2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

EIMC:Efficient Instance-aware Multi-modal Collaborative Perception

Kang Yang[1], Peng Wang[1], Lantao Li[2], Tianci Bu[3], Chen Sun[2], Deying Li[1], Yongcai Wang[1*]

***Abstract*— Multi-modal collaborative perception calls for great attention to enhancing the safety of autonomous driving. However, current multi-modal approaches remain a "local fusion → communication" sequence, which fuses multi-modal data locally and needs high bandwidth to transmit an individual's feature data before collaborative fusion. EIMC innovatively proposes an early collaborative paradigm. It injects lightweight collaborative voxels, transmitted by neighbor agents, into the ego's local modality-fusion step, yielding compact yet informative 3D collaborative priors that tighten cross-modal alignment. Next, a heatmap-driven consensus protocol identifies exactly where cooperation is needed by computing per-pixel confidence heatmaps. Only the Top-K instance vectors located in these low-confidence, high-discrepancy regions are queried from peers, then fused via cross-attention for completion. Afterwards, we apply a refinement fusion that involves collecting the top-K most confident instances from each agent and enhancing their features using self-attention. The above instance-centric messaging reduces redundancy while guaranteeing that critical occluded objects are recovered. Evaluated on OPV2V and DAIR-V2X, EIMC attains 73.01% AP@0.5 while reducing byte bandwidth usage by 87.98% compared with the best published multi-modal collaborative detector. Code publicly released at `https://github.com/sidiangongyuan/EIMC`.**

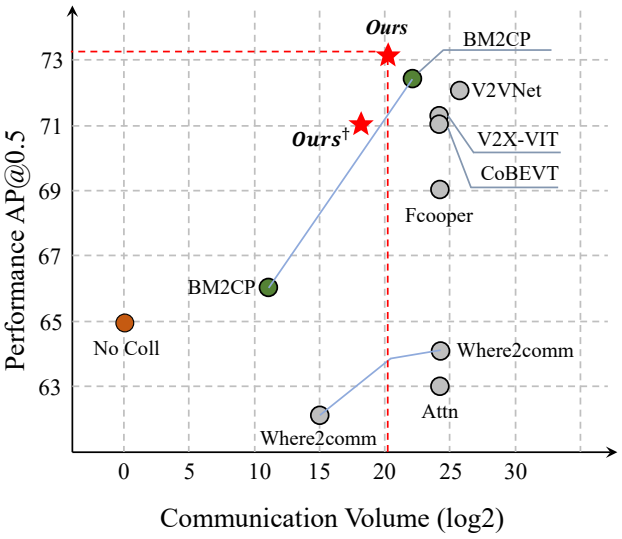


Fig. 1. Compared with other intermediate fusion methods, EIMC achieves lower communication volume while still attaining the best performance. The Ours† variant represents the version without the Mix-Voxel module. BM2CP [3] is the multimodal-based collaborative perception work.

I. INTRODUCTION

Precise 3D scene perception is critical for intelligent agents in autonomous driving, drone technology, and robotics, enabling accurate situational awareness and real-time decision-making vital for operational safety. However, individual agent perceptual capabilities are inherently limited by restricted sensing ranges and occlusion effects, impeding comprehensive scene understanding. For instance, a single vehicle's sensors in autonomous driving may fail to detect occluded objects or long-range threats. These limitations necessitate a more holistic approach that aggregates and leverages information across multiple vantage points. Consequently, researchers have turned to collaborative perception, allowing agents to share complementary information for a more accurate and robust perception of their surroundings , thereby addressing occlusion and extending effective sensing ranges [1], [2].

Current collaborative perception methodologies confront two critical challenges: 1) The multi-modal frameworks that achieved remarkable success in single-agent systems have not been sufficiently explored or effectively adapted for collaborative perception tasks; 2) The inherent communication overhead introduced by collaborative perception poses significant challenges to meet the stringent real-time requirements of autonomous driving systems. Early approaches [4], [5], [6], [7] prioritized performance improvements while neglecting communication bandwidth constraints, resulting in limited practicality. Subsequent methods [8], [9], [10] attempted to balance communication efficiency with performance metrics, yet struggle to simultaneously maintain high performance and reduce communication demands. Such methods, for instance, have explored utilizing confidence maps for information filtering, but frequently encounter difficulties in accurately identifying all regions containing potential objects, and often still rely on transmitting dense BEV features, thereby hindering effective communication cost reduction. Recently, BM2CP [3] pioneered the exploration of multi-modal collaborative perception frameworks, demonstrating the practicality of applying such approaches in this field. However, there is still room for optimization in multimodal fusion and balancing communication overhead and system performance. Overall, the most critical problem is how to optimally select essential perceptual information and enhance cross-agent knowledge transfer while ensuring

*Corresponding author: Yongcai Wang
[1]School of Information, Renmin University of China, Beijing, China, 100872
[2]Sony Research and Development Center China, Beijing, China
[3]National University of Defense Technology, Hunan, China, 410073

communication efficiency.

In this work, we propose EIMC, a multimodal collaborative perception framework that jointly optimizes fusion accuracy and communication efficiency. EIMC advances collaboration by first injecting lightweight cross-agent voxels into the ego's modality-fusion step, creating compact, discriminative 3D collaborative priors to enhance cross-modal alignment. Building upon this, a heatmap-driven consensus protocol precisely identifies cooperation needs, allowing only critical instance vectors from low-confidence, high-discrepancy regions to be queried from peers and fused for completion and refinement. This instance-centric messaging reduces redundancy while recovering occluded objects.

