cross-scale balance network with global semantics for multi-modal 3D object detection," *Neural Netw.*, vol. 179, Nov. 2024, Art. no. 106535.

[22] S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR object candidates fusion for 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10386–10393.

[23] H. Zhang, J. Wan, Z. He, J. Song, Y. Yang, and D. Yuan, "Sparse agent transformer for unified voxel and image feature extraction and fusion," *Inf. Fusion*, vol. 110, Oct. 2024, Art. no. 102455.

[24] Y. Li et al., "Fully sparse fusion for 3D object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 11, pp. 7217–7231, Nov. 2024.

[25] Z. Liu et al., "Multi-modal 3D object detection by box matching," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19917–19928, Dec. 2024.

[26] Y. Tang, H. He, Y. Wang, and J. Wu, "Towards efficient multi-modal 3D object detection: Homogeneous sparse fuse network," *Expert Syst. Appl.*, vol. 256, Dec. 2024, Art. no. 124945.

[27] G. K. Erabati and H. Araujo, "SRFDet3D: Sparse region fusion based 3D object detection," *Neurocomputing*, vol. 593, Aug. 2024, Art. no. 127814.

[28] T. Wang, Z. H. U. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Proc. Conf. Robot Learn.*, 2022, pp. 1475–1485.

[29] W. Chen, J. Zhao, W.-L. Zhao, and S.-Y. Wu, "Shape-aware monocular 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6416–6424, Jun. 2023.

[30] J.-H. Chen, J.-L. Shieh, M. A. Haq, and S.-J. Ruan, "Monocular 3D object detection utilizing auxiliary learning with deformable convolution," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2424–2436, Mar. 2024.

[31] C. Feng, Z. Jie, Y. Zhong, X. Chu, and L. Ma, "AeDet: Azimuth-invariant multi-view 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21580–21588.

[32] Y. Kim, S. Kim, S. Sim, J. W. Choi, and D. Kum, "Boosting monocular 3D object detection with object-centric auxiliary depth supervision," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 1801–1813, Feb. 2022.

[33] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2022, pp. 180–191.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[35] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Graph-DETR3D: Rethinking overlapping regions for multi-view 3D object detection," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5999–6008.

[36] J. Philion and S. Fidler, "Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 194–210.

[37] Z. Li, S. Lan, J. M. G. Y. F. Valles, and Z. Wu, "BEVNeXt: Reviving dense BEV frameworks for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 20113–20123.

[38] J. Zhou, Y. Guo, Y. Bian, Y. Huang, and B. Li, "Lane information extraction for high definition maps using crowdsourced data," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7780–7790, Jul. 2023.

[39] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[40] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.

[41] J. Xiao et al., "Tiny object detection with context enhancement and feature purification," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118665.

[42] C. R. Qi, Y. Li, H. Su, and L. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 1–10.

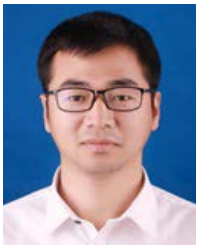
[43] C. Wang, C. Ma, M. Zhu, and X. Yang, "PointAugmenting: Cross-modal augmentation for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11794–11803.

[44] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[45] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3D object detection," 2019, *arXiv:1908.09492*.

Jinsheng Xiao (Senior Member, IEEE) received the Ph.D. degree in computational mathematics from Wuhan University, Wuhan, China, in 2001. From August 2014 to August 2015, he was a Visiting Scholar with the University of California at Santa Barbara, Santa Barbara, CA, USA. He is currently an Associate Professor of information and communication engineering with the School of Electronic Information, Wuhan University. He has authored or co-authored more than 50 scientific articles in journals, books, and conference proceedings. His research interests include image and video processing, computer vision, and artificial intelligence.

Shurui Wang received the B.S. degree from the School of Electronic Information, Wuhan University, Wuhan, China, in 2022, where he is currently pursuing the master's degree. His research interests include computer vision, 3D perception, and multi-modal fusion.



Jian Zhou (Member, IEEE) received the Ph.D. degree in cartography and geographic information system from Wuhan University, Wuhan, China, in 2019. He is currently an Associate Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include high definition map, sensor fusion localization, and computer vision.

Ziyue Tian received the bachelor's degree in computer science and technology from Wuhan Textile University in 2023. He is currently pursuing the master's degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include high-definition maps, deep learning, and autonomous vehicles.

Hongping Zhang received the Ph.D. degree in astrometry and celestial mechanics from Shanghai Astronomical Observatory, Chinese Academy of Sciences, China, in 2006. He is currently a Professor with the GNSS Research Center, Wuhan University. His research interests include GNSS precise positioning and GNSS engineering survey.

Yuan-Fang Wang received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin. He joined the Department of Computer Science, University of California at Santa Barbara, in 1987, where he is currently a Professor. His research interests include computer vision, machine learning, computer graphics, and robotics.