2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

MBFusion: A New Multi-modal BEV Feature Fusion Method for HD Map Construction

Xiaoshuai Hao[1], Hui Zhang[1], Yifan Yang[1], Yi Zhou[1], Sangil Jung[2], Seung-In Park[2] and ByungIn Yoo[2]

***Abstract*— HD map construction is a fundamental and challenging task in autonomous driving to understand the surrounding environment. Recently, Camera-LiDAR BEV feature fusion methods have attracted increasing attention in HD map construction task, which can significantly boost the benchmark. However, existing fusion methods ignore modal interaction and utilize very simple fusion strategy, which suffers from the problems of misalignment and information loss. To tackle this, we propose a novel Multi-modal BEV feature fusion method named *MBFusion*. Specifically, to solve the semantic misalignment problem between Camera and LiDAR features, we design Cross-modal Interaction Transform (CIT) module to make these two feature spaces interact knowledge with each other to enhance the feature representation by the cross-attention mechanism. Then, we propose a Dual Dynamic Fusion (DDF) module to automatically select valuable information from different modalities for better feature fusion. Moreover, MBFusion is simple, and can be plug-and-played into existing pipelines. We evaluate MBFusion on three architectures, including HDMapNet, VectorMapNet, and MapTR, to show its versatility and effectiveness. Compared with the state-of-the-art methods, MBFusion achieves 3.6% and 4.1% absolute improvements on mAP on the nuScenes and the Argoverse2 datasets, respectively, demonstrating the superiority of our method.**

I. Introduction

High-definition (HD) map provides abundant and precise static environmental information of the driving scene, which is vital and challenging for planning in autonomous driving system. HD map construction methods [14], [20], [18], [25], [29], [42], [26] 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.

HD map construction methods are generally classified into three groups based on input modality: Camera-based methods [27], [24], [28], [40], LiDAR-based approaches [14], [20], [12] and Camera-LiDAR fusion methods [14], [18], [22], [30]. Camera-based methods learn to project the perspective view (PV) features onto BEV space through the geometric prior, which often have the spatial distortions inevitably. LiDAR-based methods directly capture the spatial information for the unified BEV feature representation, which suffers from the data sparsity and the sensing noise. Recently, Camera-LiDAR BEV feature fusion methods have attracted increasing attention in HD map construction task, which can make use of both the semantic-rich information from camera and the explicit geometric information from LiDAR. However, existing methods for multi-modal BEV feature fusion ignore modal interaction and utilize very simple fusion strategy, which suffers from the problems of misalignment and information loss.

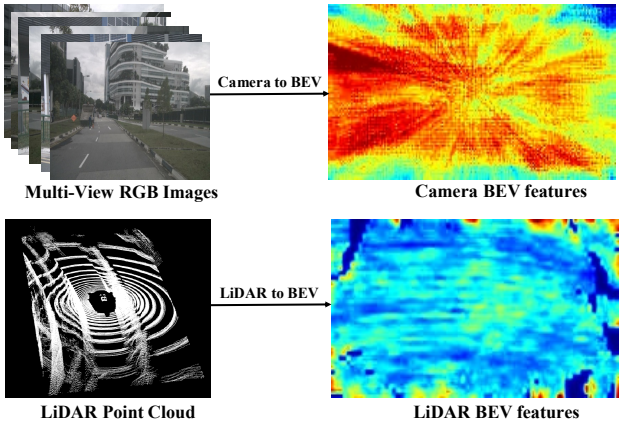

Fig. 1: Illustration of different modalities BEV features. Though in the same BEV space, LiDAR BEV features and camera BEV features can still be semantically misaligned to some extent due to the large modality gap. (Best viewed in color. Blue color means small values and red means large.)