Specifically, the Modality Fusion stage employs the Mix-Voxel module to create a collaborative geometric prior from aggregated LiDAR voxels, which an occupancy head uses to re-weight the ego camera voxel before BEV feature collapse. The Heterogeneous Modality Fusion module then aligns the resulting camera and LiDAR BEV features in latent space. In the Collaboration stage, the Instance Completion module locates uncertain regions via heatmap discrepancies to retrieve complementary instance vectors. Instance Refinement further refines these vectors through self-attention and integrates them into the BEV representation via cross-attention. Furthermore, a multi-scale approach bolsters cooperative perception for diverse scenarios. As illustrated in Fig. 1, EIMC achieves robust and efficient multimodal cooperative perception, effectively balancing performance and communication overhead.

To summarize, our contributions are:

- EIMC offers an efficient yet robust multimodal-based collaborative perception framework, enhancing 3D perception performance while reducing communication overhead.
- The Mix-Voxel and HMF modules construct compact and expressive scene-level representations, addressing modality gaps and alignment issues. The Instance Completion and Refinement modules focus on essential instance-level message transmission and fusion, yielding more robust and holistic scene understanding.
- Extensive experiments on 3D detection benchmarks demonstrate that EIMC consistently outperforms existing methods in accuracy, communication efficiency, and robustness.

II. Related Work

A. Collaborative perception

Collaborative perception aims to enhance agents' perceptual capabilities by sharing information across a communication network. Mainstream research in this domain typically utilized single-modality sensors, such as LiDAR or cameras, for 3D object detection. These methods can be broadly categorized into early, intermediate, and late fusion approaches. Early fusion usually involves transmitting raw sensor data [11], [12], but requires substantial bandwidth, while late fusion, which shares network outputs, often fails to deliver optimal performance. In contrast, recent studies have turned to intermediate fusion techniques, seeking a better balance between efficiency and accuracy [4], [7], [8], [5], [13], [9], [14], [**?**]. For instance, V2VNet [4] employs a spatially-aware graph neural network (GNN) to aggregate features across multiple agents, and AttnFuse [6] is the first to introduce an attention mechanism for modeling multi-agent interactions. Moreover, V2X-VIT [5] adapts vision transformers (ViTs) [15] for vehicle-to-everything (V2X) communication using heterogeneous self-attention, while Where2comm [8] and CoSDH [16] focuses on determining the optimal fusion points to minimize communication bandwidth. Additionally, DiscoNet [7] leverages knowledge distillation to combine the advantages of both early and intermediate fusion methods, and CoBEVT [13] presents the first generic collaborative perception framework for multi-camera-based cooperative BEV semantic segmentation. BM2CP [3] pioneers the exploration of multimodal cooperative perception tasks. Building upon these foundations, instance-aware methodologies have emerged through works like TransIFF's [9], [**?**] transformer-based feature fusion and QUEST's [14] interpretable query cooperation. Beyond core perception tasks, subsequent research addresses practical deployment challenges including pose errors [17], latency constraints [18], [19], and multi-agent alignment [20], [21]. Recent studies, such as HEAL [22], HM-VIT [23], CodeFilling [24] and STAMP [25], focus on ensuring compatibility across heterogeneous agents.

B. Multi-modal 3D Detection

LiDAR and image data provide complementary information, enriching 3D scene understanding by combining precise geometric details with rich semantic cues. The integration of these modalities for 3D detection is an active area of research [26], [27]. One line of work is early fusion, exemplified by methods such as PointPainting and PointAugmenting [28], [29], which augment LiDAR point clouds with features from camera images. Beyond early fusion, modern multimodal fusion strategies can be broadly divided into two categories. The first category uses a shared bird's-eye-view (BEV) representation to fuse dense features from both LiDAR and camera modalities [30], [31], [32], [33], [34]. For example, BEVFusion [30], [31] leverages the Lift-Splat-Shoot (LSS) backbone [35] to project image features into the BEV space and then concatenates them with LiDAR features, while DeepInteraction [32] keeps each modality's features separate to enable efficient cross-modal interaction. The second category relies on sparse object queries to implicitly integrate features from both modalities [36], [37], [38], [39], [40]. TransFusion [39] illustrates this approach by introducing a query-based fusion strategy with a soft-association mechanism to handle poor image conditions, while Futr3D [38] presents a unified query-driven fusion framework. In this paper, we present EIMC, a novel and highly robust and efficient framework for multimodal collaborative perception that not only minimizes communication overhead but also sets a new benchmark in performance.

III. METHOD

A. Problem Formulation

The architecture of the EIMC is depicted in Fig. 2. In this scenario, we consider N agents, and a communication graph G models the interactions between all agents as vertices. Let $\mathbf{I}_n$ represent the RGB image collected by the camera, which may be captured by surrounding cameras, and let $\mathbf{L}_n$ denote the point cloud collected by the LiDAR of the n-th agent. Additionally, let $\mathcal{M}$ represent the message sent from neighboring agents to the ego agent $\mathcal{E}$, and $\mathbf{Y}_\mathcal{E}$ represent the perception supervision for the ego agent. The objective of collaborative perception is to achieve the maximized perception performance of all agents while hoping the communication cost is limited, that is:

$$\begin{aligned} \arg\max_{\theta,\mathcal{M}} h\left(\Phi_\theta\left(\mathcal{M}_\mathcal{E}, \{\mathcal{M}_{n\to\mathcal{E}}\}_{n=1}^{N\neq\mathcal{E}}\right), \mathbf{Y}_\mathcal{E}\right), \\ \text{s.t.} \sum_{i=1}^{N} |\mathcal{M}_{i\to j}| \leq B, \end{aligned} \tag{1}$$

where $h(\cdot,\cdot)$ is the perception evaluation metric, Φ_θ is the perception network with trainable parameter θ. The process of our framework can be divided into two stages:

$$\begin{cases} \mathbf{B}_{\text{MF}}^n = f_{\text{MF}}(f_{\text{enc}}(\mathbf{I}_n, \mathbf{L}_n)), n = 1, \cdots, N \quad \text{(stage-1)} \\ \hat{\mathbf{Y}}_\mathcal{E} = f_{\text{dec}}\left(f_{\text{Col}}\left(\mathbf{B}_\mathcal{E}, \{\mathcal{M}_{n\to\mathcal{E}}\}_{n\neq\mathcal{E}}^{N}\right)\right) \text{(stage-2)} \end{cases} \tag{2}$$

Here, f_{enc}, f_{MF}, f_{dec}, and f_{Col} denote the encoder, Modality Fusion stage, decoder, and collaboration stage, respectively. $\mathbf{B}_{\text{MF}}$ is the result of the modality fusion process, and $\hat{\mathbf{Y}}_\mathcal{E}$ represents the detection results.

