

Eliminating Cross-modal Conflicts in BEV Space for LiDAR-Camera 3D Object Detection

Jiahui Fu¹, Chen Gao^{1†}, Zitian Wang¹, Lirong Yang², Xiaofei Wang², Beipeng Mu², Si Liu¹

Abstract—Recent 3D object detectors typically utilize multi-sensor data and unify multi-modal features in the shared bird’s-eye view (BEV) representation space. However, our empirical findings indicate that previous methods have limitations in generating fusion BEV features free from cross-modal conflicts. These conflicts encompass extrinsic conflicts caused by BEV feature construction and inherent conflicts stemming from heterogeneous sensor signals. Therefore, we propose a novel Eliminating Conflicts Fusion (ECFusion) method to explicitly eliminate the extrinsic/inherent conflicts in BEV space and produce improved multi-modal BEV features. Specifically, we devise a Semantic-guided Flow-based Alignment (SFA) module to resolve extrinsic conflicts via unifying spatial distribution in BEV space before fusion. Moreover, we design a Dissolved Query Recovering (DQR) mechanism to remedy inherent conflicts by preserving objectness clues that are lost in the fusion BEV feature. In general, our method maximizes the effective information utilization of each modality and leverages inter-modal complementarity. Our method achieves state-of-the-art performance in the highly competitive nuScenes 3D object detection dataset. The code is released at <https://github.com/fjzhxixi/ECFusion>.

I. INTRODUCTION

3D object detection is crucial for enabling safe and effective autonomous driving, allowing vehicles to locate and recognize objects in real-world 3D environments accurately. To achieve precise and dependable 3D object detection, some methods combine information from LiDAR point clouds and camera RGB images via various multi-modal fusion strategies. Specifically, point clouds provide accurate 3D localization information, while RGB images offer rich contextual details. Thus combining these complementary modalities improves the accuracy and robustness of 3D object detection. Recently, advanced methods [1], [2] attempt to fuse LiDAR-Camera features in the unified bird’s-eye view (BEV) space.

In general, BEV space provides a suitable intermediate representation for multi-modal feature fusion. Nevertheless, existing fusion strategies only consider benefits brought by the complementary BEV features between modalities, while ignoring the interference caused by cross-modal conflicts. However, we argue that multi-modal fusion operations, influenced by conflicts among heterogeneous cross-modal features, potentially undermine accurate predictions. In particular, the cross-modal conflicts are mainly from two aspects,

¹Jiahui Fu, Chen Gao, Zitian Wang, and Si Liu are with Institute of Artificial Intelligence and Hangzhou Innovation Institute, Beihang University. Email: {jiahuifu, gaochen, wangzt, liusi}@buaa.edu.cn

²Lirong Yang, Xiaofei Wang, and Beipeng Mu are with Meituan Inc. Email: {yanglirong, wangxiaofei19, mubeipeng}@meituan.com

[†]Corresponding author: Chen Gao

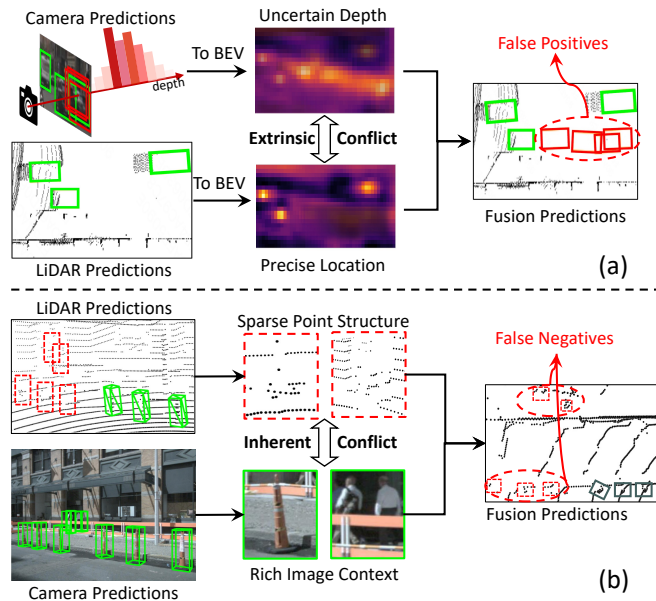


Fig. 1. Cross-modal conflicts hinder LiDAR-Camera 3D object detection. Green boxes represents the correct prediction. Red dotted boxes and red solid boxes represent false negative and false positive respectively. (a) The single-modal LiDAR prediction is accurate, yet extrinsic conflicts caused by the uncertain depth of images lead to false positive results in fusion predictions. (b) The single-modal camera prediction is accurate, yet inherent conflicts caused by the sparse points structure of small objects lead to false negative results in fusion predictions. Best viewed in color.

