Information Fusion 119 (2025) 103018

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

MapFusion: A novel BEV feature fusion network for multi-modal map construction

Xiaoshuai Hao [a], Yunfeng Diao [b,*], Mengchuan Wei [c], Yifan Yang [c], Peng Hao [c], Rong Yin [d,*], Hui Zhang [c], Weiming Li [c], Shu Zhao [e], Yu Liu [f]

[a] Beijing Academy of Artificial Intelligence, 150 Chengfu Road, Haidian District, 100083, Beijing, China
[b] School of Computer Science and Information Engineering, Hefei University of Technology, Shushan District, 230009, Hefei, China
[c] Samsung R&D Institute China–Beijing, No. 12, Taiyangong Middle Road, 100028, Beijing, China
[d] Institute of Information Engineering, Chinese Academy of Sciences, No. 19, Rd. Shucun, 100195, Beijing, China
[e] Pennsylvania State University, State College, 16801, PA, United States
[f] Department of Biomedical Engineering, Hefei University of Technology, Shushan District, 230009, Hefei, China

ARTICLE INFO

Keywords:
BEV feature fusion
Cross-modal interaction
Dual dynamic fusion
Multi-modal map construction

ABSTRACT

Map construction task plays a vital role in providing precise and comprehensive static environmental information essential for autonomous driving systems. Primary sensors include cameras and LiDAR, with configurations varying between camera-only, LiDAR-only, or camera-LiDAR fusion, based on cost-performance considerations. While fusion-based methods typically perform best, existing approaches often neglect modality interaction and rely on simple fusion strategies, which suffer from the problems of misalignment and information loss. To address these issues, we propose *MapFusion*, a novel multi-modal Bird's-Eye View (BEV) feature fusion method for map construction. Specifically, to solve the semantic misalignment problem between camera and LiDAR BEV features, we introduce the Cross-modal Interaction Transform (CIT) module, enabling interaction between two BEV feature spaces and enhancing feature representation through a self-attention mechanism. Additionally, we propose an effective Dual Dynamic Fusion (DDF) module to adaptively select valuable information from different modalities, which can take full advantage of the inherent information between different modalities. Moreover, *MapFusion* is designed to be simple and plug-and-play, easily integrated into existing pipelines. We evaluate *MapFusion* on two map construction tasks, including High-definition (HD) map and BEV map segmentation, to show its versatility and effectiveness. Compared with the state-of-the-art methods, MapFusion achieves 3.6% and 6.2% absolute improvements on the HD map construction and BEV map segmentation tasks on the nuScenes dataset, respectively, demonstrating the superiority of our approach.

1. Introduction

Map construction task provides abundant and precise static environmental information of the driving scene, which is vital yet challenging for planning in autonomous driving systems. Recently, researchers have focused on two crucial tasks: High-definition (HD) map construction and semantic map construction. Both tasks increasingly utilize the Bird's Eye View (BEV) representation as an ideal feature space for multi-view perception, thanks to its effectiveness in addressing scale ambiguity and occlusion challenges. Specifically, HD map construction methods [1–7] consider this task as the problem of predicting a collection of vectorized static map elements in bird's-eye view (BEV), such as pedestrian crossing, lane divider, road boundaries, etc. On the other hand, semantic map construction methods [8–11] treat map construction as a BEV semantic segmentation task, where each pixel in the BEV plane is assigned a semantic label.

Based on the input sensor modality, map construction methods can be categorized into camera based [6,12,13], LiDAR based [14,15] and camera-LiDAR fusion [1–3,11,16,17] methods. Camera sensors capture rich semantic information, but methods relying solely on them often struggle with spatial distortions when projecting Perspective View (PV) features into Bird's Eye View (BEV) using geometric priors. In contrast, LiDAR provides explicit geometric data with point-wise depth information, though it faces challenges related to data sparsity

* Corresponding authors.
E-mail addresses: xshao@baai.ac.cn (X. Hao), diaoyunfeng@hfut.edu.cn (Y. Diao), mc.wei@samsung.com (M. Wei), yifan.yang@samsung.com (Y. Yang), peng1.hao@samsung.com (P. Hao), yinrong@iie.ac.cn (R. Yin), hui123.zhang@samsung.com (H. Zhang), weiming.li@samsung.com (W. Li), smz5505@psu.edu (S. Zhao), yuliu@hfut.edu.cn (Y. Liu).

https://doi.org/10.1016/j.inffus.2025.103018
Received 21 July 2024; Received in revised form 3 December 2024; Accepted 4 February 2025
Available online 18 February 2025

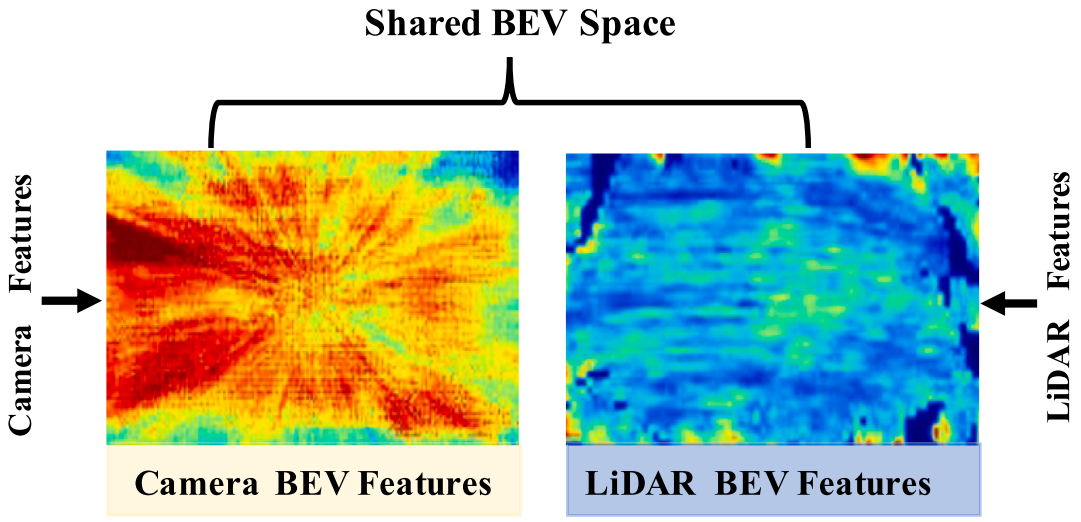


Fig. 1. Illustration of different modalities BEV features. Although both LiDAR and camera BEV features are presented in the shared BEV space, they may still be semantically misaligned due to the significant modality gap. (Blue color means small values and red means large).

and sensing noise. To maximize the advantages of both modalities, recent advancements in camera-LiDAR BEV feature fusion have gained traction, effectively leveraging the semantic richness of camera data alongside the precise geometric information from LiDAR.

Recently, BEV-level fusion methods have gained significant attention for their ability to encode raw inputs from camera and LiDAR sensors into features within the same BEV space using two independent streams. These methods are popular because they harmonize information from different modalities while maintaining spatial consistency. However, as illustrated in Fig. 1, LiDAR and camera BEV features can still exhibit semantic misalignment within the shared BEV space due to the substantial modality gap. Furthermore, existing BEV-level fusion approaches often overlook modality interaction, relying on simple element-wise operations to combine modalities, such as summation [18], weighted averaging [19], or concatenation [1,11]. These naive fusion strategies fail to effectively address modality misalignment and do not mitigate information loss during the fusion process. Addressing these challenges is the motivation behind our work.

To effectively mitigate modality misalignment and information loss, a multi-modal fusion method should incorporate the following characteristics. First, it should enable interaction and integration across multiple modalities. Modality interaction involves enhancing features from one modality using information from another, thereby reducing misalignment. In contrast, modality integration fuses the well-aligned features from different modalities to produce the final output. Second, the method should employ a variety of operations, such as attention for global information exchange, convolution for effective local information aggregation, and weighting across both spatial and channel domains. This combination allows for the accumulation of each operation's strengths, leading to high-quality fusion. Currently, existing approaches only incorporate some of these elements, resulting in sub-optimal fusion performance. To address these issues, we propose a novel multi-modal BEV feature fusion method for map construction, named *MapFusion*, which consists of the CIT and DDF modules to include both modality interaction and integration. To tackle the semantic misalignment between camera and LiDAR BEV features, we propose the new Cross-modal Interaction Transform (CIT) module, which facilitates interaction between the two BEV feature spaces and enhances feature representation using a self-attention mechanism. Specifically, we utilize a correlation matrix to weight each position in the input multi-modal BEV features. This allows the CIT module to perform simultaneous intra-modality and inter-modality fusion across spatial locations, effectively capturing complementary information across different BEV modalities and mitigating modality misalignment. To further refine the feature fusion from different modalities, we propose an effective Dual Dynamic Fusion (DDF) module to adaptively select valuable information from different modalities in a soft manner. In summary, CIT acts as modality interaction, providing flexibility by allowing fusion across both spatial locations and modalities, while DDF refines and fuses the CIT results by concentrating on modality-specific information. Both modules are indispensable for achieving optimal performance. Importantly, the core components of MapFusion, *i.e.*, CIT module and DDF module, are simple yet effective plug-and-play techniques compatible with existing pipelines for various map tasks. Extensive experiments on several benchmarks demonstrate the superiority of our method.

Our main contributions are summarized as follows:

- To address the Bird's-Eye View (BEV) feature fusion challenge in the multi-modal map construction task, we introduce *MapFusion*, a novel method that leverages complementary information from BEV features across different modalities with both modality interaction and integration.
- To solve the semantic misalignment problem between camera and LiDAR BEV features, we propose the Cross-modal Interaction Transform (CIT) module, facilitating interaction between the two BEV feature spaces and enhancing feature representation through a self-attention mechanism.
- For better feature fusion, we propose an effective Dual Dynamic Fusion (DDF) module to adaptively select valuable information from different modalities.
- Compared with the state-of-the-art methods, MapFusion achieves 3.6% and 6.2% absolute improvements on the HD map construction and BEV map segmentation tasks on the nuScenes dataset, respectively, demonstrating the superiority of our approach.