B. Modality Encoding

1) LiDAR branch.: Given the point clouds $\{\mathbf{P}_n\}_{n=1}^N$, we follow the mainstream approaches [8], [4], [6] and apply $f_{\text{point}}(\cdot)$ using PointPillars and VoxelNet. Formally, the LiDAR BEV feature $\mathbf{B}_\text{L}$ and the voxel feature $\mathbf{V}_\text{L}$ are derived as follows:

$$\begin{aligned} \mathbf{B}_\text{L} &= f_{\text{point}}(\mathbf{P}_n) \in \mathbb{R}^{N\times H\times W\times C_\text{L}}, \\ \mathbf{V}_\text{L} &= f_{\text{voxel}}(\mathbf{P}_n) \in \mathbb{R}^{N\times H\times W\times L\times C_\text{L}}, \end{aligned} \tag{3}$$

where (H, W, L) and C_L represent voxel size and feature channel of LiDAR. Furthermore, following the common depth supervision methods in single vehicle [41], [42], we obtain a sparse depth map from the point clouds and use depth completion [43] to supervise the depth predicted by the camera branch.

2) Camera branch.: Given the images $\{\mathbf{I}_n\}_{n=1}^N$, we generate image features $\mathbf{F}_\text{I} \in \mathbb{R}^{N\times H\times W\times C_I}$ using a standard image encoder [44]. We then use a Depth Net [35], [42], [41] to predict the depth distribution interval $\mathbf{D} \in \mathbb{R}^{N\times H\times W\times L}$. Next, we compute the outer product between $\mathbf{F}_\text{I}$ and $\mathbf{D}$, and apply intrinsic and extrinsic parameters for the view transformation, resulting in the image voxel representation $\mathbf{V}_\text{I} \in \mathbb{R}^{N\times H\times W\times L\times C_\text{I}}$.

C. Modality Fusion

To bridge distributional and spatial discrepancies between LiDAR and camera features, we propose the Occ-Guided Image Voxel Representation and the Heterogeneous Modality Fusion (HMF) module.

1) Occ-Guided Image Voxel Representation.: This component leverages LiDAR-based occupancy to enhance camera depth estimation. As shown in Fig. 3, the Mix-Voxel module facilitates shared information by aligning all voxels to the ego-agent coordinate system and constructing a local graph. Self-attention on this graph models inter-agent voxel interactions to capture representative features:

$$\mathbf{V}_{\text{mix}} = f_{\text{Mix-Voxel}}(\mathbf{V}_\text{L}^{\text{rc}}, \mathbf{V}_\text{L}^{\text{sd}_1}, \cdots, \mathbf{V}_\text{L}^{\text{sd}_n}). \tag{4}$$

Here, $\mathbf{V}_\text{L}^{\text{rc}}$ and $\mathbf{V}_\text{L}^{\text{sd}}$ represent the voxel associated with the ego agent and the neighboring agents, respectively. After that, $\mathbf{V}_{\text{mix}}$ is transmitted to Φ_{Occhead} to generate the occupancy probability of the 3D scene.

$$\mathbf{O}_\text{L} = \Phi_{\text{Occhead}}(\mathbf{V}_{\text{mix}}), \in \mathbb{R}^{N\times H\times W\times L\times 1}. \tag{5}$$

$\mathbf{O}_\text{L}$ shares the same resolution as $\mathbf{V}_\text{I}$. The camera voxel $\mathbf{V}_\text{I}$ is then multiplied by $\mathbf{O}_\text{L}$ to obtain the occ-guided image voxel representation:

$$\mathbf{V}_\text{I}' = \mathbf{V}_\text{I} \odot \mathbf{O}_\text{L}. \tag{6}$$

Where $\odot$ denotes element-wise multiplication operation. To effectively aggregate geometric and depth information to $\mathbf{V}_\text{I}'$, we add $\mathbf{V}_{\text{mix}}$ to $\mathbf{V}_\text{I}'$. Finally, we obtain the camera BEV feature $\mathbf{B}_\text{I}$ through voxel collapse.

2) Heterogeneous Modality Fusion.: Unlike conventional methods that simply concatenate BEV features , our HMF module (Fig. 4) effectively aggregates semantic and geometric information in a unified BEV space. The detailed architecture is illustrated in Fig. 4. We use 1×1 convolutions to expand $\mathbf{B}_\text{L}$ and $\mathbf{B}_\text{I}$ to appropriate channels C, then concatenate the expanded BEV features to obtain feature $\mathbf{B}_{\text{cat}}$. Additionally, we employ attention mechanisms to facilitate interaction between the LiDAR and image BEV features. Since the LiDAR BEV feature is more reliable than the camera's, we define $\mathbf{B}_\text{L}$ as the query. Formally, it can be expressed as:

$$\mathbf{B}_{\text{attn}} = \text{MLP}(\text{softmax}(\frac{\mathbf{B}_\text{L}\cdot\mathbf{B}_\text{I}}{\sqrt{C}})\cdot\mathbf{B}_\text{I}) + \mathbf{B}_\text{L} \tag{7}$$