i.e., **extrinsic conflict** and **inherent conflict**. (I) The extrinsic conflicts arise from variations across different modalities in the construction process of BEV features. To elaborate, LiDAR and camera modalities exhibit spatially misaligned BEV feature distributions since they are extracted separately by independent encoders and mapped to BEV using different projection methods. Consequently, these misalignments inevitably lead to incorrect object information when merged. For instance shown in Fig. 1(a), the car that can be located correctly in LiDAR predictions has obviously misaligned spatial features in camera BEV since redundant objects are projected based on the uncertain depth. Such extrinsic conflicts from feature projection result in false positives in fusion predictions. (II) The inherent conflicts stem from diverse patterns of sensor signals between modalities. Specifically, multi-modal features exhibit asymmetric perceptual capabilities for different objects, influenced by factors such as object distance, illumination, weather conditions, occlusion situation, *etc.* Previous approaches expect that the modality with superior perceptual capabilities would dominate the fusion process. However, we reveal that excessively weak object confidence from the other modality can also hinder

correct results. As shown in Fig. 1(b), the distant and small pedestrians and traffic cones can be recalled by the camera due to rich image visual clues but are missed in LiDAR predictions, which are limited by sparse point structure. Such inherent conflicts from sensor signals lead to false negatives in the fusion predictions. Thence, cross-modal conflicts are non-negligible factors when utilizing multi-modal features to achieve accurate and robust detection.

In this paper, we propose an **Eliminating Conflicts Fusion (ECFusion)** method to avoid the degradation of perception capabilities caused by conflicts during fusion. *Firstly*, to eliminate extrinsic conflicts, we propose a **Semantic-guided Flow-based Alignment (SFA)** module, which aligns LiDAR and camera BEV features to the consistent distribution by utilizing spatial flow derived from semantic correspondence. Specifically, we first associate positions whose semantic information from class-aware heatmaps corresponds to the other modality. Then, the correspondence is converted into flow fields, which are employed to propagate BEV features for alignment. In this way, the fusion interference caused by extrinsic conflicts can be alleviated by alignment before fusion. *Secondly*, to eliminate inherent conflict, we introduce a **Dissolved Query Recovering (DQR)** mechanism, which aims at discovering object queries that are dissolved in the fusion heatmaps due to inherent conflicts and recover them from individual LiDAR and camera BEV heatmaps. Concretely, besides generating object queries from the fusion heatmap as previous method [3], we also explore potential single-modal object queries. We focus on the positions that exhibit high objectness inconsistently with fused features via a masked heatmap strategy. Our design is tailored to ensure maximum utilization of perceptual capability from single-modal features. Our contributions are summarized as follows:

- We investigate ignored cross-modal conflicts when fusing multi-modal features to the unified BEV space and how they hinder LiDAR-Camera 3D object detection.
- We propose the ECFusion method to eliminate conflicts between multi-modal BEV features, including a SFA module for spatial alignment before fusion and a DQR mechanism to recover useful object queries after fusion.
- Extensive experiments show our method achieves state-of-the-art performance for LiDAR-Camera 3D object detection on the nuScenes dataset.

II. RELATED WORK

A. Camera-based 3D Object Detection

Camera-based 3D object detection aims to estimate the 3D bounding boxes of objects from 2D images. Methods [4], [5], [6], [7] focus on predicting 3D objects from monocular images by extending 2D detectors to 3D by regressing the depth and orientation of objects. Other methods [8], [9], [10], [11] adapt the query-based end-to-end paradigm of DETR [12] to 3D space with 3D position queries and perspective projection. Methods [13], [14], [15], [16], [17], [18] construct BEV feature from multi-view images via interaction with predefined grid-shaped BEV queries or view transformation [19] to lift image features into voxel features.

B. LiDAR-based 3D Object Detection

LiDAR-based 3D object detection aims to localize 3D objects from point clouds scanned by LiDAR sensors. Point-based methods [20], [21], [22], [23], [24], [25], [26] directly process point clouds without discretization or projection. They preserve fine-grained geometric information but face challenges such as irregularity and sparsity of point clouds, high computational complexity, and memory consumption. Voxel-based methods [27], [28], [29], [30], [31], [32] convert point clouds into regular 3D/2D grids and encode to features by efficient 3D/2D convolution. They can reduce computational costs and memory consumption by sparsifying the voxel space but also introducing quantization errors and information loss. Range view-based methods [33], [31] project 3D point clouds onto a spherical surface and encode attributes into pixel values. Range view is more compact and faster to process, but also includes some challenges such as scale variation, occlusion, and coordinate inconsistency.

C. LiDAR-Camera Fusion-based 3D Object Detection

Some methods [34], [35], [36], [37], [38], [39], [40], [41], [42] decorate point cloud features with the corresponding image information (*e.g.*, RGB values, context features, semantic categories) by mapping instructive clues from camera space to LiDAR space. Alternatively, some methods [43], [44] project point clouds from LiDAR space to camera space and render images analogous to RGB-D data. However, these methods are mainly limited to the expression form of a particular modality and inevitably lose the information of another modality. Recently, some methods [1], [2], [3], [45], [46], [47], [48], [49] attempt to fuse multi-modal features in the shared BEV space. Unlike the fusion methods above, they adopt two independent branches to encode inputs from the camera and LiDAR sensors into modality-specific features and unify multi-modal features into the BEV representation to fuse. Such a process preserves both the geometric structure from point clouds and semantic information from images.

While existing BEV-based fusion methods have achieved remarkable strides, they still struggle with cross-modal conflicts in the BEV space. As a consequence, their performances are inferior even compared to single-modal detectors in some scenarios. To bridge these conflicts, we delve into exploring how specific conflicts disrupt the fusion process and pursue more fusion profit by eliminating these conflicts.