The rest of this paper is organized as follows. We briefly review related works in Section 2. In Section 3, we introduce our proposed method. We then present a variety of experimental results and analyses in Section 4. Finally, Section 5 concludes this paper. This paper is an extension of our preliminary work [20] published on ICRA 2024. The main differences between this paper and the conference version are: **(1)** ***General Algorithm for BEV-based Multi-Modal Map Construction.*** More BEV-based Multi-Modal Map Construction task is realized to validate that our MapFusion, namely the CIT and DDF modules, are effective plug-and-play techniques compatible with existing pipelines for various map tasks. In the conference version we only experimented with vectorized HD map construction task, and in this paper we further include BEV map segmentation task as the new evidence of the versatility and effectiveness of our method. **(2)** ***Enhanced Insights into***

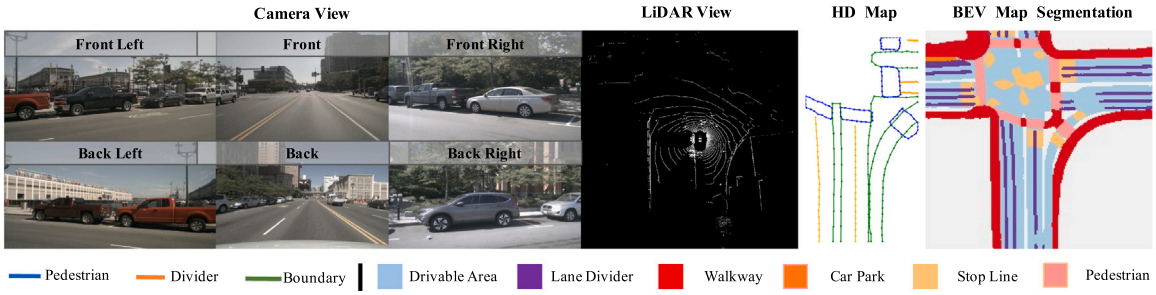


Fig. 2. Illustration of different map construction tasks (HD map construction and BEV map segmentation).

the CIT Module. We provide an internal diagram (see Fig. 4) and equation (see Eq. (5)) to clarify the theoretical foundations of the CIT module, enhancing understanding of its mechanisms. The main idea behind our CIT module is to leverage the self-attention mechanism to learn the binary relationships between camera and LiDAR modalities. Specifically, we utilize a correlation matrix to weight each position of the input feature maps, formulated as Eq. (5), where $\alpha_{i,j}$ represents the correlation between the ith and jth positions on the feature maps. This leads to the inference of four matrix blocks when calculating the correlation matrix α: two intra-modality correlation matrix blocks (for Camera and LiDAR) and two inter-modality correlation matrix blocks, as illustrated in Fig. 4. Consequently, the CIT module can adaptively perform simultaneous intra-modality and inter-modality information fusion, comprehensively capturing complementary information between BEV features of different modalities. **(3)** ***Extensive Ablation Studies and Analysis.*** We conduct additional ablation studies to validate the effectiveness of each proposed component across two BEV-based multi-modal map construction tasks. These studies include: the contributions of CIT and DDF (See Tables 5 and 6), variations of different fusion methods (See Tables 7 and 8), compatibility with other HD map construction methods (See Table 9), and an analysis of the accuracy-computation trade-off using our proposed CIT module and different fusion strategies (See Fig. 7). Based on these ablation experiments, we also conduct a deeper analysis of the working mechanism of our method. **(4)** ***More Visualization Results.*** We include additional visualization results to further illustrate our findings. Fig. 8 shows the visualization results of the t-SNE and the feature maps before and after the CIT module, demonstrating the CIT module's ability to mitigate the misalignment between different modalities. Fig. 9 illustrates the feature maps before and after the CIT module, which integrates various modes of BEV features into a unified space. In addition, we present the qualitative results of the CIT and DDF modules for the BEV map segmentation and HD map tasks in Figs. 6 and 10, respectively.

2. Related work

Our work is highly related to map construction task (See Fig. 2) and multi-sensor fusion methods, which will be discussed thoroughly in the following.

2.1. Map construction task

HD map construction. HD map construction is a critical and extensively researched area in autonomous driving. Based on input sensor modalities, HD map construction models can be categorized into camera-based [21–25], LiDAR-based [14,15] and camera-LiDAR fusion [1–3,26,27] models. Camera-only methods [21–25] have increasingly adopted the Bird's-eye view (BEV) representation as an ideal feature space for multi-view perception, owing to its remarkable ability to mitigate scale ambiguity and occlusion challenges. Various techniques have been proposed to project perspective view (PV) features onto the BEV space by leveraging geometric priors, such as LSS [28], Deformable Attention [29], and GKT [30]. Nevertheless, camera-only methods lack explicit depth information, which forces them to rely on higher resolution images or larger backbone models to achieve enhanced accuracy [29,31–36]. In contrast, LiDAR-only approaches [14,15] benefit from the accurate 3D geometric information provided by LiDAR input. However, they face challenges related to data sparsity and sensing noise.

Recently, camera-LiDAR fusion methods [1–3] leverage the semantic richness of camera data and the geometric information from LiDAR in a collaborative manner. BEV-level fusion, which uses two independent streams to encode raw inputs from camera and LiDAR sensors into features within the same BEV space, has gained significant attention [11,19]. This approach incorporates complementary modality features, outperforming uni-modal input approaches. Existing HD map construction multi-sensor fusion methods—HDMapNet [1], VectorMapNet [2], and MapTR [3]— utilize straightforward channel concatenation and convolution for multi-modal feature fusion. However, these methods overlook modality interaction and employ very simple fusion strategies, leading to issues of misalignment and information loss.

BEV map segmentation. Semantic map construction methods [8–11] take map construction as a BEV semantic segmentation task, assigning semantic labels to each pixel in the BEV plane. Building on Perspective View (PV) segmentation [4,37], early approaches utilize homography transformations to convert camera images into bird's-eye view (BEV) representations, followed by the estimation of segmentation maps [38–41]. However, homography transformation introduces strong artifacts, and BEV-based methods [6,11,12], i.e. performing segmentation directly on BEV plane, have received extensive attention. CVT [6] employs a learned map embedding and an attention mechanism between map queries and camera features. Furthermore, BEVFusion [11], BEVerse [39] and M^{2}BEV [12] explore multi-task learning with 3D object detection. However, these approaches lack explicit utilization of depth information, resulting in unsatisfactory performance.

Existing fusion methods [16,17] primarily focus on object-centric and geometry-oriented approaches. For instance, PointPainting [16] enhances only the foreground LiDAR points, while MVP [17] concentrates solely on densifying foreground 3D objects. Both methods also assume that LiDAR is the more effective modality for sensor fusion, which may not be valid for map construction tasks [11]. Additionally, X-Align [18] employs an integration method that combines the features of the two modalities before applying attention, neglecting modality interactions and relying on overly simplistic fusion strategies. In summary, these methods utilize basic feature concatenation to merge multi-modal features, necessitating the network to implicitly reconcile information from misaligned features.

2.2. Multi-sensor fusion

Multi-sensor fusion has garnered significant attention in the field of autonomous driving. Existing approaches can be broadly categorized into three types: point-level fusion, feature-level fusion, and BEV-level fusion. Point-level fusion methods [16,17,42–44] typically

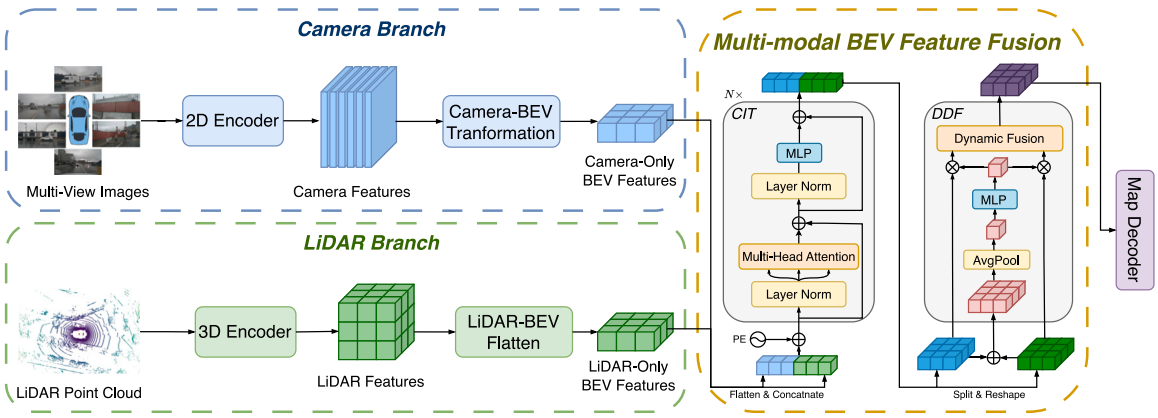


Fig. 3. An overview of MapFusion framework. First, we extract features from multi-modal inputs and convert them into a shared bird's-eye view (BEV) space efficiently using view transformations. To fuse the BEV features from different modalities, we first propose Cross-modal Interaction Transform (CIT) module to enhance one modality from another modality by self-attention mechanism. Afterwards, we propose a Dual Dynamic Fusion (DDF) module to automatically select valuable information from different modalities for better feature fusion. Finally, the fused multi-modal BEV features are fed into a shared decoder and prediction heads for map construction tasks.

project image semantic features onto foreground LiDAR points, enabling LiDAR-based detection on the enhanced point cloud. While effective for 3D object detection tasks, these methods are less suitable for semantically driven tasks such as BEV map segmentation [6,11,12,19] and HD map construction [1–3]. This limitation stems from the lossy projection of camera features to LiDAR, where only about 5% of camera features align with points from a typical 32-beam LiDAR scanner, resulting in significant information loss. Feature-level fusion methods [45,46] first project LiDAR points into a feature space or generate proposals, query the corresponding camera features, and then concatenate them back into the feature space. However, both point-level and feature-level fusion approaches encounter generalization challenges. Specifically, point-level fusion is not easily extendable to other sensor modalities, while feature-level fusion struggles with generalization across different tasks.

Recently, multi-modal feature fusion in a unified BEV space has gained considerable attention [1–3,11,19]. BEV-level fusion employs two independent streams to encode raw inputs from camera and LiDAR sensors into features within the same BEV space. This approach offers a straightforward yet effective means to integrate BEV-level features from both streams, facilitating their use in various downstream tasks. However, existing BEV-level fusion methods often overlook modality interactions, relying on element-wise operations (such as summation or mean) or simple concatenation. This can lead to issues of misalignment and information loss. In this paper, we propose a simple and effective camera-LiDAR BEV feature fusion method that simultaneously integrates complementary information from different modalities, specifically targeting multi-modal map construction tasks.

Comparison with Existing Works. This work differs from prior literature in *three* key aspects. Firstly, we focus on the BEV-based multi-modal map construction task, distinct from other BEV perception tasks [42,43,47], as it aims at predicting map elements, such as pedestrian crossing, lane divider, road boundaries, etc. In fact, the map construction task is a semantic-oriented task, which pays more attention to the semantic information in the image. Therefore, the performance of directly using the fusion method on the 3D object detection task to the map task is not satisfactory. Secondly, to solve the semantic misalignment problem between Camera and LiDAR BEV features, we propose Cross-modal Interaction Transform (CIT) module to enable the two BEV feature spaces to interact with each other and enhance feature representation through a self-attention mechanism. Additionally, to further fuse features from different modalities, we propose an effective Dual Dynamic Fusion (DDF) module to adaptively select valuable information from different modalities. To the best of our knowledge, *MapFusion* is the first to explore the effectiveness of interactive modules on multi-modal map construction tasks. Last but not least, the core components of *MapFusion*, *i.e.*, CIT module and DDF module, are simple yet effective plug-and-play techniques compatible with existing pipelines for various map tasks, such as HD map and semantic map construction.