As illustrated in Fig. 1, though in the same BEV space, LiDAR BEV features and camera BEV features can still be semantically misaligned to some extent due to the large modality gap. In addition, it is also very important to design effective cross-modal fusion strategies to adaptively select valuable information from different modalities for better feature fusion. These are rather challenging and thus become the motivation of our work, i.e. a novel multi-modal BEV feature fusion method named MBFusion. Specifically, to solve the semantic misalignment problem between Camera and LiDAR features, we design Cross-modal Interaction Transform (CIT) module to make these two feature space interact knowledge with each other to enhance the feature representation by the cross-attention mechanism. Then, we propose a Dual Dynamic Fusion (DDF) module to automatically select valuable information from two modalities for better feature fusion. Extensive experiments on several benchmarks demonstrate the superiority of our method. Our main contributions are summarized as follows:

- To tackle the multi-modal BEV feature fusion problem in HD map construction task, we develop a novel method named MBFusion, which can take advantage of the complementary information between BEV features of different modalities.

[1]Samsung R&D Institute China–Beijing, Beijing, 100108, China. E-mail:{xshuai.hao, hui123.zhang, yifan.yang, yi0813.zhou}@samsung.com.
[2]Computer Vision TU, SAIT, SEC, Korea. E-mail:{sang-il.jung, si14.park, byungin.yoo}@samsung.com.

- 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.
- Our method achieves 3.6% and 4.1% absolute improvements on mAP compared with the state-of-the-art method on the nuScenes and the Argoverse2 datasets, respectively, demonstrating the superiority of our method.

II. Related Work

HD Map Construction. High-definition (HD) map construction techniques have been extensively researched in autonomous driving field. There are three main types of HD map construction approaches: Camera-based methods, LiDAR-based approaches and Camera-LiDAR fusion methods. Camera-based methods [27], [24], [28], [40] learn to project the perspective view (PV) features onto BEV space through the geometric prior, which often have the spatial distortions inevitably. Besides, the camera-based methods rely on high-resolution images and large pretrained models for better accuracy [21], [15], [37], [41], which brings serious challenges to the practical scenarios. LiDAR-based methods [14], [20], [12] directly capture the accurate spatial information for the unified BEV feature representation which suffers from the data sparsity and the sensing noise. Recently, Camera-LiDAR BEV feature fusion methods [14], [18], [22], [30], [19], [7], [8], [9] make use of both the semantic-rich information from camera and the explicit geometric information from LiDAR. They achieve better results than those approaches with single modality input. However, existing methods for multi-modal BEV feature fusion ignore modal interaction, which suffers from the problem of misalignment problem. In this paper, we focus on a simple and effective Camera-LiDAR BEV feature fusion method to simultaneously fuse complementary information between different modalities.

Multi-sensor Fusion. Recently, multi-sensor fusion attracts increased attention in autonomous driving field. Existing approaches can roughly be divided into three categories: point-level fusion, feature-level fusion, and BEV-level fusion. Point-level fusion methods usually paint image semantic features onto foreground LiDAR points and perform LiDAR-based detection on the decorated point cloud inputs, such as PointAugmenting [33], PointPainting [32], FusionPainting [35], AutoAlign [6], and MVP [38]. Feature-level fusion methods firstly project the LiDAR points into a feature space [39] or generate proposals [1], [4], query the associated camera features and then concatenate back to the feature space [5], [16]. However, these two fusion types have downsides regarding generalization. While point-level fusion is not generically extendable to other sensor modalities, feature-level fusion is not easily generalizable to other tasks. Recently, multi-modal feature fusion in the unified BEV space has attracted some attention [22], [17]. BEV-level fusion uses two independent streams that encode the raw inputs from the camera and LiDAR sensors into features within the same BEV space. This provides a simple effective method to fuse these BEV-level features after these two streams, so that the final feature can be used in various downstream tasks. However, the BEV-fusion methods usually fuse them together with element-wise operations (summation, mean) or concatenation, resulting in sub-optimal fusion results. In this paper, we design an effective cross-modal fusion strategy to adaptively select valuable information from two modalities for better feature fusion. Moreover, our method can be easily plug-and-played into existing pipelines.

III. Methodology

In this section, we first introduce the preliminaries of this paper in Section III-A. Then, we elaborate details of our method, including the Cross-modal Interaction Transform (CIT) module, the Dual Dynamic Fusion (DDF) module and the overall training procedure in Sections III-B, III-C, III-D, respectively.