III. METHODOLOGY

As shown in Fig. 4, our ECFusion method starts by utilizing LiDAR and camera BEV feature extraction branches to produce modal-specific BEV features from individual modalities. Then, we leverage the multi-modal BEV feature fusion branch to integrate LiDAR and camera BEV features, forming a unified Fusion BEV feature. In the fusion branch, we propose a **Semantic-guided Flow-based Alignment (SFA)** module to first mitigate spatial distribution disparities (*i.e.*, extrinsic conflict) between LiDAR and camera BEV features before fusing them. Then, based on the LiDAR, camera, and fusion BEV features, we devise a **Dissolved Query**

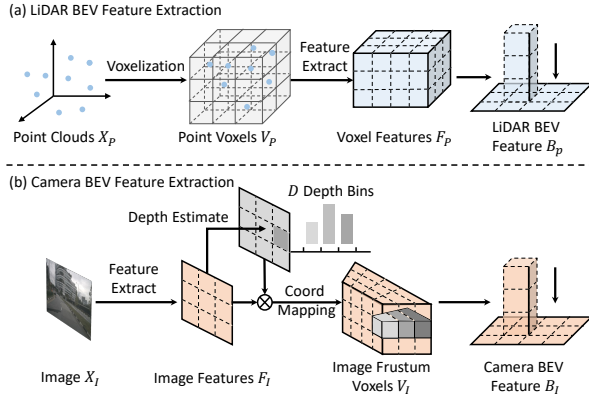


Fig. 2. The LiDAR and camera BEV feature extraction process.

Recovering (DQR) mechanism to generate comprehensive object queries. Concretely, the DQR mechanism aims to recover the dissolved object queries caused by asymmetric perceptual capabilities (*i.e.*, inherent conflict) between two modalities from single-modal features. Finally, a transformer decoder is utilized to predict the final 3D bounding boxes based on the derived object queries.

A. Single-Modal BEV Feature Extraction

The details of LiDAR and Camera BEV feature extraction branches are shown in Fig. 2, in which LiDAR and Camera BEV features are produced separately.

LiDAR BEV Feature Extraction. Given the input point clouds \mathbf{X}_P , as shown in Fig. 2(a), we first partition them into regular voxels $\mathbf{V}_P \in \mathbb{R}^{X_v \times Y_v \times Z_v}$ and use voxel encoder with 3D sparse convolution [27] to extract features $\mathbf{F}_P \in \mathbb{R}^{X \times Y \times Z \times C}$, where (X, Y, Z) denotes the size of the 3D voxel grids. Then we project \mathbf{F}_P to BEV along the Z -axis and adopt several 2D convolutional Layers to obtain the LiDAR BEV feature map $\mathbf{B}_P \in \mathbb{R}^{X \times Y \times C}$.

Camera BEV Feature Extraction. Given the input N_v view images \mathbf{X}_I , as shown in Fig. 2(b), we first extract features $\mathbf{F}_I \in \mathbb{R}^{N_v \times H \times W \times C}$ by image encoder, where (H, W) denotes the size of the image feature map. To construct the camera BEV feature map $\mathbf{B}_I \in \mathbb{R}^{X \times Y \times C}$, we apply $2D \rightarrow 3D$ view transform to image features of each view by Lift-Splat-Shoot (LSS) module [19]. Finally, The features within the same BEV grid are aggregated by BEVPool [1] operation.

B. Semantic-guided Flow-based Alignment

In the multi-modal BEV feature fusion branch, we aim to combine the LiDAR BEV features \mathbf{B}_P and camera BEV features \mathbf{B}_I to construct fusion BEV features \mathbf{B}_F . However, previous methods [1], [2] neglect the extrinsic conflicts between the two BEV features, *i.e.*, inconsistent spatial/semantic distribution patterns, and directly concatenate the two BEV features as fusion results. Note that the extrinsic conflicts are caused by the difference in origin signals coordinate, BEV feature extraction pipeline, and projection process. For example, projecting image features to BEV space requires imposed monocular depth estimation, which inevitably predicts inaccurate object depth. As a result, with such an inaccurate object depth, the projected camera BEV feature will contain

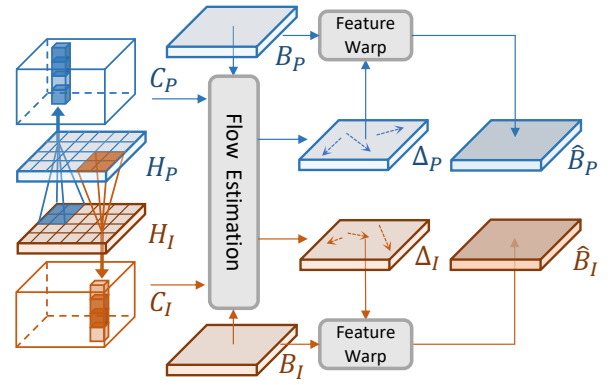


Fig. 3. Semantic-guided flow-based alignment module.

non-existent/redundant objects in the wrong positions, forming obvious misalignment with the LiDAR BEV feature.

Therefore, we elaborate a **Semantic-guided Flow-based Alignment (SFA)** module to align LiDAR and camera BEV features to obtain a consistent spatial distribution before fusion. Inspired by the optical flow methods [50], [51], we remedy the spatial discrepancy by applying appropriate flow transformation in inconsistent areas.