3. Methodology

We propose a novel multi-modal BEV map construction approach called *MapFusion*, which is a simple yet effective plug-and-play technique compatible with existing pipelines for various map construction tasks. The overview framework of *MapFusion* is shown in Fig. 3. Given different sensory inputs, we first apply modality-specific encoders to extract their features. These multi-modal features are then transformed into a unified BEV representation that preserves both geometric and semantic information. Then, we propose Cross-modal Interaction Transform (CIT) module to make these two BEV feature spaces exchange knowledge with each other to enhance the feature representation by the self-attention mechanism. Additionally, we introduce a novel Dual Dynamic Fusion (DDF) module to automatically select valuable information from different modalities, which can take full advantage of the inherent complementary information between different modalities. Finally, the fused multi-modal BEV features are fed into decoder and prediction heads for map construction tasks.

3.1. Preliminaries

For notation clarity, we first introduce some symbols and definitions used throughout this paper. Our goal is to design a novel framework taking multi-modal sensor data χ as input and predicting map elements in BEV space, and the types of the map elements (supported types are road boundary, lane divider, and pedestrian crossing, etc.). Formally, assume that we have a set of inputs, $\chi = \{Camera, LiDAR\}$, containing multi-view RGB camera images in perspective view, $Camera \in \mathbb{R}^{N^{\text{cam}} \times H^{\text{cam}} \times W^{\text{cam}} \times 3}$, N^{cam}, H^{cam}, W^{cam} denote number of cameras, image height, and image width, respectively, as well as a LiDAR point cloud, $LiDAR \in \mathbb{R}^{P \times 5}$, with number of points P. Each point consists of its 3-dimensional coordinates, reflectivity, and beam index. The detailed architectural designs are described as follows.

3.2. Map encoder

We apply modality-specific encoders to extract their features and transform multi-modal features into a unified BEV representation that preserves both geometric and semantic information. Note that our approach is compatible with other Map Encoders that can also be employed to generate camera-only and LiDAR-only BEV features.

Camera to BEV. We extract BEV features from multi-view RGB images with the BEV feature extractor. It consists of a backbone [31,48] to extract multi-scale 2D features from each perspective view, an FPN [49] to refine and fuse multi-scale features into single-scale features, and a 2D-to-BEV feature transformation module [4,30] to map 2D features into BEV features. The camera BEV features can be denoted as $\mathbf{F}^{\text{BEV}}_{\text{Camera}} \in \mathbb{R}^{H\times W\times C}$, where H, W, C refer to the spatial height, spatial width, and the number of channels of BEV feature maps, respectively.

LiDAR to BEV. For the LiDAR points, we follow SECOND [50] in using voxelization and a sparse LiDAR encoder. The LiDAR features are projected to BEV space using a flattening operation as in [11], to obtain the unified LiDAR BEV representation $\mathbf{F}^{\text{BEV}}_{\text{LiDAR}} \in \mathbb{R}^{H\times W\times C}$.

3.3. Cross-modal Interaction Transform (CIT)

Existing methods directly convert all sensory features to the shared BEV representation, and then fuse them via arithmetic or splicing operations to obtain multi-modal BEV features. However, despite being in the same BEV space, LiDAR BEV features and camera BEV features can still be semantically misaligned due to the significant modality gap, leading to a misalignment problem. To address this issue, we propose a new and powerful Cross-Modal Interaction Transformer (CIT) module to enhance one modality from another modality by the self-attention mechanism. Next, we describe in detail our proposed CIT module.

Concatenation Interaction Transformer. First, given the BEV features from both camera ($\mathbf{F}^{\text{BEV}}_{\text{Camera}} \in \mathbb{R}^{H\times W\times C}$) and LiDAR ($\mathbf{F}^{\text{BEV}}_{\text{LiDAR}} \in \mathbb{R}^{H\times W\times C}$) sensors, the BEV tokens $\mathbf{T}^{\text{BEV}}_{\text{Camera}} \in \mathbb{R}^{HW\times C}$ and $\mathbf{T}^{\text{BEV}}_{\text{LiDAR}} \in \mathbb{R}^{HW\times C}$ are obtained by flattening each BEV feature and permuting the order of the matrices. Second, we concatenate the tokens of each modality and add a learnable positional embedding, which is a trainable parameter of dimension $2HW \times C$, to get the input BEV tokens $\mathbf{T}^{\text{in}} \in \mathbb{R}^{2HW\times C}$ of the Transformer [51]. The positional embedding enables the model to differentiate spatial information between different tokens at training time. Third, the input token $\mathbf{T}^{\text{in}}$ uses linear projections for computing a set of queries, keys and values ($\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$),

$$\mathbf{Q} = \mathbf{T}^{\text{in}}\mathbf{W}^{\text{Q}}, \mathbf{K} = \mathbf{T}^{\text{in}}\mathbf{W}^{\text{K}}, \mathbf{V} = \mathbf{T}^{\text{in}}\mathbf{W}^{\text{V}}, \tag{1}$$

where $\mathbf{W}^{\text{Q}} \in \mathbb{R}^{C\times D_{\text{Q}}}$, $\mathbf{W}^{\text{K}} \in \mathbb{R}^{C\times D_{\text{K}}}$ and $\mathbf{W}^{\text{V}} \in \mathbb{R}^{C\times D_{\text{V}}}$ are weight matrices. Moreover, D_{Q}, D_{K} and D_{V} are equal in our Transformer, i.e., $D_{\text{Q}} = D_{\text{K}} = D_{\text{V}} = C$. Fourth, the self-attention layer uses the scaled dot products between $\mathbf{Q}$ and $\mathbf{K}$ to compute the attention weights and then multiply by the values to infer the refined output $\mathbf{Z}$,

$$\mathbf{Z} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_{\text{k}}}}\right)\mathbf{V}, \tag{2}$$

where $\frac{1}{\sqrt{D_{\text{k}}}}$ is a scaling factor for preventing the softmax function from falling into a region with extremely small gradients when the magnitude of dot products grows large. To encapsulate multiple complex relationships from different representation subspaces at different positions, the multi-head attention mechanism is adopted,

$$\begin{aligned}\hat{\mathbf{Z}} &= \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\mathbf{Z}_1, \ldots, \mathbf{Z}_h)\mathbf{W}^{\text{O}},\\ \mathbf{Z}_i &= \text{Attention}(\mathbf{Q}\mathbf{W}^{\text{Q}}_i, \mathbf{K}\mathbf{W}^{\text{K}}_i, \mathbf{V}\mathbf{W}^{\text{V}}_i), i \in \{1, \ldots, h\}.\end{aligned} \tag{3}$$

The subscript h denotes the number of heads, and $\mathbf{W}^{\text{O}} \in \mathbb{R}^{h\cdot C\times C}$ denotes the projected matrix of $\text{Concat}(\mathbf{Z}_1, \ldots, \mathbf{Z}_h)$. Finally, the transformer uses a non-linear transformation to calculate the output features, $\mathbf{T}^{\text{out}}$ which are of the same shape as the input features $\mathbf{T}^{\text{in}}$,

$$\mathbf{T}^{\text{out}} = \text{MLP}(\hat{\mathbf{Z}}) + \mathbf{T}^{\text{in}}. \tag{4}$$

The output $\mathbf{T}^{\text{out}}$ are converted into $\hat{\mathbf{F}}^{\text{BEV}}_{\text{Camera}}$ and $\hat{\mathbf{F}}^{\text{BEV}}_{\text{LiDAR}}$ for further feature fusion Eq. (5) is given in Box I.

Remarks: The main idea behind our CIT module is to leverage the self-attention mechanism to learn the binary relationships between Camera and LiDAR modalities. More specifically, we utilize a correlation matrix to weight each position in the input feature maps. This can be formulated as shown in Eq. (5). In this formula, $\alpha_{i,j}$ represents the correlation between the ith position and the jth position on the feature maps. According to Eq. (5), four matrix blocks can be naturally inferred when calculating the correlation matrix α. Two of these blocks represent intra-modality correlation matrices (for Camera and LiDAR), while the other two represent inter-modality correlation matrices, as illustrated in Fig. 4. Thus, we utilize the correlation matrix to weight each position of the input multi-modal BEV features. The CIT module can then adaptively perform simultaneous intra-modality and inter-modality information fusion, robustly capturing the complementary information between BEV features of different modalities.

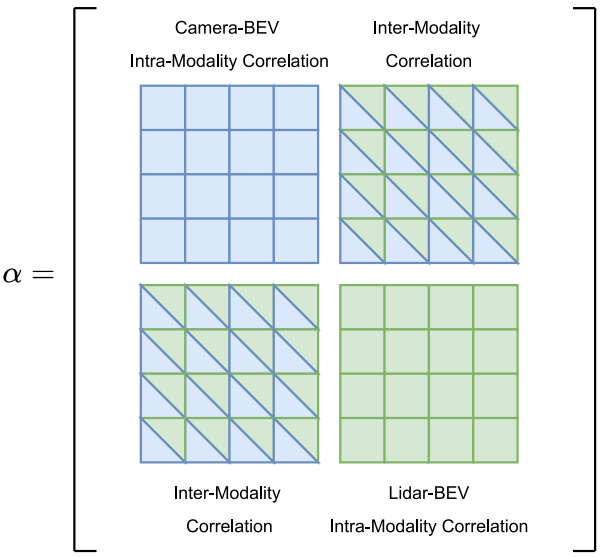


Fig. 4. Illustration of the Correlation Matrix α.

3.4. Dual Dynamic Fusion (DDF)

Despite the effectiveness of the cross-modal interaction transform module, we argue that how to design an effective cross-modal fusion strategy to adaptively select valuable information from different modalities for better feature fusion is still very important. Recently, multi-modal BEV feature fusion methods [11,19] have received much attention. It is a common approach to utilize concatenation followed by convolution to combine features from multi-modal BEV feature inputs, $\hat{\mathbf{F}}^{\text{BEV}}_{\text{Camera}}$ and $\hat{\mathbf{F}}^{\text{BEV}}_{\text{LiDAR}}$, resulting in the aggregated features $\mathbf{F}_{\text{fused}}$, as shown in Fig. 5(a) Conv Fusion. Another common method is to use CNN to convolve the BEV features of different modalities separately, and then add the convolutional features, as shown in Fig. 5(b) Add Fusion. As Fig. 5(c) illustrates, the input of the Dynamic Fusion (DF) module is the Conv Fusion output features, and then they are fused with learnable static weights, inspired by Squeeze-and-Excitation mechanism [52]. To effectively select valuable information from different modalities, we propose a Dual Dynamic Fusion (DDF) module for better feature fusion and maximum performance gain. Next, we describe in detail our proposed fusion designs.