The final fused BEV features can be presented as:

$$\mathbf{B}_{\text{fus}} = \mathbf{B}_{\text{cat}} + \mathbf{B}_{\text{attn}}, \in \mathbb{R}^{N\times H\times W\times C}. \tag{8}$$

D. Collaboration

After obtaining $\mathbf{B}_{\text{fus}}$, we focus on how to transmit messages effectively and efficiently. Existing collaborative perception paradigms often suffer from prohibitive bandwidth costs due to dense BEV feature transmission. TTransIFF [9] introduces sparse instances to reduce communication bandwidth, but it does not achieve optimal performance and is not suitable for multimodal scenarios. In this work, we propose a novel heatmap-driven instance-level communication

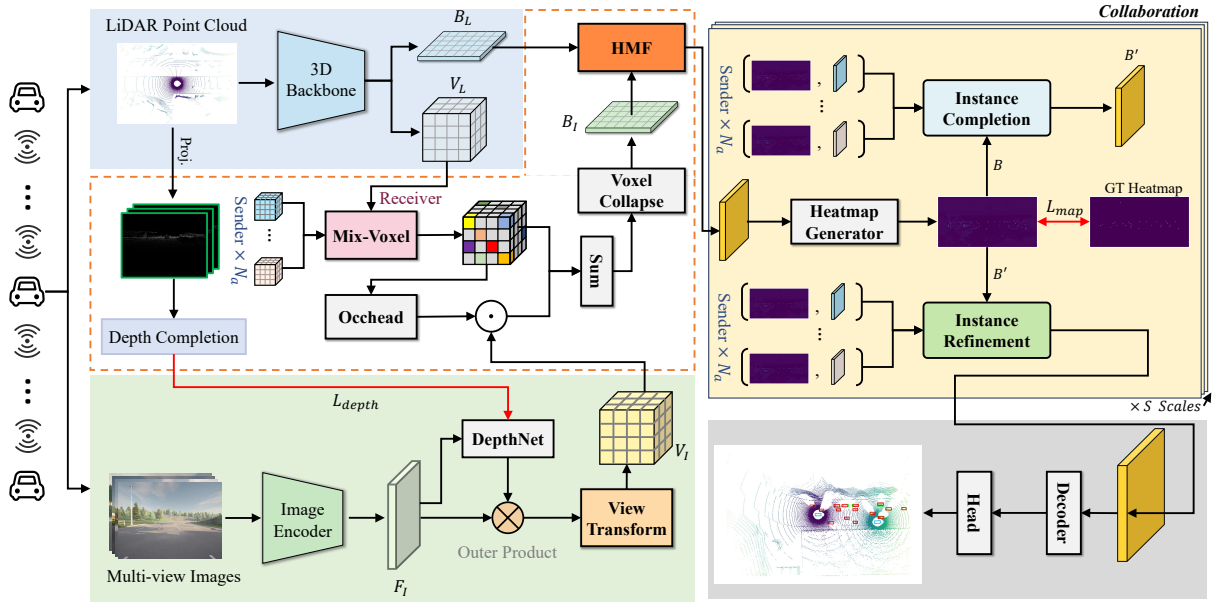


Fig. 2. Framework. Given LiDAR and camera inputs, our method first extracts heterogeneous features through dedicated modality-specific encoders. The Mix-Voxel (MV) module leverages lightweight voxel transmission as priors to build the collaborative voxel and then constructs occupancy-guided voxel-based image representations, which are compressed into BEV features and fused with LiDAR BEV features through Heterogeneous Modality Fusion (HMF). Instance Completion (IC) and Instance Refinement (IR) modules subsequently propagate instance-level messages identified from heatmap priors. The collaboration employs multi-scale feature for final detection, with predicted and ground truth bounding boxes visualized as green and red boxes respectively.

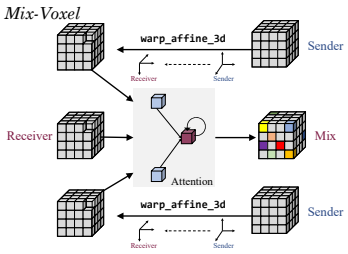


Fig. 3. **Mix-Voxel module** constructs a local graph of voxels, utilizing self-attention mechanisms to facilitate information exchange.

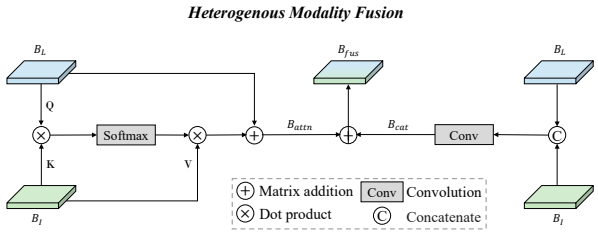


Fig. 4. **Heterogeneous Modality Fusion module**. The module consists of two streams. The left part uses an attention-based method to establish the interaction between the two modalities, while the right part employs a basic operation (e.g., concatenation) followed by convolutional layers to fuse the modalities.

strategy, comprising two modules, Instance Completion and Instance Refinement, to balance performance and bandwidth, as illustrated in Figure 5.

