HM-ViT: Hetero-modal Vehicle-to-Vehicle Cooperative Perception with Vision Transformer

Hao Xiang[1], Runsheng Xu[1], Jiaqi Ma[1]
[1]University of California, Los Angeles
{haxiang, rxx3386, jiaqima}@g.ucla.edu

Abstract

Vehicle-to-Vehicle technologies have enabled autonomous vehicles to share information to see through occlusions, greatly enhancing perception performance. Nevertheless, existing works all focused on homogeneous traffic where vehicles are equipped with the same type of sensors, which significantly hampers the scale of collaboration and benefit of cross-modality interactions. In this paper, we investigate the multi-agent hetero-modal cooperative perception problem where agents may have distinct sensor modalities. We present HM-ViT, the first unified multi-agent hetero-modal cooperative perception framework that can collaboratively predict 3D objects for highly dynamic Vehicle-to-Vehicle (V2V) collaborations with varying numbers and types of agents. To effectively fuse features from multi-view images and LiDAR point clouds, we design a novel heterogeneous 3D graph transformer to jointly reason inter-agent and intra-agent interactions. The extensive experiments on the V2V perception dataset OPV2V demonstrate that the HM-ViT outperforms SOTA cooperative perception methods for V2V hetero-modal cooperative perception. Our code will be released at https://github.com/XHwind/HM-ViT.

1. Introduction

Recent advances in Vehicle-to-Vehicle (V2V) communication technology and intelligent transportation systems [13, 31, 27, 14, 7, 19, 28, 51, 53, 52] have allowed autonomous vehicles (AVs) to share sensory information, enabling them to perceive their surroundings better [3, 32, 54]. With the rapid growth of autonomous driving, V2V perception systems have the potential to be deployed at scale and create a safer transportation system. Cooperative perception systems, as shown in recent studies [49, 48, 40], can intelligently aggregate features from multiple vehicles within the communication range to enhance visual reasoning and overall performance.

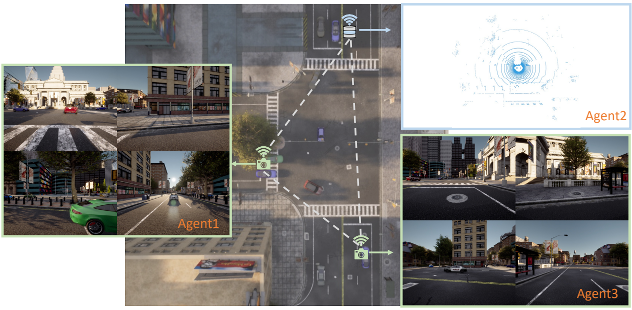


Figure 1: Illustration of multi-agent hetero-modal V2V systems where each agent may be equipped with either LiDAR or multi-view cameras.

Despite the rapid growth in this field, previous studies [49, 40, 46, 22, 37, 47] have primarily focused on homogeneous multi-agent cooperative perception, where all agents are equipped with the same type of sensors. In reality, however, agents may have different sensor modalities (hetero-modality) due to the cost and sensor preferences of ADS developers and car makers. As shown in Fig. 1, some agents are equipped with only LiDARs (LiDAR agents), while others only have multiple cameras (camera agents). Enabling collaboration between these heterogeneous agents could improve the sensing capability by allowing agents to see through occlusions and increase the scale and reliability of V2V systems. Additionally, LiDAR agents can provide accurate geometric information, while camera agents can provide rich semantic context. Thus, the collaboration between these agents could leverage the distinct but complementary environment attributes captured by each sensor modality to enhance the V2V perception systems. Furthermore, compared with the single-agent solution where multiple LiDARs and cameras are installed in a single vehicle, distributing different types of sensors across distinct agents could also potentially decrease the costs for each agent while still achieving satisfying performance. Nevertheless, whether, when, and how multi-agent hetero-modal V2V cooperation can benefit the perception system of heterogeneous traffic has not yet been studied.