Specifically, as shown in Fig. 3, we start by establishing spatial correspondence between two modalities. Since LiDAR and camera BEV features $\mathbf{B}_P, \mathbf{B}_I$ are generated from two individual/heterogeneous branches, it is impractical to directly establish the correspondence between \mathbf{B}_P and \mathbf{B}_I . Thus we leverage the normalized LiDAR and camera BEV heatmaps $\mathbf{H}_P, \mathbf{H}_I \in \mathbb{R}^{X \times Y \times N_c}$, where N_c is the number of object categories, to capture the pixel-wise spatial correspondence. Technically, we construct the spatial correspondence for each pixel based on the cross-modal semantic similarity in the $q \times q$ neighborhood.

We first obtain cross-modal cost volumes $\mathbf{C}_P, \mathbf{C}_I \in \mathbb{R}^{X \times Y \times q^2}$ from $\mathbf{H}_P, \mathbf{H}_I$, which can be formulated as:

$$\begin{cases} \mathbf{C}_P(i, j) = \bigcup_{m \in [-q/2, q/2], n \in [-q/2, q/2]} \mathbf{H}_P(i, j)^T \mathbf{H}_I(i+m, j+n), \\ \mathbf{C}_I(i, j) = \bigcup_{m \in [-q/2, q/2], n \in [-q/2, q/2]} \mathbf{H}_I(i, j)^T \mathbf{H}_P(i+m, j+n). \end{cases} \quad (1)$$

Then we use a lightweight convolution block to estimate the flow fields $\Delta_P, \Delta_I \in \mathbb{R}^{X \times Y \times 2}$, which serve as spatial correspondence between modalities:

$$\{\Delta_P, \Delta_I\} = \text{Conv}(\text{Concat}(\mathbf{B}_P, \mathbf{C}_P, \mathbf{B}_I, \mathbf{C}_I)). \quad (2)$$

Next, we adopt the differentiable bilinear sampling operation to warp features based on $\{\Delta_P, \Delta_I\}$, which linearly interpolates the features of neighbors around the warped position. Formally, the aligned BEV feature $\hat{\mathbf{B}}_P, \hat{\mathbf{B}}_I$ are obtained as:

$$\hat{\mathbf{B}}_z(p) = \text{Interp}(\mathbf{B}_z(p + \Delta_z(p))), \quad z \in \{P, I\}, \quad (3)$$

where $\text{Interp}(\cdot)$ denotes neighborhood bilinear interpolation. Then we fuse them as $\mathbf{B}_F = \text{Conv}(\text{Concat}(\hat{\mathbf{B}}_P, \hat{\mathbf{B}}_I))$. Benefiting from the flow-based spatial alignment before fusion, it avoids uncoordinated features caused by extrinsic conflicts.