1) Instance Completion.: The core objective of this module is to determine what complementary perceptual message the ego agent (receiver) requires from collaborating agents (sender). Specifically, for each agent pair (receiver rc and one sender sd), we generate heatmaps $\mathbf{H}_{\text{rc}}$ and $\mathbf{H}_{\text{sd}} \in \mathbb{R}^{H\times W\times 1}$ through a lightweight CNN Φ_{hm}. The target discrepancy heatmap is computed as:

$$\mathbf{H}_{\text{tg}} = \mathbf{H}_{\text{rc}} - \mathbf{H}_{\text{sd}} \in \mathbb{R}^{(N-1)\times H\times W\times 1}. \tag{9}$$

We then identify the $\mathbf{K}_{\text{IC}}$ spatial positions with minimal values in $\mathbf{H}_{\text{tg}}$, indicating regions where the receiver's perception is least confident compared to the sender. Then, for each selected position (h, w), we extract receiver's BEV feature $\mathbf{F}_{\text{rc}}^{(h,w)} \in \mathbb{R}^C$ as query $\mathbf{Q}$ and sender's BEV feature $\mathbf{F}_{\text{sd}}^{(h,w)}$ as key/value $(\mathbf{K}, \mathbf{V})$. These features undergo cross-attention:

$$\mathbf{F}_{rc}^{\text{updated}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V}. \tag{10}$$

When multiple senders update the same spatial position in $\mathbf{B}_{\text{rc}}$, we employ element-wise summation for feature fusion:

$$\mathbf{B}_{rc}^{(h,w)} \leftarrow \sum_{s=1}^{N_{\text{senders}}} \mathbf{F}_{rc,s}^{\text{updated}} \tag{11}$$

The insight behind this module is that the key aspect of collaborative perception lies in the transmission of complementary information. This complementary information is reflected in the heatmap as areas where the ego has low confidence in the presence of an object, while the sender exhibits high confidence. As a result, subtracting the sender's heatmap from the ego's heatmap highlights the areas of interest for the ego. Through the heatmap-driven approach,

the most relevant and critical regions can be effectively identified and prioritized for further processing.

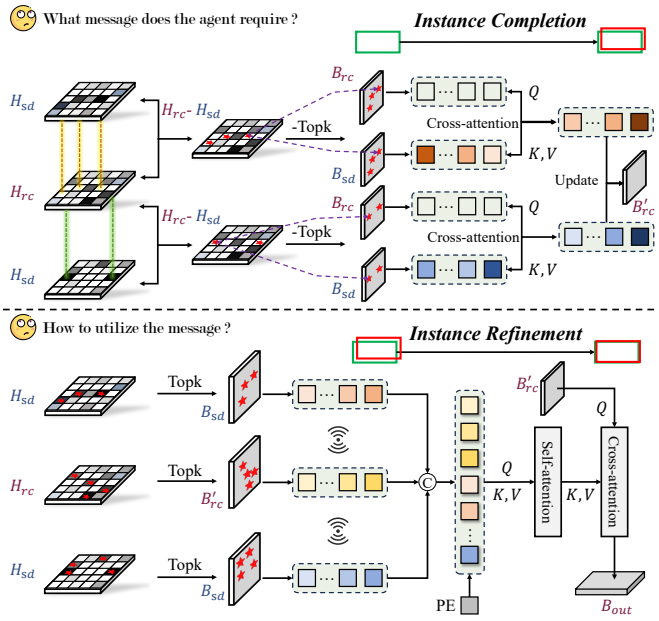


Fig. 5. **Instance-level communication**. The Instance Completion (IC) module prioritizes critical regions by analyzing cross-agent heatmap discrepancies, performing instance completion via cross-attention. The IR module first selects agent-specific instances from heatmaps, then refines them via self-attention. Finally, it aggregates instance-to-scene context by cross-attending BEV features (query) to instance representations (key/value).

2) Instance Refinement.: Following instance completion, this module aims to holistically refine the potential instances through attention mechanisms. First, we extract Top-k critical region (h, w) from heatmap $\mathbf{H}_n \in \mathbb{R}^{H\times W}$ for each agent. Then we retrieve instance features $\mathbf{F}_n \in \mathbb{R}^{\mathbf{K}_{\text{IR}}\times C}$ from corresponding BEV features $\mathbf{B}_n$, at position (h, w). We concatenate all instances:

$$\mathbf{F}_{\text{all}} = [\mathbf{F}_{\text{rc}}; \mathbf{F}_{\text{sd}_1}; \cdots; \mathbf{F}_{\text{sd}_\mathbf{N}}] \in \mathbb{R}^{N\cdot\mathbf{K}_{\text{IR}}\times C}. \tag{12}$$

To enable information exchange from a data-driven perspective, we employ self-attention for inter-instance communication:

$$\mathbf{F}_{\text{all}}^{\text{updated}} = \text{SelfAttn}(\mathbf{F}_{\text{all}}, \mathbf{F}_{\text{all}}, \mathbf{F}_{\text{all}}), \tag{13}$$

where $\mathbf{F}_{\text{all}}$ is addition with positional encoding $E_{\text{pos}}(h, w)$ maintaining spatial relationships. Finally, cross-attention enables each BEV grid feature to acquire valuable information from potentially relevant instances:

$$\mathbf{B}_{\text{out}} = \text{CrossAttn}(Q_{\text{scence}}, K_{\text{ins}}, V_{\text{ins}}), \tag{14}$$

$$Q_{\text{scence}} = \mathbf{B}'_{\text{rc}} \in \mathbb{R}^{H\times W\times C}, \tag{15}$$

$$K_{\text{ins}}, V_{\text{ins}} = \mathbf{F}_{\text{all}}^{\text{updated}}. \tag{16}$$

This module enables adaptive message passing between instances via self-attention and context-aware scene reconstruction through instance-to-BEV cross-attention with minimal communication bandwidth.

To capture instances and scenes at varying granularities and enhance the information exchange between agents, we downsample the initial feature map to generate feature maps at different scales for use in the collaboration stage. Finally, the feature maps from all scales are fused to produce an output feature map that matches the dimensions of the original input.