Dual Dynamic Fusion. As shown in Fig. 5(d), our Dual Dynamic Fusion (DDF) module takes two sets of features from the camera BEV features and LiDAR BEV features as input. In order to generate meaningful attention weights that can effectively select informative features from both inputs, we first sum the features from both branches before performing the squeeze and excitation operations that generate the attention weights. We can formulate this process as:

$$\mathbf{w} = \sigma\left(\gamma\left(\text{AvgPool}\left(\hat{\mathbf{F}}^{\text{BEV}}_{\text{Camera}} + \hat{\mathbf{F}}^{\text{BEV}}_{\text{LiDAR}}\right)\right)\right), \tag{6}$$

$$\boldsymbol{\alpha} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{D_K}}\right) = \begin{array}{c} \\ L_1^{BEV} \\ \vdots \\ L_{HW}^{BEV} \\ L_{HW+1}^{BEV} \\ \vdots \\ L_{2HW}^{BEV} \end{array} \begin{array}{c} \begin{array}{cccc} C_1^{BEV} & & \cdots & C_{HW}^{BEV} \end{array} \\ \left(\begin{array}{cccc} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,HW} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,HW} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{HW,1} & \alpha_{HW,2} & \cdots & \alpha_{HW,HW} \\ \alpha_{HW+1,1} & \alpha_{HW+1,2} & \cdots & \alpha_{HW+1,HW} \\ \alpha_{HW+2,1} & \alpha_{HW+2,2} & \cdots & \alpha_{HW+2,HW} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{2HW,1} & \alpha_{2HW,2} & \cdots & \alpha_{2HW,HW} \end{array} \right. \end{array} \begin{array}{c} \begin{array}{cccc} C_{HW+1}^{BEV} & & \cdots & C_{2HW}^{BEV} \end{array} \\ \left.\begin{array}{cccc} \alpha_{1,HW+1} & \alpha_{1,HW+2} & \cdots & \alpha_{1,2HW} \\ \alpha_{2,HW+1} & \alpha_{2,HW+2} & \cdots & \alpha_{2,2HW} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{HW,HW+1} & \alpha_{HW,HW+2} & \cdots & \alpha_{HW,2HW} \\ \alpha_{HW+1,HW+1} & \alpha_{HW+1,HW+2} & \cdots & \alpha_{HW+1,2HW} \\ \alpha_{HW+2,HW+1} & \alpha_{HW+2,HW+2} & \cdots & \alpha_{HW+2,2HW} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{2HW,HW+1} & \alpha_{2HW,HW+2} & \cdots & \alpha_{2HW,2HW} \end{array} \right) \end{array}, \tag{5}$$

Box I.

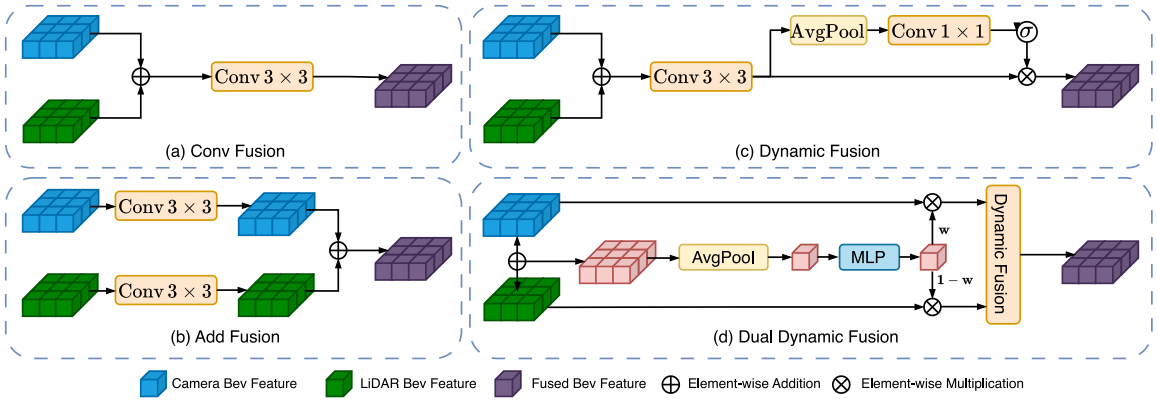


Fig. 5. Three existing fusion strategies and our proposed Dual Dynamic Fusion (DDF) strategy.

where σ and γ represent the sigmoid function and linear layers respectively, AvgPool is the global average pooling operation, and $\mathbf{w}$ denotes the attention weights. We then multiply $\mathbf{w}$ and $\mathbf{1} - \mathbf{w}$ to both input features before the summation so that the fusion process essentially acts as a self-gating mechanism to adaptively select useful information from different BEV features:

$$\mathbf{F}_{\text{fused}} = \text{Adaptive}\left(\text{Conv}_{3\times3}\left(\left[\mathbf{w}\cdot\hat{\mathbf{F}}^{\text{BEV}}_{\text{Camera}}, (\mathbf{1}-\mathbf{w})\cdot\hat{\mathbf{F}}^{\text{BEV}}_{\text{LiDAR}}\right]\right)\right), \tag{7}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the channel dimension. $\cdot$ is element-wise multiplication. $\text{Conv}_{3\times3}$ fuses the channel and spatial information with a 3×3 convolution layer to reduce the channel dimension of concatenated feature to C. With input feature $\hat{\mathbf{F}} \in \mathbb{R}^{H\times W\times C}$, the Adaptive operation is formulated as:

$$\text{Adaptive}(\hat{\mathbf{F}}) = \sigma\left(\mathbf{W}\,\text{AvgPool}(\hat{\mathbf{F}})\right)\cdot\hat{\mathbf{F}}, \tag{8}$$

where $\mathbf{W}$ denotes linear transform matrix (e.g., 1×1 convolution) and σ denotes sigmoid function. Therefore, the DDF module can adaptively select valuable information from two modalities for better feature fusion. The output fused feature $\mathbf{F}_{\text{fused}}$ will be used for map construction task, with the decoder and prediction heads.

Remarks: DF only performs channel-wise fusion, while DDF first conducts spatial fusion and then channel-wise fusion. DDF enhances DF by incorporating global average pooling, utilizing global weights to reduce information loss. In DDF module, the AvgPool in Eq. (6) is performed in the spatial domain with an input dimension of $W \times H \times C$ and an output of C; The AvgPool in Eq. (8) is performed in the channel domain with an input of $W \times H \times C$ and an output of $W \times H$.

3.5. Map-task heads

We apply specific heads for different map tasks to the fused BEV features. We show two examples: HD map construction and BEV map segmentation.

HD map construction head. HD map constructors formulate this task as predicting a collection of vectorized static map elements in bird's eye view (BEV), *i.e.*, pedestrian crossings, lane dividers, road boundaries. We follow MapTR [3] to train the map head with the classification loss [53], the point2point loss [54], and the edge direction loss [3].

BEV map segmentation head. Different map categories may overlap (e.g., crosswalk is a subset of drivable space). Therefore, we formulate this problem as multiple binary semantic segmentation, one for each class. We follow BEVFusion [11] to train the segmentation head with the standard focal loss [53].

4. Experiments

4.1. Dataset

NuScenes Datasets. We evaluate our method on the widely-used challenging nuScenes [55] dataset following the standard settings of previous methods [3,11]. The nuScenes dataset contains 1000 sequences of recordings collected by autonomous driving cars. Each sample is annotated at 2 Hz and contains 6 camera images covering 360° horizontal FOV of the ego-vehicle. For the HD map construction task, we following MapTR [3] and three kinds of map elements are chosen for fair evaluation — pedestrian crossing, lane divider, and road boundary. Moreover, for the BEV map segmentation task, we following BEVFusion [11], we predict six semantic classes: drivable lanes, pedestrian crossings, walkways, stop lines, carparks, and lane dividers.

Argoverse2 Dataset. There are 1000 logs in the Argoverse2 dataset [56]. Each log contains 15 s of 20 Hz RGB images from 7 cameras, 10 Hz LiDAR sweeps, and a 3D vectorized map. The train, validation, and test sets contain 700, 150, and 150 logs, respectively. For both HD map construction and BEV map segmentation tasks, we select three map elements for fair evaluation: pedestrian crossing, lane divider, and road boundary.

Table 1
Comparisons with state-of-the-art methods on nuScenes val set for the HD map construction task. We compare with existing methods from literature, where the numbers are taken from MapTR [3]. We also provide information on the backbones, epochs and input modalities in the table. Our proposed MapFusion outperforms all existing approaches in both single-class APs and the overall mAP by a significant margin.

Method	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet [1]	C	Efficient-B0	30	14.4	21.7	33.0	23.0
HDMapNet [1]	L	PointPillars	30	10.4	24.1	37.9	24.1
HDMapNet [1]	C & L	Efficient-B0 & PointPillars	30	16.3	29.6	46.7	31.0
VectorMapNet [2]	C	ResNet-50	110	36.1	47.3	39.3	40.9
VectorMapNet [2]	L	PointPillars	110	25.7	37.6	38.6	34.0
VectorMapNet [2]	C & L	ResNet-50 & PointPillars	110	37.6	50.5	47.5	45.2
MapTR [3]	C	ResNet-50	24	46.3	51.5	53.1	50.3
MapTR [3]	L	SECOND	24	48.5	53.7	64.7	55.6
MapTR [3]	C & L	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
MapFusion (Ours)	C & L	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	**66.1**$_{+3.6}$

4.2. Evaluation metrics

HD map construction task. We adopt the evaluation metrics consistent with previous works [1–3], where average precision (AP) is used to evaluate the map construction quality and Chamfer distance D_{Chamfer} determines the matching between predictions and ground truth. We calculate the AP_{τ} under several D_{Chamfer} thresholds ($\tau \in T = \{0.5\text{ m}, 1.0\text{ m}, 1.5\text{ m}\}$), and then average across all thresholds as the final mean AP (mAP) metric,

$$mAP = \frac{1}{|T|}\sum_{\tau \in T} \text{AP}_{\tau}. \tag{9}$$

The perception ranges are $[-15.0\text{ m}, 15.0\text{ m}]/[-30.0\text{ m}, 30.0\text{ m}]$ for X/Y-axes.

BEV map segmentation task. For the BEV map segmentation task, our primary evaluation metric is the mean Intersection over Union (mIoU). Due to potential overlaps between classes, we apply binary segmentation separately to each class and choose the highest IoU over different thresholds. We then average these values over all semantic classes to produce the mIoU. This evaluation protocol was proposed in BEVFusion [11].