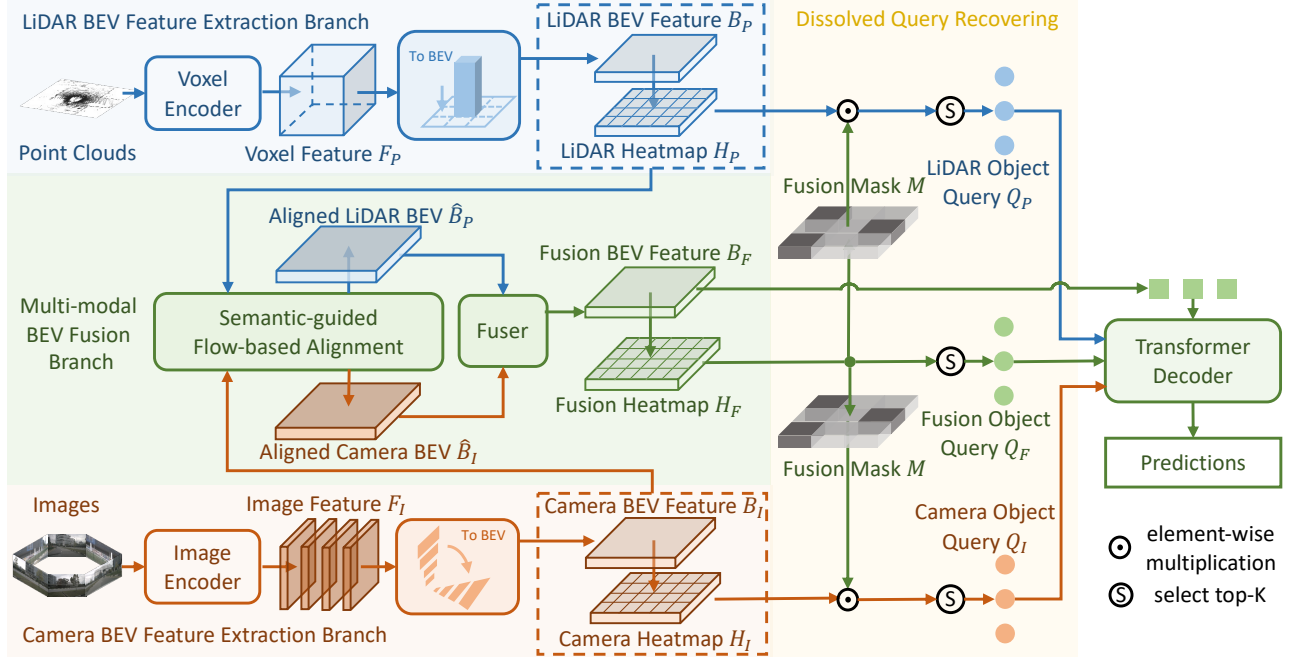


Fig. 4. Overview of our framework. Given inputs of point clouds and multi-view images: (I) We process them by individual LiDAR/Camera Feature Extraction Branch to obtain modal-special BEV features $\mathbf{H}_P, \mathbf{H}_I$. (II) We utilize Multi-modal BEV Fusion Branch to align \mathbf{H}_P and \mathbf{H}_I by SFA module and integrate them into unified fusion BEV features \mathbf{H}_F . (III) We generate fusion object queries \mathbf{Q}_F and modal-special object queries $\mathbf{Q}_P, \mathbf{Q}_I$ by DQR mechanism. Finally, all queries are aggregated together to predict 3D bounding boxes through a transformer decoder.

C. Dissolved Query Recovering Mechanism

Based on the fusion BEV feature \mathbf{B}_F , following [3], most previous methods directly generate class-specific fusion heatmap $\mathbf{H}_F \in \mathbb{R}^{X \times Y \times N_c}$ and select the Top- K_F local maximum candidate indexes. The information of selected candidates is used to initialize the context feature and position embedding of object queries $\mathbf{Q}_F \in \mathbb{R}^{K_F \times C}$, which are used to aggregate relevant context and prediction box parameters through DETR-style decoder layers [12].

Therefore, ensuring the high quality of initial queries is crucial to accurate detection since the object is unlikely to be recalled if there is no corresponding query. Theoretically, we expect fusion queries \mathbf{Q}_F should integrate all valuable objectness clues from both point clouds and image modalities, enabling them to inherit the unique detection ability of each modality. However, we found that in current method $\{GT_P \cup GT_I\} \not\subset GT_F$ as shown in Fig. 5, which means numerous objects in $\{\widetilde{GT}_P \cup \widetilde{GT}_I\}$ that are not matched by fusion queries can be recalled by modal-special queries. Our investigation shows that although the fusion strategy learned by current methods can indeed identify new objects, it would sacrifice a non-negligible portion of the single-modal detection ability. Therefore, it is pivotal to investigate how to maintain the single-modal detection capability while leveraging the cross-modal complementary.

Hence, we propose a **Dissolved Query Recovering (DQR)** mechanism to explicitly maintain the single-modal detection capability. Our fundamental concept revolves around the exploration of queries that are dissolved in the fusion heatmap due to conflicts but can be recovered from single-modal heatmaps. Further, we comprehensively integrate

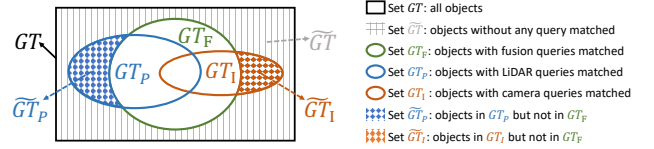


Fig. 5. Analysis of queries set from LiDAR, camera, and fusion heatmaps. GT means ground-truth objects. Note that our DQR aims at recovering queries to match with \widetilde{GT}_P and \widetilde{GT}_I .

multi-source queries for high recall.

Concretely, In addition to queries \mathbf{Q}_F from \mathbf{H}_F , we also generate modal-specific queries $\mathbf{Q}_P, \mathbf{Q}_I$ from single-modal heatmaps $\mathbf{H}_P, \mathbf{H}_I$ as complement. We commence by generating a fusion mask M based on the fusion query positions \mathcal{P}_F . By applying this mask to single-modal heatmaps, we circumvent redundant query generation for the same object, thereby focusing on the exploration of distinctive object positions $\mathcal{P}_P, \mathcal{P}_I$ compared to the fusion heatmap as below:

$$\begin{cases} \mathcal{P}_F = \operatorname{argmax}(\mathbf{H}_F, K_F), \\ \mathbf{M}[p] = \begin{cases} 0 & \text{if } p \in \mathcal{P}_F, \\ 1 & \text{else,} \end{cases} \\ \mathcal{P}_z = \operatorname{argmax}(\mathbf{H}_z \odot M, K_z) \quad z \in \{P, I\}, \end{cases} \quad (4)$$

where \odot represents element-wise multiplication. After obtaining multiple sets of query positions, we sample the corresponding BEV features with respective position embedding to construct the unified queries \mathbf{Q}_u as:

$$\mathbf{Q}_u = \bigcup_{z \in \{F, P, I\}} \{\mathbf{Q}_z\}, \quad \mathbf{Q}_z = \mathbf{B}_z[\mathcal{P}_z] + \operatorname{PE}(\mathcal{P}_z), \quad (5)$$

where $\operatorname{PE}(\cdot)$ means position encoding function. Then in the transformer decoder, we adopt the self-attention between

queries for relations between different objects and the cross-attention between queries and fusion features to aggregate relevant context. Finally, a feed-forward network (FFN) is utilized for predicting 3D bounding boxes $\{b_i\}_{i=1}^{K_F+K_P+K_I}$ according to the aggregated query features.