A. 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 vectorized map elements in BEV space, and the set of map elements classes: road boundary, lane divider, and pedestrian crossing. As illustrated in Fig. 2, the set of inputs, $\chi = \{\textit{Camera}, \textit{LiDAR}\}$, contains multi-view RGB camera images in perspective view, $\textit{Camera} \in \mathbb{R}^{N^{cam} \times H^{cam} \times W^{cam} \times 3}$, N^{cam}, H^{cam}, W^{cam} denote number of cameras, image height, and image width, respectively, as well as a LiDAR point cloud, $\textit{LiDAR} \in \mathbb{R}^{P \times 5}$, with number of points P. Each point consists of its 3-dimensional coordinates, reflectivity, and ring index. The overall framework of our method is illustrated in Fig. 2.

Multi-Modal Feature Extractors. We first establish a baseline method of fusion-based HD map construction based on MapTR [18]. As shown in Fig. 2, initial features are extracted from both sensor inputs. For camera images, we firstly utilize Resnet50 [10] as backbone to extract the multi-view features. Then we adopt GKT [3] as the 2D-to-BEV transformation module to convert the multi-view features into BEV space. The generated BEV features can be denoted as $\mathbf{F}^{\mathbf{BEV}}_{\mathbf{Camera}} \in \mathbb{R}^{H \times W \times C}$, where H, W, C represents the height, width and the number of channels of BEV features respectively. For the LiDAR points, we follow SECOND [36] in using voxelization and a sparse LiDAR encoder. The LiDAR features are projected to BEV space using a flattening operation as in [22], to obtain the unified BEV representation $\mathbf{F}^{\mathbf{BEV}}_{\mathbf{LiDAR}} \in \mathbb{R}^{H \times W \times C}$.

B. Cross-modal Interaction Transform (CIT)

Existing methods directly convert all sensory features to the shared BEV representation, and then fuse them together with arithmetic or splicing operations to obtain multi-modal

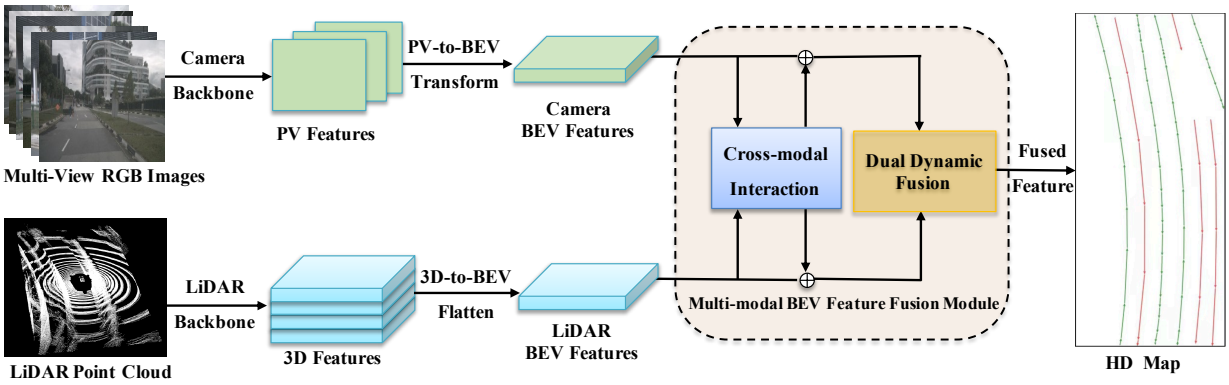

Fig. 2: An overview of MBFusion framework. First, we extract features from multi-modal inputs and convert them into a shared bird's-eye view (BEV) space efficiently using view transformations. To fuse the BEV features from different modalities, we first propose Cross-modal Interaction Transform (CIT) module to enhance one modality from another modality by cross-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 detector and prediction heads for HD Map Construction.

BEV features. Though in the same BEV space, LiDAR BEV features and camera BEV features can still be semantically misaligned to some extent due to the large modality gap, which leads to the 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 cross-attention mechanism. Next, we describe in detail our proposed CIT module.