4.3. Experimental setting

MapFusion is trained with 4 NVIDIA RTX A6000 GPUs. For the HD map construction task, we build upon MapTR [3] as the baseline. Specifically, we adopt ResNet50 [57] and SECOND [50] as the backbone and employ GKT [30] as the default 2D-to-BEV module. Training losses include classification loss, point2point loss, and edge direction loss. with weights of 2.0, 5.0, and 0.005, respectively. The model is trained for 24 and 6 epochs on the nuScenes and Argoverse2 datasets respectively. All the data pre-processing steps for both datasets follow MapTR [3]. We set the mini-batch size to 16, and use a step-decayed learning rate with an initial value of $4e^{-3}$. For the BEV map segmentation task, we use BEVFusion [11] as our baseline and train our networks within the mmdetection3d framework [58]. Specifically, we adopt Swin-T [31] and VoxelNet [50] as the backbone, and utilize LSS [4] as the default 2D-to-BEV module. The model is trained for 20 and 6 epochs on the nuScenes and Argoverse2 datasets, respectively. The baseline is trained using the hyperparameters reported in [11], following a learning schedule of 20 epochs with a cyclic learning rate, starting for $1e^{-4}$ and performing a single cycle with target ratios 10, $1e^{-4}$ and a step of 0.4 For the CIT module described in Section 3.3 of the paper, we added this module before the fuser operation in the baseline model. To implement the cross-modal interaction, we first obtain BEV tokens by flattening each BEV feature and permuting the order of the matrix. Then, we concatenate the tokens of each modality and add a learnable positional embedding. This step is followed by a multi-head self-attention block as described in [51], containing 8 heads and an embedding dimension of 256. For the DDF module described in Section 3.4 of the paper, we replace the naive convolutional fuser with the DDF module in the baseline model.

4.4. Comparison with the state-of-the-arts

4.4.1. HD map construction task

We compare MapFusion with state-of-the-art HD map construction methods on nuScenes and Argoverse2 datasets. Our proposed MapFusion outperforms all existing approaches in both single-class APs and the overall mAP by a significant margin.

Experimental Settings. We adopt average precision (AP) to evaluate the map construction quality. Chamfer distance $D_{Chamfer}$ is used to determine whether the prediction and GT are matched or not. We calculate the AP_{τ} under several $D_{Chamfer}$ thresholds ($\tau \in T, T = \{0.5, 1.0, 1.5\}$, unit is meter), and then average across all thresholds as the final AP metric. The resolution of source images is 1600×900. During the training phase, we resize the source images using a ratio of 0.5. Moreover, we set the maximum number of map elements in one frame, the number of points in one map element, the size of each BEV grid, and the number of transformer decoder layers to 100, 20, 0.75 m, and 2, respectively. We follow the experimental settings of existing methods from MapTR [3].

Experimental Results. With the same settings and data partition, we compare the proposed MapFusion method with several state-of-the-art methods, i.e., HDMapNet [1], VectorMapNet [2] and MapTR [3]. Tables 1 and 2 show the overall performance of MapFusion and all the baselines on nuScenes and Argoverse2 datasets, respectively. Note that re-implementation is needed because the reference methods do not report results on Argoverse2 data set, which has different input data format from nuScenes. The experimental results reveal a number of interesting points: (1) The performance of multi-modal methods are obviously better than that of single-modal methods, which proves the significance of utilizing complementary cues from camera and LiDAR to improve the HD map construction performance. (2) In the multi-modality setting, the proposed MapFusion approach achieves a 3.6% absolute improvement in mAP over the previous state-of-the-art MapTR [3] on the nuScenes dataset. Similarly, it shows a 4.1% absolute improvement in mAP compared to MapTR [3] on the Argoverse2 dataset. This advantage arises from the limitations of the three compared HD map construction methods—HDMapNet [1], VectorMapNet [2], and MapTR [3]—which rely on straightforward channel concatenation and convolution for multi-modal feature fusion, as shown in Fig. 5(a) (Conv Fusion). These methods neglect modality interaction and employ overly simplistic fusion strategies, resulting in misalignment and information loss.

In a nutshell, MapFusion shows significant superiority over other multi-modal methods, indicating the benefit of cross-modal interaction transform (CIT) module and dual dynamic fusion (DDF) module. This is due to the fact that the CIT module enables the two feature spaces to interact with each other and enhances feature representation through a self-attention mechanism, while the DDF module automatically selects valuable information from different modalities and can make full use of the inherent complementary information between different modalities.

Table 2
Results of the HD map construction task on the Argoverse2 dataset.

Method	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet [1]	C	Efficient-B0	30	13.1	5.70	37.6	18.8
VectorMapNet [2]	C	ResNet-50	110	38.3	36.1	39.2	37.9
MapTR[a] [3]	C	ResNet-50	6	58.7	59.3	60.3	59.4
MapTR[a] [3]	C & L	ResNet-50 & SECOND	6	65.1	61.6	75.1	67.3
MapFusion (Ours)	C & L	ResNet-50 & SECOND	6	**69.4**	**65.8**	**78.9**	**71.4**$_{+4.1}$

[a] Denotes our re-implementation following the setting in the paper.

Table 3
Results of the BEV map segmentation task on the nuScenes dataset. We compare with existing methods from literature, where the numbers are taken from BEVFusion [11]. We also provide information on the backbones and input modalities in the table. MapFusion outperforms the state-of-the-art multi-sensor fusion methods and achieves consistent improvements across different categories. **Note that, we use BEVFusion [11] as the baseline model.**

Method	Modality	Backbone	Drivable	Ped. Cross.	Walkway	Stop line	Carpark	Divider	mIoU
OFT [37]	C	ResNet18	74.0	35.3	45.9	27.5	35.9	33.9	42.1
LSS [4]	C	ResNet18	75.4	38.8	46.3	30.3	39.1	36.5	44.4
CVT [6]	C	EfficientNet-B4	74.3	36.8	39.9	25.8	35.0	29.4	40.2
M^2BEV [12]	C	ResNet101	77.2	✗	✗	✗	✗	40.5	✗
BEVFusion [11]	C	Swin-T	81.7	54.8	58.4	47.4	50.7	46.4	56.6
X-Align [18]	C	Swin-T	82.4	55.6	59.3	49.6	53.8	47.4	58.0
PointPillars [14]	L	VoxelNet	72.0	43.1	53.1	29.7	27.7	37.5	43.8
CenterPoint [15]	L	VoxelNet	75.6	48.4	57.5	36.5	31.7	41.9	48.6
PointPainting [16]	C & L	ResNet-101 & PointPillars	75.9	48.5	57.1	36.9	34.5	41.9	49.1
MVP [17]	C & L	ResNet-101 & VoxelNet	76.1	48.7	57.0	36.9	33.0	42.2	49.0
BEVFusion [11]	C & L	Swin-T & VoxelNet	85.5	60.5	67.6	52.0	57.0	53.7	62.7
X-Align [18]	C & L	Swin-T & VoxelNet	86.8	65.2	70.0	58.3	57.1	58.2	65.7
MapFusion (Ours)	C & L	Swin-T & VoxelNet	**88.9**	**69.6**	**74.0**	**63.0**	**56.5**	**61.5**	**68.9**$_{+6.2}$

Table 4
Results of the BEV map segmentation task on the Argoverse2 dataset.

Method	Modality	Backbone	Drivable	*Ped.Cross.*	Divider	mIoU
BEVFusion[a] [11]	C & L	Swin-T & VoxelNet	78.1	30.7	46.3	51.7
MapFusion (Ours)	C & L	Swin-T & VoxelNet	**83.5**	**37.4**	**53.7**	**58.2**$_{+6.5}$

[a] Denotes our re-implementation following the setting in the paper.

4.4.2. BEV map segmentation task

We further compare MapFusion with state-of-the-art BEV map segmentation models, where MapFusion outperforms the state-of-the-art multi-sensor fusion methods and achieves consistent improvements across different categories.

Experimental Settings. We report the Intersection-over-Union (IoU) on 6 background classes (drivable space, pedestrian crossing, walkway, stop line, car-parking area, and lane divider) on nuScenes dataset and 3 background classes (drivable space, pedestrian crossing and lane divider) on Argoverse2 dataset. The class-averaged mean IoU as our evaluation metric. For each frame, we only perform the evaluation in the $[-50\text{ m}, 50\text{ m}] \times [-50\text{ m}, 50\text{ m}]$ region around the ego car following [11,16,17,19]. In MapFusion model, we use a single model that jointly performs binary segmentation for all classes instead of following the conventional approach to train a separate model for each class. We follow the experimental results of existing methods from BEVFusion [11].

Experimental Results. With the same settings and data partition, we compare the proposed MapFusion method with several state-of-the-art methods, i.e., PointPainting [16], MVP [17], and BEVFusion [11]. Tables 3 and 4 show the overall performance of MapFusion and all the baselines on nuScenes and Argoverse2 datasets, respectively. Similar to HD map construction, we also re-implemented the experimental results for the Argoverse2 dataset.

The experimental results reveal several interesting points: (1) In the single-modality setting, camera-based models perform significantly better than LiDAR-based models. This observation is the exact opposite of results in 3D object detection task [11,19]. The main reason is that the map construction task is a semantic-oriented task, which pays more attention to the semantic information in the image. Therefore, the performance of directly using the fusion method on the 3D object detection task for the map task is not satisfactory. (2) In the multi-modality setting, MapFusion outperforms existing state-of-the-art multi-sensor fusion methods, consistently across various categories. This advantage arises from the limitations of these methods: PointPainting [16] is object-centric, focusing solely on enhancing foreground LiDAR points, while MVP [17] is geometry-oriented, concentrating exclusively on densifying foreground 3D objects—neither effectively segments map components. Furthermore, BEVFusion [11] and X-Align [18] neglect modality interactions and rely on overly simplistic fusion strategies (see Fig. 5(a) Conv Fusion and Fig. 5(c) Dynamic Fusion), resulting in misalignment and information loss. Our proposed MapFusion approach achieves a 6.2% absolute improvement in mean Intersection over Union (mIoU) compared to the previous state-of-the-art BEVFusion [11] on the nuScenes dataset, and a 6.5% absolute improvement on the Argoverse2 dataset. Notably, we re-implemented the BEVFusion method following the original settings outlined in their paper. Overall, MapFusion consistently enhances the performance of existing fusion methods on both the nuScenes and Argoverse2 datasets, demonstrating the effectiveness of our proposed CIT and DDF components.

4.5. Ablation studies

4.5.1. Contribution of each component

To systematically evaluate the effectiveness of each module of our proposed MapFusion, we train the model using different components and show the experimental results of the HD map construction task and BEV map segmentation task in Tables 5 and 6 respectively. In the main ablation study, we design the following model variants: (1) MapFusion (Baseline) : we train the model without the cross-modal interaction transform module and dual dynamic fusion module. (2) MapFusion (w/ DDF) : we train the model with the dual dynamic fusion module. (3) MapFusion (w/ CIT) : we train the model with the cross-modal interaction transform module. (4) MapFusion (full) : we train the model

Table 5
An ablation study of the proposed MapFusion components is performed on the nuScenes dataset HD map construction task. "DDF" and "CIT" respectively denote Dual Dynamic Fusion module and Cross-modal Interaction Transform module. We show the effects of our proposed modules.