E. Loss Function

Finally, the training loss of the model is simply the sum of the regression loss L_{reg}, classification loss L_{cls}, depth estimation loss L_{depth}, direction loss L_{dir} and heatmap loss L_{hm}:

$$L_{\text{total}} = \lambda_{\text{reg}}L_{\text{reg}} + \lambda_{\text{cls}}L_{\text{cls}} + \lambda_{\text{depth}}L_{\text{depth}} + \lambda_{\text{dir}}L_{\text{dir}} + \lambda_{\text{hm}}L_{\text{hm}}. \tag{17}$$

Where the λ are the weighting factors of the different losses used in the optimization process.

IV. Experimental Results

TABLE I

Comparison of mainstream works on the OPV2V and DAIR-V2X dataset. The best performance is highlighted in **bold**. Ours† indicates the model without the Mix-Voxel module.

Method	OPV2V			DAIR-V2X			Bandwidth
	AP30	AP50	AP70	AP30	AP50	AP70	Comm ($\log_2$)
No Coll	83.63	63.74	58.32	69.99	65.02	53.82	0.00
Fcooper	93.88	89.03	74.28	76.61	69.29	51.37	24.00
Attn	88.08	86.30	75.32	68.78	63.16	49.17	24.00
V2VNet	93.56	93.13	89.00	77.36	72.22	52.95	25.43
V2X-VIT	95.09	93.66	86.06	77.29	71.87	55.46	24.00
CoBEVT	94.54	93.03	84.64	77.86	71.70	55.85	24.00
Where2comm	88.36	86.77	76.34	68.18	63.16	51.04	24.00
BM2CP	93.34	93.04	88.94	**77.91**	72.37	56.18	23.18
Ours†	93.59	92.27	83.96	76.82	71.81	56.76	18.97
Ours	**95.29**	**94.71**	**89.16**	75.01	**73.01**	**58.37**	20.16

A. Performance

As evidenced in Table I, our EIMC framework establishes new state-of-the-art detection performance across both benchmark datasets. On the OPV2V dataset, EIMC achieves superior results with 95.29% AP30, 94.71% AP50, and 89.16% AP70, outperforming existing approaches by margins of +0.20%, +1.58%, and +0.16%, respectively, in these metrics. The framework exhibits particularly strong robustness on the DAIR-V2X dataset, attaining 58.37% AP70, a 2.19% absolute improvement over the previous best method (BM2CP: 56.18%). This significant advancement at the strictest IoU threshold (0.7) demonstrates EIMC's ability to ensure reliable perception under real-world challenges. As shown in the table, EIMC effectively balances communication overhead and performance. Following the [8], the communication volume is computed as follows:

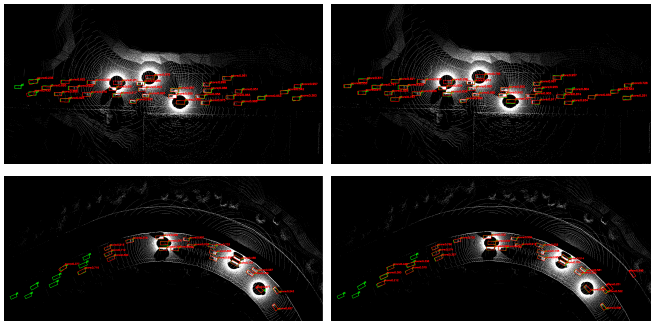
$$Comm(B) = \log_2(H \times W \times C \times \text{float32}/8). \tag{18}$$

Fig. 6 shows the comparison with BM2CP and EIMC. EIMC achieves more complete and accurate detection. To evaluate the effectiveness of our collaborative perception

TABLE II

ROBUSTNESS ANALYSIS UNDER POSE NOISE ON OPV2V AND DAIR-V2X. THE POSE NOISE FOLLOWS A GAUSSIAN DISTRIBUTION.

Method	AP50 under Noise Level (σ_p/σ_r)								AP70 under Noise Level (σ_p/σ_r)							
	OPV2V				DAIR-V2X				OPV2V				DAIR-V2X			
	0/0	0.2/0.2	0.4/0.4	0.6/0.6	0/0	0.2/0.2	0.4/0.4	0.6/0.6	0/0	0.2/0.2	0.4/0.4	0.6/0.6	0/0	0.2/0.2	0.4/0.4	0.6/0.6
No Coll	63.74	-	-	-	65.02	-	-	-	58.32	-	-	-	53.82	-	-	-
Fcooper	89.03	80.77	66.53	59.30	69.29	67.43	64.18	62.25	74.28	62.62	51.56	47.29	51.37	50.18	48.46	47.56
Attn	88.08	86.31	80.57	77.98	63.16	57.32	54.19	52.06	75.32	73.33	69.09	60.23	49.17	45.77	43.17	42.75
V2VNet	93.13	91.04	74.16	61.45	72.32	68.76	62.21	58.48	89.00	67.33	37.96	28.55	52.95	46.97	43.00	40.85
V2X-VIT	93.66	91.87	86.29	80.33	71.87	69.13	64.86	61.95	86.06	77.63	65.64	60.04	55.46	52.47	50.44	49.17
CoBEVT	93.03	91.14	84.29	76.27	71.70	69.26	64.75	62.46	84.64	77.69	65.28	57.41	55.85	53.44	51.12	50.14
Where2comm	86.77	80.34	76.57	72.19	63.16	63.12	62.47	60.14	76.34	68.73	60.32	54.77	51.04	51.04	51.03	51.02
BM2CP	93.04	91.34	81.55	73.54	72.17	**70.31**	**65.07**	61.78	88.94	75.35	57.43	49.53	56.18	53.24	49.97	48.65
Ours	**94.71**	**92.44**	**89.78**	**87.87**	**73.01**	68.23	64.86	**63.82**	**89.16**	**80.60**	**76.33**	**73.98**	**58.37**	**55.18**	**52.91**	**52.21**



BM2CP Ours

Fig. 6. Visualization of predictions from BM2CP and EIMC on OPV2V dataset.

method, we compared the performance of different LiDAR-only methods, as shown in Table III. Our method achieves competitive results, with AP50 reaching 68.46% and AP70 reaching 54.02%. Note that the performance of LiDAR-only methods may outperform that of multimodal methods due to the following reasons: 1) DAIR-V2X is a real-world dataset, which introduces sensor alignment noise; 2) Unlike the OPV2V dataset, DAIR-V2X only provides front-view cameras for vehicles and roadside cameras, limiting the camera branch's ability to obtain surround-view BEV features.