First, given the BEV features from both camera ($\mathbf{F}^{\mathbf{BEV}}_{\mathbf{Camera}} \in \mathbb{R}^{H\times W\times C}$) and LiDAR ($\mathbf{F}^{\mathbf{BEV}}_{\mathbf{LiDAR}} \in \mathbb{R}^{H\times W\times C}$) sensors, the BEV tokens $\mathbf{T}^{\mathbf{BEV}}_{\mathbf{Camera}} \in \mathbb{R}^{HW\times C}$ and $\mathbf{T}^{\mathbf{BEV}}_{\mathbf{LiDAR}} \in \mathbb{R}^{HW\times C}$ 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 $\mathbf{T}^{in} \in \mathbb{R}^{2HW\times C}$ of the Transformer [31]. The positional embedding enables the model to differentiate spatial information between different tokens at training time. Third, the input token $\mathbf{T}^{in}$ uses linear projections for computing a set of queries, keys and values ($\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$),

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

where $\mathbf{W}^{Q} \in \mathbb{R}^{C\times D_Q}$, $\mathbf{W}^{K} \in \mathbb{R}^{C\times D_K}$ and $\mathbf{W}^{V} \in \mathbb{R}^{C\times D_V}$ are weight matrices. Moreover, D_Q, D_K and D_V are equal in our Transformer, i.e., $D_Q = D_K = D_V = C$. 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 $\mathbf{Z}$,

$$\mathbf{Z} = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{D_k}})\mathbf{V}, \tag{2}$$

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

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

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

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

The output $\mathbf{T}^{out}$ are converted into $\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{Camera}}$ and $\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{LiDAR}}$ for further feature fusion.

The key idea behind our module is leveraging the self attention mechanism to incorporate the global information for camera and LiDAR modalities given their complementary nature. Specifically, we leverage the correlation matrix to weight each position of the input multi-modal BEV features. Therefore, the CIT module can automatically perform simultaneous intra-modality and inter-modality information fusion and robustly capture the complementary information between BEV features of different modalities.

C. 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 strategies to adaptively select valuable information from different modalities for better feature fusion is still very important. Recently, multi-modal BEV feature fusion methods [17], [22], have received much attention. It is a common approach to utilize concatenation followed by convolution to combine features from multi-modal BEV feature inputs, $\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{Camera}}$ and $\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{LiDAR}}$, resulting in the aggregated features $\mathbf{F}_{fused}$, as shown in Fig. 3 (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. 3 (b) Add Fusion.

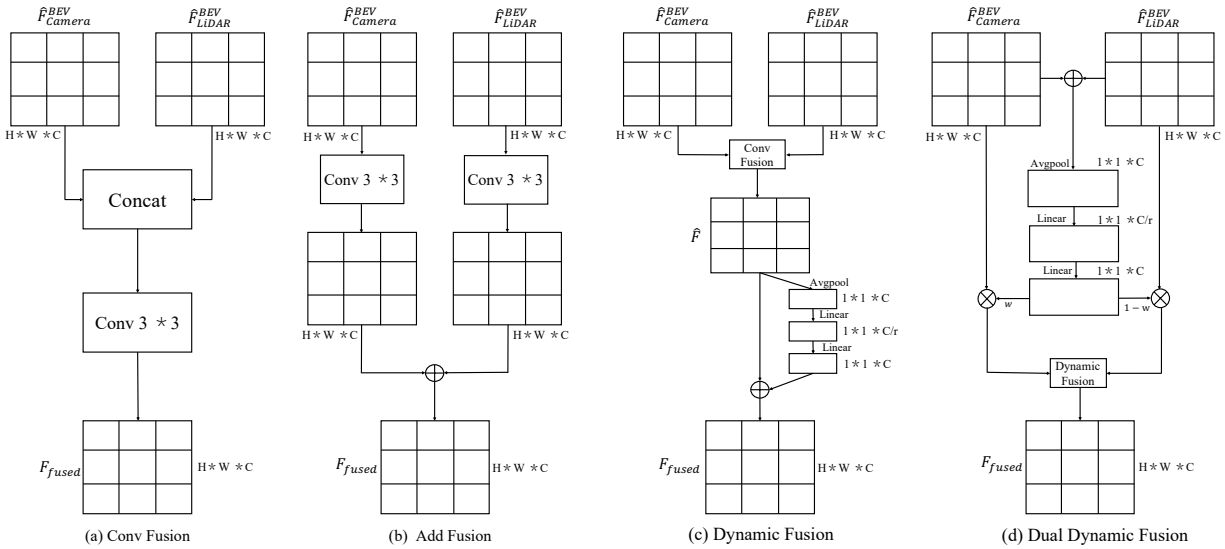

Fig. 3: Three existing fusion strategies and our proposed Dual Dynamic Fusion (DDF) strategy.