DDF	CIT	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
✗	✗	C & L	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
✔	✗	C & L	ResNet-50 & SECOND	24	58.4	64.1	72.5	$65.0_{+2.5}$
✗	✔	C & L	ResNet-50 & SECOND	24	60.2	64.3	72.1	$65.5_{+3.0}$
✔	✔	C & L	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	$\mathbf{66.1}_{+3.6}$

Table 6
An ablation study of the proposed MapFusion components is performed on the nuScenes dataset BEV map segmentation task.

DDF	CIT	Modality	Backbone	Drivable	Ped. Cross.	Walkway	Stop line	Carpark	Divider	mIoU
✗	✗	C & L	ResNet-50 & VoxelNet	85.5	60.5	67.6	52.0	57.0	53.7	62.7
✔	✗	C & L	ResNet-50 & VoxelNet	86.2	62.2	68.9	54.4	56.4	56.0	$64.1_{+1.4}$
✗	✔	C & L	ResNet-50 & VoxelNet	88.8	68.3	73.6	62.6	56.0	60.5	$68.3_{+5.6}$
✔	✔	C & L	ResNet-50 & VoxelNet	**88.9**	**69.6**	**74.0**	**63.0**	**56.5**	**61.5**	$\mathbf{68.9}_{+6.2}$

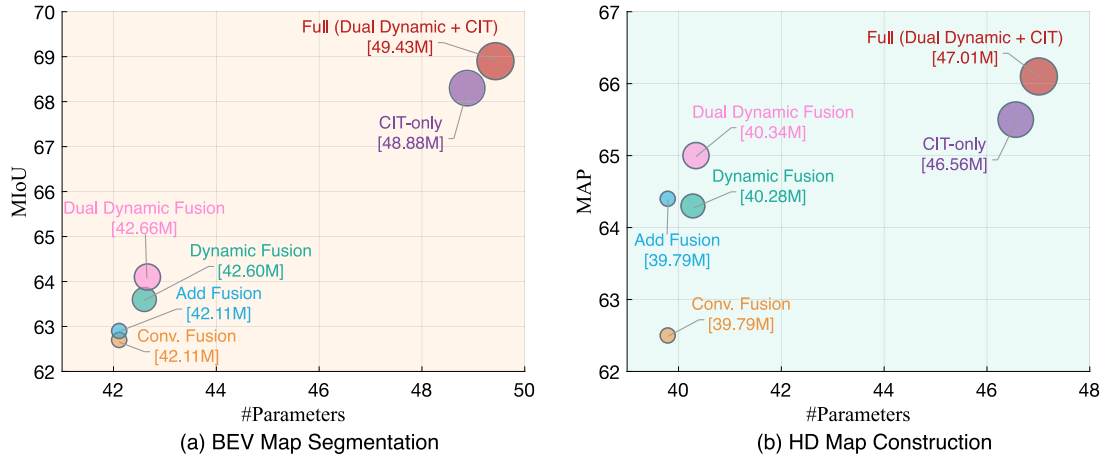


Fig. 6. Qualitative results on BEV map segmentation task. We present a sample scene from nuScenes: (a) six camera inputs, (b) LiDAR scan, (c) ground-truth BEV map segmentation map, (d) baseline BEV segmentation map (BEVFusion [11]), (e) BEV segmentation map of only using CIT module, and (f) BEV segmentation map of MapFusion (full).

with the cross-modal interaction transform module and dual dynamic fusion module.

The experimental results reveal some interesting findings: (1) The results of both MapFusion (w/ DDF) and MapFusion (w/ CIT) are significantly better than the MapFusion (Baseline), verifying the effectiveness of CIT and DDF components for improving multi-modal BEV map construction. Compared with the baseline model, DDF and CIT modules achieve 2.5% and 3.0% absolute improvements respectively on HD map construction task, demonstrating the superiority of our approach. Similarly, compared with the baseline model, DDF and CIT modules achieve 1.4% and 5.6% absolute improvements respectively on BEV map segmentation task. (2) The results of MapFusion (w/ DDF) and MapFusion (w/ CIT) are inferior to the MapFusion (full), verifying the effectiveness of using both CIT and DDF simultaneously. MapFusion (full) achieves 3.6% and 6.2% absolute improvements on the HD map construction and semantic map construction tasks, respectively, demonstrating the superiority of our method.

These experimental results demonstrate that the CIT module enables the camera and LiDAR BEV space to interact with each other to enhance feature representation through the cross-attention mechanism. Moreover, it is verified that the DDF module can automatically select valuable information from different modalities, thereby making full use of the inherent complementary information between different modalities.

4.5.2. Analysis on different fusion methods

To systematically evaluate the effectiveness of the dual dynamic fusion (DDF) method, we train the model using different fusion methods detailed in Section 3.4. Tables 7 and 8 show the experimental results on the HD map construction task and BEV map segmentation task using different fusion methods, respectively. For instance, the proposed DDF method achieves 2.5% absolute improvements compared with Baseline model (Conv. Fusion) method on HD map construction task. Similarly, DDF method achieves 1.4% absolute improvement compared with Baseline model (Conv. Fusion) on BEV map segmentation task. Experimental results show that the DDF module plays a vital role in multi-modal BEV feature fusion and can automatically select valuable information from different modalities for better feature fusion.

4.5.3. Compatibility with other HD map construction methods

We show MapFusion is compatibility with other HD Map Construction methods, i.e., HDMapNet [1], VectorMapNet [2], and MapTR [3]. Besides adding MapFusion, we do not modify their original training settings. For all experiments, we report the result of the nuScenes val set. As shown in Table 9, simply adding MapFusion on top of these strong baselines consistently improves state-of-the-art performance. MapFusion demonstrates a significant accuracy boost (absolute): HDMapNet(+7.6%), VectorMapNet (+5.5%), and MapTR (+3.6%). This shows the versatility of MapFusion as a multi-modal BEV feature fusion method.

4.5.4. Accuracy-computation analysis

In Fig. 7, we report the accuracy-computation trade-off by utilizing our proposed CIT module (See Section 3.3) and different fusion strategies (See Section 3.4). It can be seen that when using the CIT module, we achieve the highest accuracy improvement at a higher computational cost, while the DDF module introduces less additional cost but

Table 7
Performance comparison of different fusion strategies on HD map construction task. Our proposed dual dynamic fusion strategy outperforms all existing approaches by a significant margin.

Method	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
Baseline(Conv. Fusion)	C & L	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
Add fusion	C & L	ResNet-50 & SECOND	24	**61.1**	60.3	71.8	$64.4_{+1.9}$
Dynamic Fusion	C & L	ResNet-50 & SECOND	24	58.4	63.1	71.5	$64.3_{+1.8}$
Dual Dynamic Fusion	C & L	ResNet-50 & SECOND	24	58.4	**64.1**	**72.5**	$\mathbf{65.0}_{+2.5}$

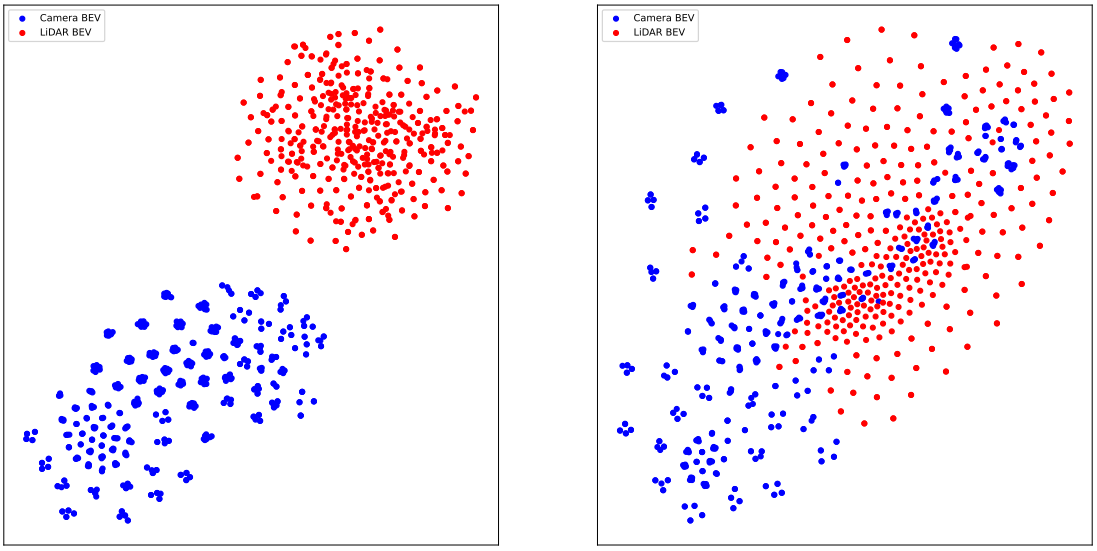
Table 8
Performance comparison of different fusion strategies on BEV map segmentation task.

Method	Modality	Backbone	Drivable	Ped. Cross.	Walkway	Stop line	Carpark	Divider	mIoU
Baseline(Conv. Fusion)	C&L	ResNet-50 & VoxelNet	85.5	60.5	67.6	52.0	57.0	53.7	62.7
Add fusion	C&L	ResNet-50 & VoxelNet	85.4	60.6	67.8	52.3	57.5	53.9	$62.9_{+0.2}$
Dynamic fusion	C&L	ResNet-50 & VoxelNet	86.1	**62.5**	68.7	53.9	54.7	55.6	$63.6_{+0.9}$
Dual dynamic fusion	C&L	ResNet-50 & VoxelNet	**86.2**	62.2	**68.9**	**54.4**	**56.4**	**56.0**	$\mathbf{64.1}_{+1.4}$

Table 9
Compatibility to other HD map construction methods. Adding MapFusion leads to consistent performance boost on nuScenes val set in terms of mAP.

Method	Venue	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet[a] [1]	ICRA 22	C & L	Efficient-B0 & PointPillars	30	13.3	26.9	44.3	28.2
HDMapNet + MapFusion	–	C & L	Efficient-B0 & PointPillars	30	**21.1**	**34.2**	**52.1**	$\mathbf{35.8}_{+7.6}$
VectorMapNet[a] [2]	ICML 23	C & L	ResNet-50 & PointPillars	110	35.8	48.2	45.3	43.1
VectorMapNet + MapFusion	–	C & L	ResNet-50 & PointPillars	110	**41.1**	**53.7**	**50.9**	$\mathbf{48.6}_{+5.5}$
MapTR [3]	ICLR 23	C & L	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
MapTR + MapFusion	–	C & L	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	$\mathbf{66.1}_{+3.6}$

[a] Denotes our re-implementation following the setting in the original papers.