D. Model optimization

During training, the matching cost and loss function as [1] are applied. In detail, detection loss L_D for all bounding box predictions is computed to optimize, including classification loss for all results and box regression loss for only positive pairs with Ground-Truth (GT) boxes matched. Besides, we adopt focal losses $\{L_{H_z}|z \in F, P, I\}$ for three heatmap predictions with the gaussian-distributed GT map generated by GT boxes center [26]. Noted that the GT map for L_P, L_I should be masked by fusion mask M synchronously. The total loss is $L = L_{det} + \sum_{z \in \{F, P, I\}} L_{H_z}$.

IV. EXPERIMENTS

A. Datasets and Metrics

We validate the effectiveness of our method on the large-scale nuScenes dataset[52], which is currently the most popular benchmark for multi-modal 3D object detection. nuScenes contains 700 sequences for training, 150 sequences for validation, and 150 sequences for testing. Each sequence is approximately 20 seconds long, including annotated 3D bounding boxes from 10 categories of sampled key frames. Each sample consists of point clouds from 32-beam LiDAR scans and 6 surround-view 1600×900 resolution images that provide 360° horizontal FOV. For metrics, nuScenes provides the official evaluation metrics for the 3D detection task including mean Average Precision (mAP) and nuScenes Detection Score (NDS).

B. Implementation Details

We implement our single-modal BEV feature extraction branches following BEVFusion [1]. We set the image size to 448×800 as the official settings of the camera-based methods [16], [17] and the voxel size to 0.075m as the official settings of the LiDAR-based methods [26], [28]. Our training consists of two stages as BEVFusion: 1) We respectively train the LiDAR and camera branch with single-modal inputs. 2) We jointly train the entire model that loads weights from two pre-trained branches and frozen modal encoders. We set the query numbers to $K_F = 200, K_P = 50, K_I = 50$ for training and testing. We directly use predictions from a single model without any test-time augmentation or extra post-processing for both val and test results.

C. Comparison with State-of-the-Art Methods

We show the main results of nuScenes in Table I, our method surpasses all previous LiDAR-camera fusion methods and achieves the state-of-the-art performance of 73.4% NDS on *val* set and 73.9% NDS on *test* set. Our method has a significant improvement over LiDAR or camera methods. Moreover, By contrast with baseline method BEVFusion [1], our method exceeds it by a large margin with +2.2% mAP

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE nuSCENES validation AND test SET. WE REPORT METRICS MAP(%) AND NDS(%). MODALITY ‘L’ AND ‘C’ REPRESENT LiDAR AND CAMERA.

Method	Modality	validation		test	
		mAP↑	NDS↑	mAP↑	NDS↑
BEVDet4D [16]	C	42.1	54.5	45.1	56.9
BEVFormer [13]	C	-	-	48.1	56.9
Ego3RT [53]	C	47.8	53.4	42.5	47.9
PolarFormer [15]	C	50.0	56.2	49.3	57.2
CenterPoint [26]	L	59.6	66.8	60.3	67.3
Focals Conv [41]	L	61.2	68.1	63.8	70.0
Transfusion-L [3]	L	65.1	70.1	65.5	70.2
LargeKernel3D [54]	L	63.3	69.1	65.3	70.5
FUTR3D [55]	L+C	64.5	68.3	-	-
PointAug [35]	L+C	-	-	66.8	71.0
MVP [37]	L+C	67.1	70.8	66.4	70.5
AutoAlign [46]	L+C	66.6	71.1	-	-
AutoAlignV2 [47]	L+C	67.1	71.2	68.4	72.4
TransFusion [3]	L+C	67.5	71.3	68.9	71.6
VFF [48]	L+C	-	-	68.4	72.4
BEVFusion [2]	L+C	67.9	71.0	69.2	71.8
UVTR [56]	L+C	65.4	70.2	67.1	71.1
Focals Conv-F [41]	L+C	-	-	67.8	71.8
BEVFusion [1]	L+C	68.5	71.4	70.2	72.9
DeepInteraction [57]	L+C	69.9	72.6	70.8	73.4
ECFusion(Our)	L+C	70.7	73.4	71.5	73.9

TABLE II
ABLATION STUDIES ON THE MULTI-MODAL.

LiDAR	Camera	mAP↑	NDS↑
✓		64.6	69.2
	✓	36.1	35.9
✓	✓	70.7	73.4

TABLE III
ABLATION STUDIES ON MAIN DESIGNS.

SFA	DQR	mAP↑	NDS↑
		68.4	71.7
✓		69.1	72.2
	✓	70.5	73.1
✓	✓	70.7	73.4

and +2.0% NDS on *val* set. We ascribe this performance gain to eliminating cross-modal conflicts in BEV space.

D. Ablation Studies

In this subsection, we present ablation studies and in-depth analyses of our designs. The mAP and NDS are evaluated on the nuScenes *val* set.

Multi Modalities. We verify the effectiveness of multi-modal fusion in Table II with improvements to LiDAR-based results (+6.1% mAP and +4.2% NDS) and camera-based results (+34.6% mAP and +37.5% NDS).

Main Contributions. To illustrate our improvements for fusion, we conduct ablation experiments on our two primary components SFA and DQR in Table III. As a result, our SFA module can improve the baseline by +0.7% mAP and +0.5% NDS. Furthermore, the method with the DQR mechanism demonstrates significant improvements, elevating

TABLE IV

ABLATION STUDIES ON THE SFA MODULE. Δ_P AND Δ_I ARE FLOWS FOR LiDAR AND CAMERA. C_P AND C_I DENOTE THE COST VOLUME.