TABLE III

LIDAR-ONLY PERFORMANCE COMPARISON OF DIFFERENT METHODS ON DAIR-V2X DATASET.

Method	AP30	AP50	AP70
No Coll	65.26	59.88	48.52
Fcooper	73.10	65.36	44.52
Attn	69.85	64.33	51.12
V2VNet	66.92	53.60	49.83
V2X-VIT	67.14	63.30	49.34
CoBEVT	**75.07**	67.70	47.09
Where2comm	69.48	64.46	51.77
Ours	73.22	**68.47**	**54.02**

B. Ablation

Components analysis. Our ablation studies on DAIR-V2X, as shown in Table IV, demonstrate critical insights:

TABLE IV

ABLATION STUDIES OF COMPONENTS ON THE DAIR-V2X DATASET.

MV	IC	IR	MS	AP30	AP50	AP70	#Params (M)
-	-	-	-	70.25	67.73	50.91	∼25.9
✓	-	-	-	71.02	68.39	51.44	∼26.1
✓	✓	-	-	74.12	69.34	54.78	∼26.8
✓	-	✓	-	71.77	68.24	52.09	∼28.0
-	✓	✓	✓	**76.82**	71.81	56.76	∼42.0
✓	✓	✓	-	74.22	70.54	56.28	∼30.9
✓	✓	✓	✓	75.01	**73.01**	**58.37**	∼42.0

the IC module drives the most substantial performance gain, boosting AP70 by +3.34% (51.44% → 54.78%) through targeted cross-agent feature retrieval. The MV module significantly enhances perception accuracy. Ultimately, the full framework with MS integration achieves optimal results (58.37% AP70, 42M parameters), highlighting the synergistic benefits of our design.

TABLE V

ABLATION STUDY OF TOP-K SELECTION STRATEGIES ON DAIR-V2X DATASET.

Instance Completion				Instance Refinement			
$\mathbf{K}_{\text{IC}}$	AP30	AP50	AP70	$\mathbf{K}_{\text{IR}}$	AP30	AP50	AP70
10	73.46	68.69	53.51	200/100/50	**75.93**	71.19	56.52
15	73.28	68.16	52.68	150/100/50	75.09	70.82	56.72
20	75.01	**73.01**	**58.37**	100/50/25	75.01	**73.01**	**58.37**
25	**77.78**	72.15	54.96	100/100/100	74.34	69.84	54.70
30	75.71	70.68	54.57	50/50/50	72.39	68.96	54.12

1) Analysis on Parameters.: As shown in Table V, we perform the analysis on the number of Top-k selection. It can be observed that if $\mathbf{K}_{\text{IC}}$ is too small, it may fail to capture all potential key regions adequately, whereas an excessively large $\mathbf{K}_{\text{IC}}$ may reduce the accuracy of the bounding box. Moreover, dynamically adjusting the number of $\mathbf{K}_{\text{IR}}$ according to scale variations can improve detection performance. We analyze how detection accuracy (AP) varies with the number of agents (Fig. 7). While most methods initially improve with more agents, gains saturate beyond 3–4 agents due to scene coverage limits. Our framework effectively enhances multi-agent cooperation, achieving clear accuracy improvements as agent numbers increase within

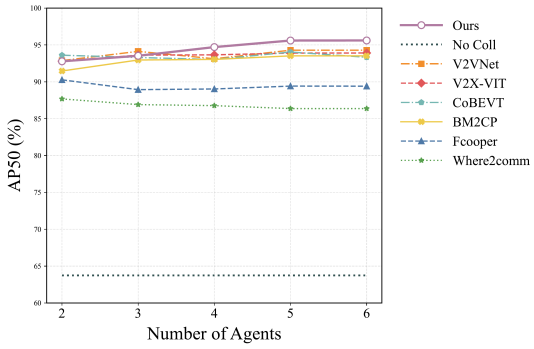


Fig. 7. Average Precision at IoU=0.5 with respect to agent number. Note that all methods were trained using the default agent setting of 4, and the results were obtained through inference with varying agent numbers. With the exception of CoBEVT, which, due to the inherent limitations of the method, requires retraining based on the number of agents.

environmental constraints.

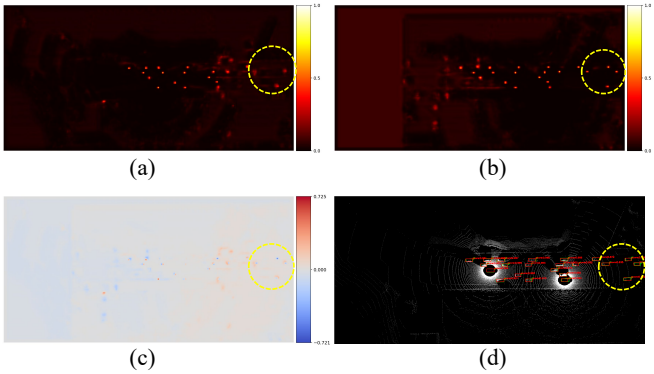


Fig. 8. **Visualization of heatmap and result.** (a) Ego agent heatmap; (b) Communication agent heatmap; (c) Target heatmap; (d) Output. The area within the yellow circle highlights regions of interest.