(a) Before CIT Module **(b) After CIT Module**

Fig. 7. Accuracy-Computation Analysis. We report the accuracy-computation trade-off by utilizing our proposed cross-modal interaction transform module and different fusion strategies.

provides less performance gain. It can be seen that all our proposed fusion modules achieve better trade-offs compared with the baselines. Furthermore, we find that the CIT module significantly outperforms existing BEV fusion strategies, which again verifies that the baseline fusion using simple concatenation and convolutions does not provide the suitable capacity for the model to align and aggregate multi-modal features.

4.6. Visualization

t-SNE. We randomly choose 500 samples from the nuScenes validation dataset and show the t-SNE [59] visualizations of (a) Before CIT module and (b) After CIT module in Fig. 8. Red/Blue denotes camera BEV feature/LiDAR BEV feature. As can be seen, Fig. 8(a) Before CIT module shows that blue and red features are clearly separated, indicating that although in the same space, camera BEV features and LiDAR BEV features can still be misaligned to some extent due to the inaccurate depth in the view transformer and the large modality gap. Fig. 8(b) After the CIT module, the BEV features from different modalities are aligned in a shared space, *i.e.*, red and blue dots are close after the CIT module.

Feature map visualizations. In order to visually demonstrate the effectiveness of the CIT module, we visualize the feature map before and after the CIT module in Fig. 9. Before the CIT module, the BEV

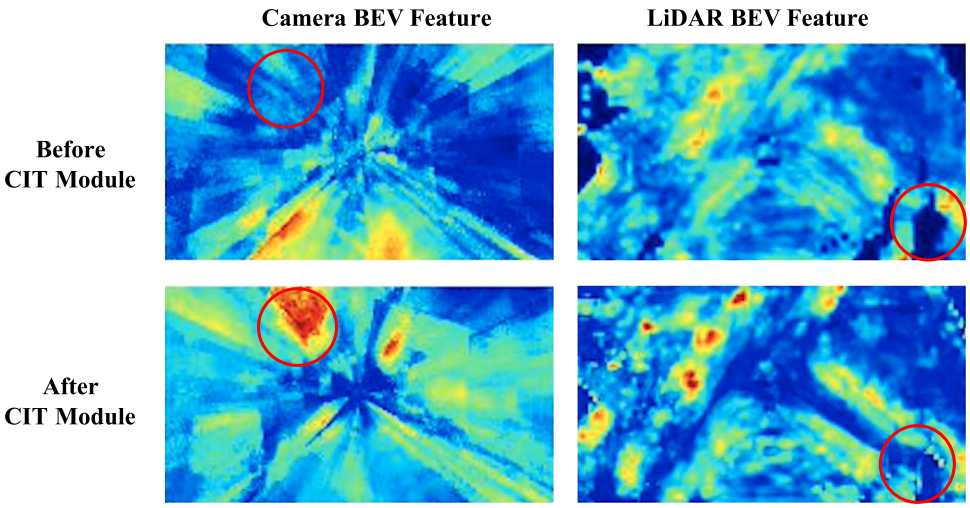


Fig. 8. The t-SNE visualizations of (a) Before CIT module and (b) After CIT module on HD map construction task. Red/Blue denotes camera BEV feature/LiDAR BEV feature. After the CIT module, the BEV features from different modalities are aligned in a shared space, *i.e*, red and blue dots are close after the CIT module.

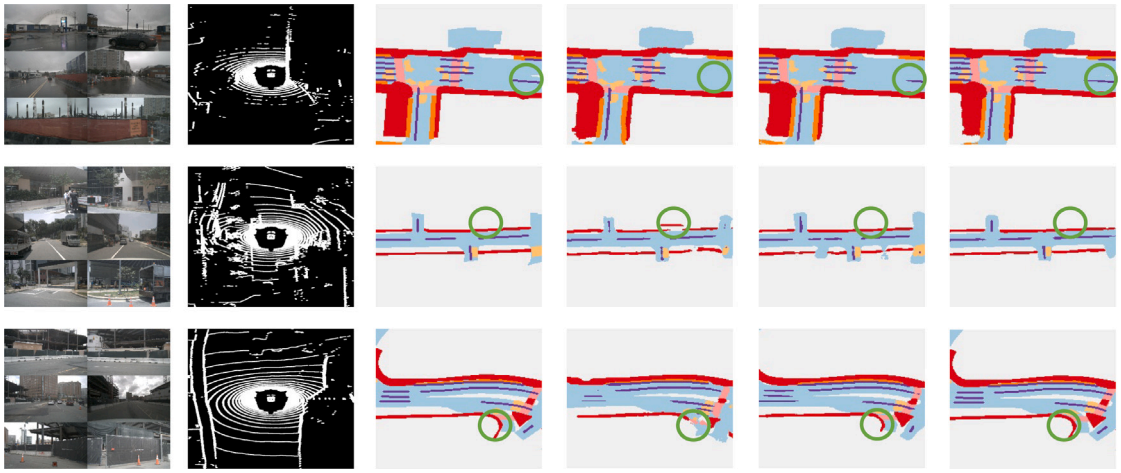


Fig. 9. Visualization of feature maps before and after the CIT module for the HD map construction task.

features of different modalities look quite different. While after the CIT module, they look more similar, verifying mitigated modality misalignment. We can also find that: (1) The camera feature map is enhanced, as shown in the red circles of the left top and left bottom images, making the feature representation more powerful; (2) Missing features in the LiDAR feature map are recovered, as shown in the red circles of the right top and right bottom images. In summary, the CIT module integrates different modes of BEV features into a shared space, thereby enhancing representation learning and overall model performance.

Qualitative Results. In Fig. 6, we present more sample scenes from nuScenes on the BEV map segmentation task. Each scene consists of 5 parts: (a) six surround camera inputs (b) LiDAR scan, (c) ground-truth BEV segmentation map, (d) baseline BEV segmentation (BEVFusion [11]), (e) BEV segmentation using CIT module, and (d) BEV segmentation of MapFusion (full). In Fig. 10, we present qualitative results on a sample scene from nuScenes on the HD map construction task, showing both LiDAR and camera inputs. We compare the predicted vectorized HD map results of different models, including HDMapNet [1], VectorMapNet [2], the baseline (MapTR [3]), MapFusion (only using the CIT module), and the full MapFusion. We observe that the baseline model prediction is highly erroneous. By using the CIT module can already correct substantial errors in the baseline prediction, and the full MapFusion model further improves accuracy. Qualitative results demonstrate the advantages of the CIT module and the DDF module on the multi-modal map construction task.

5. Conclusion

To tackle the multi-modal BEV feature fusion problem in multi-modal map construction task, we propose a novel method named MapFusion, which can take advantage of the complementary information between BEV features of different modalities. Specifically, we first propose Cross-modal Interaction Transform (CIT) module to enhance one modality from another modality by the cross-attention mechanism. Moreover, we propose a Dual Dynamic Fusion (DDF) module to adaptively select valuable information from two modalities for better feature fusion. Extensive experiments on several benchmarks demonstrate the superiority of our method. We also verified the effectiveness of the MapFusion components via an extensive ablation study.

This paper provides a novel multi-modal BEV feature fusion method MapFusion for optimal fusion of RGB and LiDAR information. As shown in our experiments, our MapFusion brings consistent accuracy improvements for two different types of map reconstruction tasks in different datasets. Our MapFusion model can be simply integrated into existing pipelines in plug-and-play manner. Besides the map reconstruction task, we believe that MapFusion can also benefit other multi-modal perception tasks, which we leave for future work.

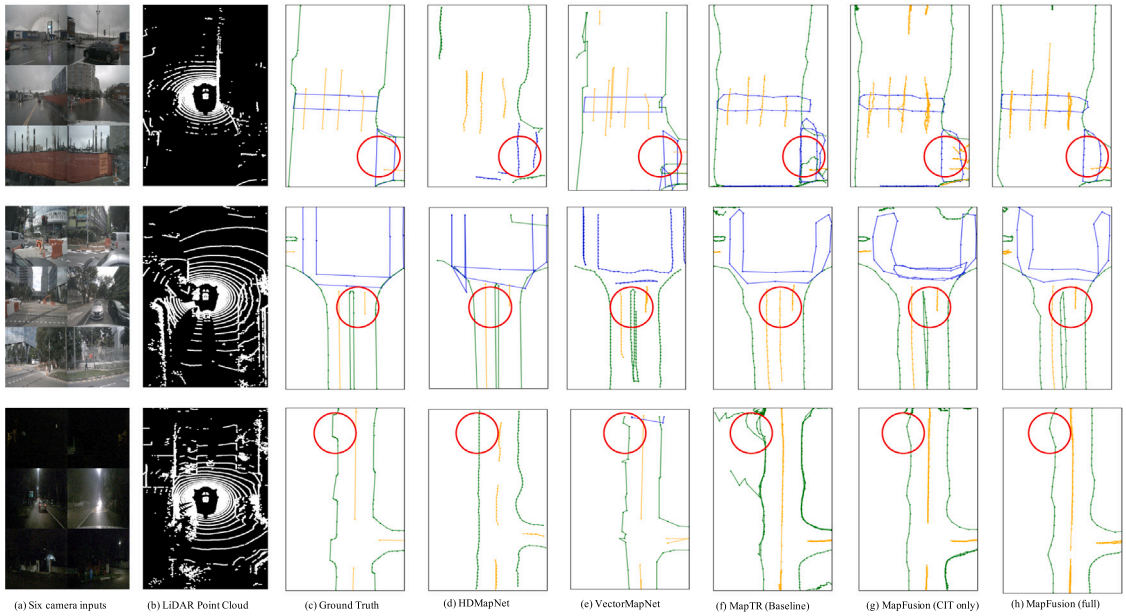


Fig. 10. Qualitative results on HD map task. We present a sample scene from nuScenes: (a) six camera inputs, (b) LiDAR scan, (c) ground-truth BEV vectorized HD map, (d) HDMapNet [1], (e) VectorMapNet [2], (f) baseline BEV vectorized HD map (MapTR [3]), (g) BEV vectorized HD map of only using CIT module, and (h) BEV vectorized HD map of MapFusion (full).

CRediT authorship contribution statement

Xiaoshuai Hao: Conceptualization. **Yunfeng Diao:** Methodology. **Mengchuan Wei:** Software. **Yifan Yang:** Resources. **Peng Hao:** Visualization. **Rong Yin:** Writing – original draft. **Hui Zhang:** Writing – original draft. **Weiming Li:** Writing – review & editing. **Shu Zhao:** Writing – review & editing. **Yu Liu:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: NA If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62302139), the National Natural Science Foundation of China (No. 62106259), the National Natural Science Foundation of China (No. 62176081) and the FRFCU-HFUT (JZ2023HGTA0202, JZ2023HGQA0101).

Data availability

The data that has been used is confidential.