As Fig. 3 (c) illustrates, the input of the dynamic fusion module is the Conv Fusion output features and then fuse them with learnable static weights, inspired by Squeeze-and-Excitation mechanism [13]. 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. 3(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 the squeeze and excitation operations that generate the attention weights. We can formulate this process as:

$$w = \sigma(\gamma(AvgPool(\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{Camera}} + \hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{LiDAR}}))), \tag{5}$$

where σ and γ represent the sigmoid function and linear layers respectively, and w denotes the attention weights. We then multiply **w** and **1-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:

$$\mathbf{F}_{fused} = f_{adaptive}(f_{concat}([w\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{Camera}}, (1-w)\hat{\mathbf{F}}^{\mathbf{BEV}}_{\mathbf{LiDAR}}])), \tag{6}$$

where $[\cdot;\cdot]$ denotes the concatenation operation along the channel dimension. f_{concat} is a static channel and spatial fusion function implemented by a 3×3 convolution layer to reduce the channel dimension of concatenated feature to C. With input feature $\hat{\mathbf{F}} \in \mathbb{R}^{H \times W \times C}$, $f_{adaptive}$ is formulated as:

$$f_{adaptivate}(\hat{\mathbf{F}}) = \sigma(\mathbf{W} f_{avg}(\hat{\mathbf{F}})) \cdot \hat{\mathbf{F}}, \tag{7}$$

where $\mathbf{W}$ denotes linear transform matrix (e.g., 1×1 convolution), f_{avg} denotes the global average pooling 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 $\mathbf{F}_{fused}$ will be used for HD Map construction task, with the decoder and prediction heads from MapTR [18].

D. Overall Training

We follow the MapTR [18] model training loss function, which is composed of three parts, including the classification loss $\mathcal{L}_{cls}$, the point2point loss $\mathcal{L}_{p2p}$, and the edge direction loss $\mathcal{L}_{dir}$. Combining these loss terms together, the overall objective function can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{p2p} + \lambda_3 \mathcal{L}_{dir}, \tag{8}$$

where λ_1, λ_2 and λ_3 are hyper parameters for balancing these terms.

IV. Experiments

A. Experimental Setup

Dataset. There are two widely adopted datasets for HD Map Construction task, including nuScenes dataset [2] and Argoverse2 dataset [34]. NuScenes dataset contains 1000 sequences of recordings collected by autonomous driving cars. Each episode is annotated at 2Hz and contains 6 camera images and LiDAR sweeps. Our dataset setup and pre-processing steps are identical to that of MapTR [18], which includes three categories of map elements, pedestrian crossing, divider, and road boundary. Moreover, we further conduct experiments on Argoverse2 dataset. Like nuScenes, it contains 1000 logs (700, 150, 150 for training, validation and test set). Each episode provides 15s of 20Hz camera images, 10Hz LiDAR sweeps and a vectorized map. We use the same pre-processing settings as on nuScenes dataset.

Evaluation Metrics. 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.

TABLE I: Comparisons with state-of-the-art methods on nuScenes val set. We compare with existing methods from literature, where the numbers are taken from MapTR [18]. We also provide information on the backbones, epochs and input modalities in the table. Our proposed MBFusion 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 [14]	Camera	Efficient-B0	30	14.4	21.7	33.0	23.0
HDMapNet [14]	LiDAR	PointPillars	30	10.4	24.1	37.9	24.1
HDMapNet [14]	Camera & LiDAR	Efficient-B0 & PointPillars	30	16.3	29.6	46.7	31.0
VectorMapNet [20]	Camera	ResNet-50	110	36.1	47.3	39.3	40.9
VectorMapNet [20]	LiDAR	PointPillars	110	25.7	37.6	38.6	34.0
VectorMapNet [20]	Camera & LiDAR	ResNet-50 & PointPillars	110	37.6	50.5	47.5	45.2
MapTR [18]	Camera	ResNet-50	24	46.3	51.5	53.1	50.3
MapTR [18]	LiDAR	SECOND	24	48.5	53.7	64.7	55.6
MapTR [18]	Camera & LiDAR	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
MBFusion	Camera & LiDAR	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	**66.1**$_{+3.6}$
MBFusion	Camera & LiDAR	ResNet-50 & SECOND	110	**67.9**	**72.2**	**78.2**	**72.8**$_{+10.3}$