	Δ_P	Δ_I	C_P	C_I	mAP \uparrow	NDS \uparrow
#1					68.4	71.7
#2	✓		✓		68.9	72.1
#3		✓		✓	69.3	72.4
#4	✓	✓	✓	✓	69.5	72.8
#5	✓	✓			69.1	72.2

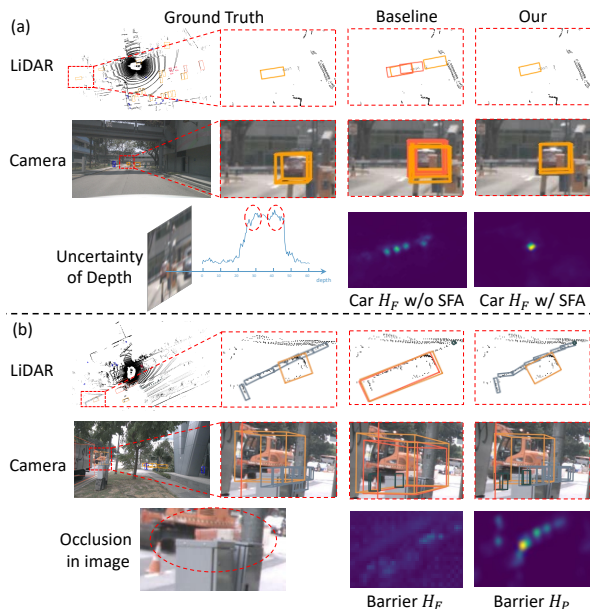


Fig. 6. Visualization of comparing our method with baseline on nuScenes val set. We show ground truth boxes on LiDAR and camera space respectively. For the baseline and our method, we show the predictions on LiDAR and camera space. Furthermore, we visualize the BEV heatmaps constructed by different methods.

the baseline by +2.1% mAP and +1.4% NDS. Meanwhile, the simultaneous utilization of our two primary design components synergistically enhances performance, resulting in a combined improvement of +2.3% mAP and +1.7% NDS.

Qualitative Results. To gain more insight into our improvements, we compare the qualitative results in Fig. 6. In sub-fig. (a), Caused by the extrinsic conflicts from inevitable uncertainty of depth estimation for the car at a long distance, fusion BEV heatmap H_F has redundant high activations of the Car category. In contrast, our heatmap can obtain accurate object position assistance by the alignment of the SFA module. Moreover, we show a case in sub-fig. (b) when the detection capability of the camera for these barriers is lost due to the occlusion, fusion BEV heatmap H_F of baseline loses valuable objectness of barriers due to inherent conflicts. Meanwhile, our method can recover object queries from LiDAR heatmap H_P to retain the right predictions.

SFA Module. To precisely demonstrate the superiority of our aligned modality-specific BEV features, we ablate the design of our SFA module as shown in Table IV. Comparisons between #2-#4 and #1 show that applying flow-based alignment to any modality can bring improvements, *i.e.*, +0.4% NDS with LiDAR BEV flow Δ_P and +0.7% NDS

TABLE V

ABLATION STUDIES ON THE DQR MECHANISM. Q_P AND Q_I ARE RECOVERED QUERIES FROM LiDAR AND CAMERA. M DENOTES OUR FUSION MASK. K_F DENOTES THE FUSION QUERY NUMBER.

	Q_P	Q_I	M	K_F	mAP \uparrow	NDS \uparrow
#1				200	68.4	71.7
#2	✓		✓		70.1	72.6
#3		✓	✓	200	69.6	72.3
#4	✓	✓	✓		70.5	73.1
#5	✓	✓		200	68.0	71.5
#6				300	68.7	71.8

with camera BEV flow Δ_I . Both utilization can also further achieve significant +1.1% NDS improvements. Moreover, we show in #5 compared with #4 that without semantic guidance represented by cost volume C_P, C_I results in performance degradation -0.6% NDS. Overall, our approach estimates multi-modal flows to facilitate cross-modal consistency that leads to improved fusion BEV features.

DQR Mechanism. We validate the efficacy of our DQR in Table V. Comparisons #2-#4 with #1 reveal that results with recovering dissolved queries from single-modal heatmaps significantly outperform the baseline. For instance, LiDAR queries Q_P bring +0.9% NDS, and camera queries Q_I bring +0.6% NDS, while using both achieve an even higher gain of +1.4% NDS. To confirm the necessity of fusion mask M , we conduct experiment #5, which independently generates three groups of queries from heatmaps without M . The results as -0.2% NDS drop compared with #1 highlight the necessity of our mask design. Furthermore, simply increasing the number of queries from 200 to 300 in experiment #6 yields only marginal improvements, indicating that our approach effectively discovers and recovers lost objects due to cross-modal conflicts, rather than using more queries.

V. CONCLUSION

In this work, we present a novel approach to eliminate widely prevalent yet long-overlooked cross-modal conflicts during the fusion process for LiDAR-Camera 3D object detection. This key idea is to align various spatial distributions before fusion and recover lost object information after fusion. To achieve that, we introduce an SFA module to exploit cross-modal correspondence and align multi-modal features via a spatial flow field. Moreover, we design a DQR mechanism to persevere object queries that are obscured in fusion features due to conflict interference. Experiments on the nuScenes dataset prove the effectiveness of our method.