References

[1] Q. Li, Y. Wang, Y. Wang, H. Zhao, Hdmapnet: An online hd map construction and evaluation framework, in: IEEE International Conference on Robotics and Automation, 2022, pp. 4628–4634.

[2] Y. Liu, T. Yuan, Y. Wang, Y. Wang, H. Zhao, Vectormapnet: End-to-end vectorized hd map learning, in: International Conference on Machine Learning, 2023, pp. 22352–22369.

[3] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, C. Huang, MapTR: Structured modeling and learning for online vectorized HD map construction, in: International Conference on Learning Representations, 2023.

[4] J. Philion, S. Fidler, Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D, in: European Conference on Computer Vision, 2020, pp. 194–210.

[5] T. Roddick, R. 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. 11135–11144.

[6] B. Zhou, P. 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.

[7] L. Qiao, W. Ding, X. Qiu, C. Zhang, End-to-end vectorized HD-map construction with piecewise Bézier curve, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13218–13228.

[8] T. Roddick, R. 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.

[9] B. Pan, J. Sun, H.Y.T. Leung, A. Andonian, B. Zhou, Cross-view semantic segmentation for sensing surroundings, IEEE Robot. Autom. Lett. 5 (3) (2020) 4867–4873.

[10] N. Gosala, A. Valada, Bird's-eye-view panoptic segmentation using monocular frontal view images, IEEE Robot. Autom. Lett. 7 (2) (2022) 1968–1975.

[11] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D.L. Rus, S. Han, BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation, in: ICRA, IEEE, 2023, pp. 2774–2781.

[12] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, J.M. Álvarez, M^2BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation, 2022, arXiv preprint arXiv:2204.05088.

[13] X. Hao, R. Li, H. Zhang, D. Li, R. Yin, S. Jung, S.-I. Park, B. Yoo, H. Zhao, J. Zhang, Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation, in: European Conference on Computer Vision, Springer, 2024, pp. 166–183.

[14] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, 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.

[15] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11784–11793.

[16] S. Vora, A.H. Lang, B. Helou, O. Beijbom, PointPainting: Sequential fusion for 3D object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4603–4611.

[17] T. Yin, X. Zhou, P. Krähenbühl, Multimodal virtual point 3D detection, in: Conference on Neural Information Processing Systems, 2021, pp. 16494–16507.

[18] S. Borse, M. Klingner, V.R. Kumar, H. Cai, A. Almuzairee, S. Yogamani, F. Porikli, X-Align: Cross-modal cross-view alignment for bird's-eye-view segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3287–3297.

[19] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, Z. Tang, BEVFusion: A simple and robust LiDAR-camera fusion framework, 2022, pp. 10421–10434,
[20] X. Hao, H. Zhang, Y. Yang, Y. Zhou, S. Jung, S.-I. Park, B. Yoo, Mbfusion: A new multi-modal bev feature fusion method for hd map construction, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 15922–15928.
[21] G. Zhang, J. Lin, S. Wu, Y. Song, Z. Luo, Y. Xue, S. Lu, Z. Wang, Online map vectorization for autonomous driving: A rasterization perspective, 2023, arXiv preprint arXiv:2306.10502.
[22] W. Ding, L. Qiao, X. Qiu, C. Zhang, PivotNet: Vectorized pivot learning for end-to-end HD map construction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3672–3682.
[23] L. Qiao, W. Ding, X. Qiu, C. Zhang, End-to-end vectorized HD-map construction with piecewise bezier curve, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13218–13228.
[24] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, X. Wang, MapTRv2: An end-to-end framework for online vectorized HD map construction, 2023, arXiv preprint arXiv:2308.05736.
[25] T. Yuan, Y. Liu, Y. Wang, Y. Wang, H. Zhao, Streammapnet: Streaming mapping network for vectorized online hd map construction, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7356–7365.
[26] X. Hao, G. Liu, Y. Zhao, Y. Ji, M. Wei, H. Zhao, L. Kong, R. Yin, Y. Liu, MSC-bench: Benchmarking and analyzing multi-sensor corruption for driving perception, 2025, arXiv preprint arXiv:2501.01037.
[27] X. Hao, M. Wei, Y. Yang, H. Zhao, H. Zhang, Y. Zhou, Q. Wang, W. Li, L. Kong, J. Zhang, Is your HD map constructor reliable under sensor corruptions? 2024.
[28] J. Philion, S. Fidler, Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, in: European Conference on Computer Vision, 2020, pp. 194–210.
[29] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, J. Dai, BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers, in: European Conference on Computer Vision, 2022, pp. 1–18.
[30] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, W. Liu, Efficient and robust 2D-to-BEV representation learning via geometry-guided kernel transformer, 2022, arXiv preprint arXiv:2206.04584.
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9992–10002.
[32] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer V2: Scaling up capacity and resolution, in: International Conference on Computer Vision and Pattern Recognition, 2022, pp. 11999–12009.
[33] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, M. Li, Mixgen: A new multi-modal data augmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 379–389.
[34] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., InternImage: Exploring large-scale vision foundation models with deformable convolutions, 2022, arXiv preprint arXiv:2211.05778.
[35] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, et al., BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision, 2022, arXiv preprint arXiv:2211.10439.
[36] X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, H. Zhao, Neural map prior for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17535–17544.
[37] T. Roddick, A. Kendall, R. Cipolla, Orthographic feature transform for monocular 3d object detection, 2018, arXiv preprint arXiv:1811.08188.
[38] S. Ammar Abbas, A. Zisserman, A geometric approach to obtain a bird's eye view from an image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[39] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, J. Lu, Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving, 2022, arXiv preprint arXiv:2205.09743.
[40] N. Garnett, R. Cohen, T. Pe'er, R. Lahav, D. Levi, 3D-lanenet: End-to-end 3d multiple lane detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2921–2930.
[41] M. Zhu, S. Zhang, Y. Zhong, P. Lu, H. Peng, J. Lenneman, Monocular 3D vehicle detection using uncalibrated traffic cameras through homography, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021, pp. 3814–3821.
[42] C. Wang, C. Ma, M. Zhu, X. Yang, PointAugmenting: Cross-modal augmentation for 3D object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 11794–11803.
[43] S. Xu, D. Zhou, J. Fang, J. Yin, B. Zhou, L. Zhang, FusionPainting: Multimodal fusion with adaptive attention for 3D object detection, in: IEEE International Intelligent Transportation Systems Conference, 2021, pp. 3047–3054.
[44] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, H. Zhao, AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection, in: International Joint Conference on Artificial Intelligence, 2022, pp. 827–833.
[45] Y. Chen, Y. Li, X. Zhang, J. Sun, J. Jia, Focal sparse convolutional networks for 3D object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5418–5427.
[46] M. Liang, B. Yang, S. Wang, R. Urtasun, Deep continuous fusion for multi-sensor 3D object detection, in: European Conference on Computer Vision, 2018, pp. 663–678.
[47] Z. Chen, H. Zhao, X. Hao, B. Yuan, X. Li, Stvit+: Improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization, Appl. Intell. 55 (5) (2025) 328.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[49] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[50] Y. Yan, Y. Mao, B. Li, SECOND: Sparsely embedded convolutional detection, Sensors 18 (10) (2018) 3337.
[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[52] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[53] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania, Calibrating deep neural networks using focal loss, Adv. Neural Inf. Process. Syst. (2020) 15288–15299.
[54] M. Malkauthekar, Analysis of Euclidean distance and Manhattan distance measure in face recognition, in: Third International Conference on Computational Intelligence and Information Technology, CIIT 2013, 2013, pp. 503–507.
[55] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, Nuscenes: A multimodal dataset for autonomous driving, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11618–11628.
[56] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J.K. Pontes, D. Ramanan, P. Carr, J. Hays, Argoverse 2: Next generation datasets for self-driving perception and forecasting, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.
[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[58] M. Contributors, MMDetection3D: OpenMMLab next-generation platform for general 3D object detection, 2020.
[59] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. (2008).

Xiaoshuai Hao received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences in 2023. He is currently a researcher at the Beijing Academy of Artificial Intelligence, focusing on embodied multimodal large models. His research interests include multimedia retrieval, multimodal learning, and embodied intelligence.

Yunfeng Diao is a Lecturer in the School of Computer Science and Information Engineering, Hefei University of Technology, China. He received his PhD from Southwest Jiaotong University, China. His current research interests include computer vision and the security of machine learning.

Mengchuan Wei received the Master's degree in Communication Engineering from Beijing University of Posts and Telecommunications in 2016. In 2022, he joined Samsung Research Center Beijing, China, where he is an algorithm engineer. His research interests include computer vision and artificial intelligence.

Yifan Yang received the B.E. degree in automation from Northwestern Polytechnical University, China, in 2018, and M.E. degree in control engineering from Harbin Engineering University, China, in 2021. In 2021, he joined Samsung Research Center Beijing, China, where he is an Engineer. His research interests include computer vision and machine learning.

Peng Hao received the B.E. degree from Tianjin University, Tianjin, China, in 2017, and the Ph.D. degree in Robotics from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was also affiliated with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. He is currently a researcher with Samsung Research China-Beijing, Beijing. His research interests include scene understanding, robotic dexterous manipulation and task planning.

Rong Yin received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, in 2020. She is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her current research interests include machine learning, data mining, statistical theory, distributed learning, self-supervised learning, and graph representation learning.

Hui Zhang received the B.S. degree in applied mathematics from Tsinghua University, China, in 1997, and Ph.D degree in computer applications from Tsinghua University, China in 2003. In 2009, he joined Samsung Research Center Beijing, China, where he is a Principal Engineer. His research interests include computer vision and machine learning.

Weiming Li received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He was with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong, from 2008 to 2011. In 2011, he joined Samsung Research Center Beijing,China, where he is currently a Principal Engineer. His current research interests include computer vision and artificial intelligence.

Shu Zhao is a Ph.D. student in the Department of Computer Science and Engineering at The Pennsylvania State University, University Park, PA, USA. He received a Bachelor of Engineering degree from Anhui University and a Master of Engineering degree from the University of Chinese Academy of Sciences. His research focuses on multi-modal large language models and their robustness.

Yu Liu received the B.S. degree and Ph. D degree from the Department of Automation, University of Science and Technology of China in 2011 and 2016, respectively. He is currently a Full Professor in the Department of Biomedical Engineering at Hefei University of Technology. His research interests include image processing, computer vision and machine learning. In particular, his current research is mainly focused on image fusion, image restoration, visual recognition, medical image segmentation, and signal/image/video-based biomedical applications. He has published over 100 scientific articles in prestigious journals and conferences. He is serving as an Editorial Board Member for Information Fusion, and an Associate Editor for IEEE Signal Processing Letters. He was a recipient of the IEEE Instrumentation and Measurement Society Andi Chi Best Paper Award in 2020 and the IET Image Processing Premium (Best Paper) Award in 2017. He was identified as a Clarivate Highly Cited Researcher in 2023 and 2024.