2) Analysis of Heatmap.: The Instance Completion module mainly addresses the occlusion problem that cannot be resolved by a single agent. As shown in Fig. 8, (a) depicts the ego agent's heatmap, (b) shows the heatmap of a neighboring agent, and (c) represents the target heatmap. It is clearly observed that, within the yellow region, the blue instances are highlighted. These instances are occluded and have low confidence from the ego's perspective. Through the Instance Completion module, these instances can ultimately be reconstructed and detected.

3) Robustness to localization noise.: We also evaluate the robustness to localization noise following the setting in [8]. The results are shown in Table II. For each noise level, we compare the AP performance of EIMC at IOU thresholds of 0.5 and 0.7 with several other methods. As shown in the table, most existing methods exhibit a noticeable decrease in performance as noise levels increase at an IOU of 0.7. In contrast, EIMC demonstrates remarkable robustness, maintaining high performance even under severe noise conditions. Furthermore, under the most extreme noise conditions, in the OPV2V dataset, EIMC outperforms the best-performing method by 9.3% at an IOU of 0.5 and by 22.8% at an IOU of 0.7. In the DAIR-V2X dataset, the performance improvements are 2.1% and 2.3%, respectively.

TABLE VI

ABLATION STUDIES OF HMF MODULE.

Strategy	**AP30**	**AP50**	**AP70**
Attn	74.18	68.80	53.94
Attn+Concat	75.01	**73.01**	**58.37**
Attn+Add	**75.23**	70.94	58.06

TABLE VII

ABLATION STUDIES OF MV MODULE.

Method	AP30	AP50	AP70	Bandwidth($\log_2$)
M1	75.01	73.01	58.37	20.16
M2	74.22	72.35	56.96	19.67
M3	**76.24**	**73.55**	**59.17**	22.46

4) HMF module.: We performed an ablation study on the HMF module. As shown in Table VI, relying solely on the attention mechanism is insufficient for fully aligning heterogeneous modalities, whereas integrating the concatenation operation yields optimal performance.

5) Mix-Voxel module: Table VII details our ablation study on the Mix-Voxel module's compression. We evaluated three strategies: an aggressive 4× downsampling (M2), a mild 2× downsampling (M3), and an asymmetric approach (M1). The results show that M3 achieves the highest accuracy but with substantial communication overhead, while M2 harms performance. Consequently, we select M1, which strikes the best balance between performance and efficiency.

V. CONCLUSIONS

By distilling sparse key instances from dense BEV features and leveraging dual completion and refinement mechanisms, the EIMC framework achieves an effective balance between communication efficiency and perception performance, while also demonstrating robust resilience to noise. In future work, we will evaluate the effectiveness of our method across additional datasets.

REFERENCES

[1] Y. Guo and J. Ma, "Leveraging existing high-occupancy vehicle lanes for mixed-autonomy traffic management with emerging connected automated vehicle applications," *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 1375–1399, 2020.

[2] J. Ma, E. Leslie, A. Ghiasi, Z. Huang, and Y. Guo, "Empirical analysis of a freeway bundled connected-and-automated vehicle application using experimental data," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 6, p. 04020034, 2020.

[3] B. Zhao, W. Zhang, and Z. Zou, "Bm2cp: Efficient collaborative perception with lidar-camera modalities," *arXiv preprint arXiv:2310.14702*, 2023.

[4] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[5] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," 2022.

[6] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

[7] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[8] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022.

[9] Z. Chen, Y. Shi, and J. Jia, "Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 205–18 214.

[10] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6876–6883.

[11] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2020.

[12] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524.

[13] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," 2022.

[14] S. Fan, H. Yu, W. Yang, J. Yuan, and Z. Nie, "Quest: Query stream for practical cooperative perception," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 18 436–18 442.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] J. Xu, Y. Zhang, Z. Cai, and D. Huang, "Cosdh: communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6834–6843.

[17] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *Conference on Robot Learning*. PMLR, 2021, pp. 1195–1210.

[18] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *European Conference on Computer Vision*. Springer, 2022, pp. 316–332.

[19] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, 2023.

[20] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4812–4818.

[21] Z. Huang, S. Wang, Y. Wang, W. Li, D. Li, and L. Wang, "Roco: Robust cooperative perception by iterative object matching and pose adjustment," in *ACM Multimedia 2024*, 2024.

[22] Y. Lu, Y. Hu, Y. Zhong, D. Wang, S. Chen, and Y. Wang, "An extensible framework for open heterogeneous collaborative perception," *arXiv preprint arXiv:2401.13964*, 2024.

[23] H. Xiang, R. Xu, and J. Ma, "Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 284–295.

[24] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 481–15 490.

[25] X. Gao, R. Xu, J. Li, Z. Wang, Z. Fan, and Z. Tu, "Stamp: Scalable task and model-agnostic collaborative perception," *arXiv preprint arXiv:2501.18616*, 2025.

[26] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.

[27] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[28] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[29] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 794–11 803.

[30] 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," 2022.

[31] 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," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

[32] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1992–2005, 2022.

[33] X. Li, B. Fan, J. Tian, and H. Fan, "Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 209–21 218.

[34] Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, "Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 643–21 652.

[35] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," 2020.

[36] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[37] Z. Zhou and S. Tulsiani, "Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 588–12 597.

[38] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 172–181.

[39] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[40] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez, "Focalformer3d: focusing on hard instance for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8394–8405.

[41] 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," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1477–1485, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25233

[42] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8551–8560.

[43] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the cpu," in *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 2018, pp. 16–22.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.