VI. ACKNOWLEDGEMENT

This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (NO. 62122010, U23B2010), Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F02022F02, Key Research and Development Program of Zhejiang Province under Grant 2022C01082.

REFERENCES

- [1] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *ICRA*, 2023.
- [2] 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," in *NeurIPS*, 2022.
- [3] 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 *CVPR*, 2022.
- [4] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3d object detection," in *ICCV*, 2021.
- [5] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *CoRL*, 2022.
- [6] Z. Wu, Y. Gan, L. Wang, G. Chen, and J. Pu, "Monopgc: Monocular 3d object detection with pixel geometry contexts," in *ICRA*, 2023.
- [7] C. Gao, J. Chen, S. Liu, L. Wang, Q. Zhang, and Q. Wu, "Room-and-object aware knowledge reasoning for remote embodied referring expression," in *CVPR*, 2021.
- [8] Y. Wang, V. C. Guizilimi, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3d object detection from multi-view images via 3d-to-2d queries," in *CoRL*, 2022.
- [9] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3d object detection," in *ECCV*, 2022.
- [10] Y. Liu, J. Yan, F. Jia, S. Li, Q. Gao, T. Wang, X. Zhang, and J. Sun, "PETRv2: A unified framework for 3d perception from multi-camera images," *arXiv preprint arXiv:2206.01256*, 2022.
- [11] Z. Wang, Z. Huang, J. Fu, N. Wang, and S. Liu, "Object as query: Lifting any 2d object detector to 3d detection," in *ICCV*, 2023.
- [12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.
- [13] 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 *ECCV*, 2022.
- [14] 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.
- [15] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformers," in *AAAI*, 2023.
- [16] J. Huang, G. Huang, Z. Zhu, Y. Yun, and D. Du, "BEVDet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.
- [17] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3d object detection," in *AAAI*, 2023.
- [18] E. Xie, Z. Yu, D. Zhou, J. Phillion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," *arXiv preprint arXiv:2204.05088*, 2022.
- [19] J. Phillion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *ECCV*, 2020.
- [20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.
- [21] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, 2017.
- [22] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *CVPR*, 2019.
- [23] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3d object detection," in *CVPR*, 2020.
- [24] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *ICCV*, 2019.
- [25] S. Shi, X. Wang, and H. Li, "Part-a² net: 3d part-aware and aggregation neural network for object detection from point cloud," in *CVPR*, 2019.
- [26] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3d object detection and tracking," in *CVPR*, 2021.
- [27] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, 2018.
- [28] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019.
- [29] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *CVPR*, 2018.
- [30] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018.
- [31] L. Fan, X. Xiong, F. Wang, N. Wang, and Z. Zhang, "Rangedet: In defense of range view for lidar-based 3d object detection," in *ICCV*, 2021.
- [32] X. Wang and K. M. Kitani, "Cost-aware evaluation and model scaling for lidar-based 3d object detection," in *ICRA*, 2023.
- [33] Z. Liang, M. Zhang, Z. Zhang, X. Zhao, and S. Pu, "Rangercnn: Towards fast and accurate 3d object detection with range image representation," *arXiv preprint arXiv:2009.00206*, 2020.
- [34] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *CVPR*, 2020.
- [35] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *CVPR*, 2021.
- [36] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *ECCV*, 2020.
- [37] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," in *NeurIPS*, 2021.
- [38] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *CVPR*, 2019.
- [39] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *CVPR*, 2022, pp. 5418–5427.
- [40] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Trans. Multimedia*, 2022.
- [41] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3d object detection," in *CVPR*, 2022.
- [42] P. Jacobson, Y. Zhou, W. Zhan, M. Tomizuka, and M. C. Wu, "Center feature fusion: Selective multi-sensor fusion of center-based objects," in *ICRA*, 2023.
- [43] H. Li, Z. Zhang, X. Zhao, Y. Wang, Y. Shen, S. Pu, and H. Mao, "Enhancing multi-modal features using local self-attention for 3d object detection," in *ECCV*, 2022.
- [44] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation," in *ICCV*, 2021.
- [45] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *CVPR*, 2020.
- [46] 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 *IJCAI*, 2022.
- [47] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection," in *ECCV*, 2022.
- [48] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, and J. Jia, "Voxel field fusion for 3d object detection," in *CVPR*, 2022.
- [49] Y. Gao, C. Sima, S. Shi, S. Di, S. Liu, and H. Li, "Sparse dense fusion for 3d object detection," in *IROS*, 2023.
- [50] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, 2017.
- [51] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, and Y. Tong, "Semantic flow for fast and accurate scene parsing," in *ECCV*, 2020.
- [52] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
- [53] J. Lu, Z. Zhou, X. Zhu, H. Xu, and L. Zhang, "Learning ego 3d representation as ray tracing," in *ECCV*, 2022.
- [54] Y. Chen, J. Liu, X. Qi, X. Zhang, J. Sun, and J. Jia, "Scaling up kernels in 3d cnns," in *CVPR*, 2023.
- [55] 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*, 2022.
- [56] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," in *NeurIPS*, 2022.
- [57] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," in *NeurIPS*, 2022.