TABLE II: Results on Argoverse2 dataset. † denotes our re-implementation following the setting in the paper.

Method	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet [14]	Camera	Efficient-B0	30	13.1	5.70	37.6	18.8
VectorMapNet [20]	Camera	ResNet-50	110	38.3	36.1	39.2	37.9
MapTR† [18]	Camera	ResNet-50	6	58.7	59.3	60.3	59.4
MapTR† [18]	Camera & LiDAR	ResNet-50 & SECOND	6	65.1	61.6	75.1	67.3
MBFusion	Camera & LiDAR	ResNet-50 & SECOND	6	**69.4**	**65.8**	**78.9**	**71.4**$_{+4.1}$

TABLE III: Ablation study on the proposed MBFusion components. "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
✗	✗	Camera & LiDAR	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
✓	✗	Camera & LiDAR	ResNet-50 & SECOND	24	58.4	64.1	72.5	65.0
✗	✓	Camera & LiDAR	ResNet-50 & SECOND	24	60.2	64.3	72.1	65.5
✓	✓	Camera & LiDAR	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	**66.1**

Models and Training. MBFusion is trained with 4 NVIDIA RTX A6000 GPUs. We adopt the AdamW optimizer [23] for all our experiments and λ_1 is set to 2, λ_2 is set to 5, and λ_3 is set to $5e^{-3}$ during training. Moreover, we adopt ResNet50 [11] and SECOND [36] as the backbone and employ GKT [3] as the default 2D-to-BEV module. Note that our method is orthogonal to the camera and LiDAR feature extraction, allowing us to flexibly embrace state-of-the-art camera and LiDAR encoders. We set the mini-batch size to 16, and utilize a step-decayed learning rate with initialization value $6e^{-4}$.

B. Comparison with the State-of-the-Arts

With the same settings and data partition, we compare the proposed MBFusion method with several state-of-the-art methods, i.e., HDMapNet [14], VectorMapNet [20] and MapTR [18]. Table I and Table II show the overall performance of MBFusion and all the baselines on nuScenes and Argoverse2 datasets, respectively. 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) The proposed MBFusion approach achieves 3.6% absolute improvement compared with the prior state-of-the-art MapTR [18] method in mAP on the nuScenes dataset. Similarly, MBFusion achieves 4.1% absolute improvement compared with the MapTR [18] method in mAP on the Argoverse2 dataset. In a nutshell, MBFusion shows significant superiority over other multi-modal methods, indicating the benefit of cross-modal interaction transform (CIT) module and dual dynamic fusion (DDF) module.

C. Ablation Studies

Analysis on different modules. To systematically evaluate the effectiveness of each module of our proposed MBFusion, we train the model by removing each component solely and present the results in Tab. III. In the main ablation study, we design the following ablation models: (1) **MBFusion (w/o CIT)** : we remove the cross-modal interaction transform module from MBFusion; (2) **MBFusion (w/o DDF)**: we remove the dual dynamic fusion from MBFusion; (3) **MBFusion (full)**: our full MBFusion model. The results of MBFusion (w/o CIT) and MBFusion (w/o DDF) are inferior to the full MBFusion method, verifying the effectiveness of both components.

Analysis on different Fusion methods. In Sec. III-C, we introduce three existing fusion strategies and our proposed Dual Dynamic Fusion (DDF) strategy. Detailed comparisons

TABLE IV: Comparison with different fusion strategies. 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)	Camera & LiDAR	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
Add Fusion	Camera & LiDAR	ResNet-50 & SECOND	24	**61.1**	60.3	71.8	64.4$_{+1.9}$
Dynamic Fusion	Camera & LiDAR	ResNet-50 & SECOND	24	58.4	63.1	71.5	64.3$_{+1.8}$
Dual Dynamic Fusion	Camera & LiDAR	ResNet-50 & SECOND	24	58.4	**64.1**	**72.5**	**65.0**$_{+2.5}$

TABLE V: Compatibility to other HD map construction methods. Adding MBFusion leads to consistent performance boost on nuScenes val set in terms of mAP. † denotes our re-implementation following the setting in the original papers.

Method	Venue	Modality	Backbone	Epochs	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
HDMapNet† [14]	ICRA 22	Camera & LiDAR	Efficient-B0 & PointPillars	30	13.3	26.9	44.3	28.2
HDMapNet + MBFusion	-	Camera & LiDAR	Efficient-B0 & PointPillars	30	**21.1**	**34.2**	**52.1**	**35.8**$_{+7.6}$
VectorMapNet† [20]	ICML 23	Camera & LiDAR	ResNet-50 & PointPillars	110	35.8	48.2	45.3	43.1
VectorMapNet + MBFusion	-	Camera & LiDAR	ResNet-50 & PointPillars	110	**41.1**	**53.7**	**50.9**	**48.6**$_{+5.5}$
MapTR-tiny [18]	ICLR 23	Camera & LiDAR	ResNet-50 & SECOND	24	55.9	62.3	69.3	62.5
MapTR-tiny + MBFusion	-	Camera & LiDAR	ResNet-50 & SECOND	24	**61.6**	**64.4**	**72.5**	**66.1**$_{+3.6}$

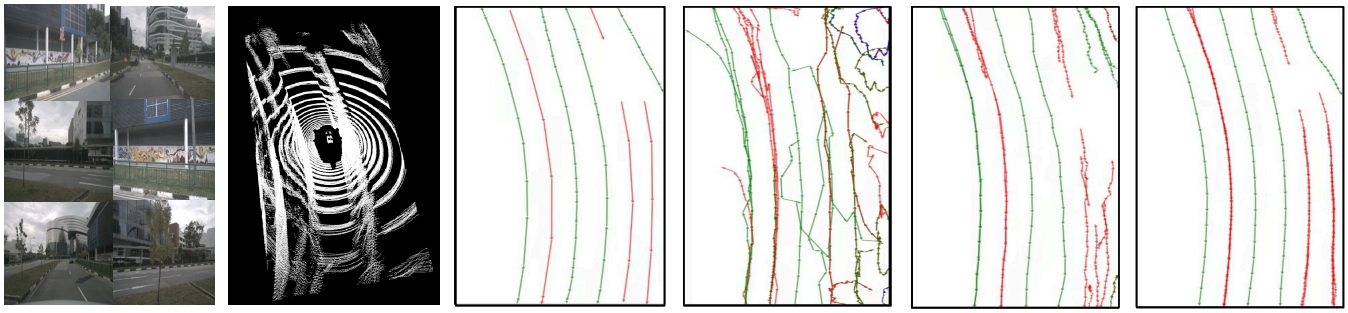

Fig. 4: Qualitative results on nuScenes. We present a sample scene from nuScenes: a) six camera inputs, b) LiDAR scan, c) ground-truth BEV vectorized HD map, d) baseline BEV vectorized HD map (MapTR [18]), e) BEV vectorized HD map of only using CIT module, and f) BEV vectorized HD map of MBFusion (full).

among strategies can be found in Table IV. Our DDF strategy achieves the overall best performance, demonstrating its effectiveness. For instance, the proposed DDF method achieves 2.5% absolute improvements compared with Conv Fusion method, which indicates that the DDF module plays an essential role in the multi-modal BEV feature fusion.

Compatibility with other HD Map Construction methods. We show MBFusion is compatibility with other HD Map Construction methods, i.e., HDMapNet [14], VectorMapNet [20] and MapTR [18]. Besides adding MBFusion, we do not modify their original training settings. For all experiment, we report the result of nuScenes val set. As shown in Table V, simply adding MBFusion on top of these strong baselines consistently improve state-of-the-art performance. MBFusion demonstrates significant accuracy boost (absolute): HDMapNet(+7.6%), VectorMapNet (+5.5%), and MapTR (+3.6%). This shows the versatility of MBFusion as multi-modal BEV feature fusion method.

D. Qualitative Results

In Fig. 4, we present qualitative results on a sample scene from nuScenes, showing both LiDAR and camera inputs. We compare the predicted vectorized HD map results of different models, including the baseline (MapTR [18]), MBFusion (only using the CIT module), and the full MBFusion. 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 MBFusion model further improves accuracy.

V. Conclusion

To tackle the multi-modal BEV feature fusion problem in HD map construction task, we propose a novel method named MBFusion, 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 MBFusion components via an extensive ablation study. As part of future work, we believe that MBFusion can further benefit other multi-modal perception tasks.

REFERENCES

[1] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C. Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1080–1089, 2022.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11618–11628, 2020.

[3] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, and W. Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*, 2022.

[4] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao. FUTR3D: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2023.

[5] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia. Focal sparse convolutional networks for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427, 2022.

[6] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. In *International Joint Conference on Artificial Intelligence*, pages 827–833, 2022.

[7] X. Hao and W. Zhang. Uncertainty-aware alignment network for cross-domain video-text retrieval. In *NIPS*, 2023.

[8] X. Hao, W. Zhang, D. Wu, F. Zhu, and B. Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *CVPR*, pages 18962–18972, 2023.

[9] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li. Mixgen: A new multi-modal data augmentation. In *WACV*, pages 379–389, 2023.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, Y. Xie, C. Wang, and J. Philbin. FISHING net: Future inference of semantic heatmaps in grids. *arXiv preprint arXiv:2006.09917*, 2020.

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

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

[15] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pages 1–18, 2022.

[16] M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *European Conference on Computer Vision*, pages 663–678, 2018.

[17] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang. Bevfusion: A simple and robust lidar-camera fusion framework. pages 10421–10434, 2022.

[18] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *International Conference on Learning Representations*, 2023.

[19] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang. Maptrv2: An end-to-end framework for online vectorized HD map construction. *arXiv preprint arXiv:2308.05736*, 2023.

[20] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369, 2023.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9992–10002, 2021.

[22] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pages 2774–2781, 2023.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5924–5932, 2023.

[25] J. Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210, 2020.

[26] L. Qiao, W. Ding, X. Qiu, and C. Zhang. End-to-end vectorized hd-map construction with piecewise bézier curve. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13228, 2023.

[27] L. Qiao, W. Ding, X. Qiu, and 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*, pages 13218–13228, 2023.

[28] Z. Qin, J. Chen, C. Chen, X. Chen, and X. Li. Uniformer: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. *arXiv preprint arXiv:2207.08536*, 2022.

[29] T. Roddick and 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*, pages 11135–11144, 2020.

[30] G. Salazar-Gomez, D. S. González, M. Diaz-Zapata, A. Paigwar, W. Liu, Ö. Erkent, and C. Laugier. Transfusegrid: Transformer-based lidar-rgb fusion for semantic grid prediction. In *International Conference on Control, Automation, Robotics and Vision*, pages 268–273, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[32] S. Vora, A. H. Lang, B. Helou, and O. Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4603–4611, 2020.

[33] C. Wang, C. Ma, M. Zhu, and X. Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[34] 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, and 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.

[35] S. Xu, D. Zhou, J. Fang, J. Yin, B. Zhou, and L. Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *IEEE International Intelligent Transportation Systems Conference*, pages 3047–3054, 2021.

[36] Y. Yan, Y. Mao, and B. Li. SECOND: sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[37] 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. *arXiv preprint arXiv:2211.10439*, 2022.

[38] T. Yin, X. Zhou, and P. Krähenbühl. Multimodal virtual point 3d detection. In *Conference on Neural Information Processing Systems*, pages 16494–16507, 2021.

[39] J. H. Yoo, Y. Kim, J. S. Kim, and J. W. Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *European Conference on Computer Vision*, pages 720–736, 2020.

[40] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.

[41] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.

[42] B. Zhou and 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*, pages 13760–13769, 2022.