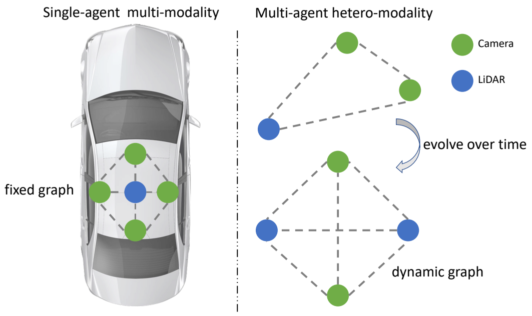


Figure 2: Comparison of the single-agent multi-modal system and multi-agent hetero-modal system. The graph structure of the former is fixed whereas for the latter the graph is both dynamic and heterogeneous.

In this work, we address the multi-agent hetero-modal cooperative perception problem where each agent could have distinct sensor types and share/receive information with each other. Notably, as shown in Fig. 2, this multi-agent hetero-modal setting is distinct from the single-agent multi-modal setting. In the hetero-modal setting, the agent sensors form a dynamic heterogeneous graph where the existence and types of sensors are random, and the relative poses vary from scene to scene. In contrast, the sensor types/numbers and relative positions (extrinsics) between sensors are fixed in the single-agent multi-modal setting. Existing multi-modal methods heavily rely on these assumptions, and most of existing works [38, 39, 2, 55, 50] transform LiDAR points or 3D proposals onto the image plane to index 2D features. Their network architectures build upon the co-existence of both LiDAR and camera inputs with fixed geometric relationships. However, the dynamic nature of hetero-modal V2V perception requires a flexible architecture that can handle varying agent numbers and types, and the transmitted neural features are also spatially misaligned. Moreover, there are semantic discrepancies in the transmitted features between camera agents and LiDAR agents. Hence, these unique characteristics pose significant challenges for designing the multi-agent hetero-modal cooperative system and prevent adapting existing multi-modal fusion methods to this new problem.

To enable collaboration between heterogeneous agents in V2V systems, we propose **H**etero-**M**odal **V**ision **T**ransformer (HM-ViT), the first unified cooperative perception framework that can leverage and fuse distributed information for hetero-modal V2V perception via a spatial-aware 3D heterogeneous vision transformer. Fig. 3 demonstrates the overall framework. Each agent first generates bird's eye view (BEV) representations through modality-specific encoders and then shares compressed features with neighboring agents. Afterward, the received features are decompressed and aggregated via the proposed HM-ViT, which conducts joint local and global heterogeneous 3D attentions with the consideration of both node and edge types. Our extensive experiments show that the HM-ViT can significantly improve the perception capability of camera agents and LiDAR agents over the single-agent baseline and outperforms SOTA cooperative perception methods by a large margin. In particular, for camera agents, performance can be boosted from 2.1% to 53.2% at AP@0.7 with the collaboration of LiDAR agents, a **23-fold** improvement. Our primary contributions can be summarized as follows:

- We present the novel transformer framework (HM-ViT) for multi-agent hetero-modal cooperative perception, capable of capturing the modality-specific characteristics and heterogeneous 3D interactions. The proposed model exhibits superior flexibility and robustness with state-of-the-art performance on highly dynamic heterogeneous traffic involving varying agent numbers and types.
- We propose a generic heterogeneous 3D graph attention (H^3GAT), tailored for extracting inter-agent and intra-agent heterogeneous interactions. We instantiate two such attentions – local attention (H^3GAT-L) and global attention (H^3GAT-G) for capturing both local and global visual cues.
- We conduct extensive benchmark experiments by varying sensor modalities, demonstrating the strong performance of the proposed method for hetero-modal V2V perception tasks. We will release all the codes and baselines to facilitate future research.

2. Related work

V2V perception. V2V perception aims to enhance the perception performance of autonomous vehicles by leveraging shared information from other connected vehicles. Existing works have primarily focused on LiDAR-based 3D object perception. The pioneer cooperative perception methods transmit raw sensing observation (*i.e.*, early fusion) or perception outputs (*i.e.*, late fusion), whereas recent works [49, 40, 16, 43, 21, 48] are exploring the use of circulating intermediate neural features for achieving better performance-bandwidth trade-off. V2VNet [40] employs graph neural networks to aggregate the shared neural features for joint detection and prediction. AttFuse [49] uses single-head attention to model the per-location multi-agent interaction. Disconet [22] presents a matrix-valued edge weight for learning the interactions and a teacher-student learning framework to facilitate the training. V2X-ViT explores [48] vision transformer for vehicle-to-everything cooperation via window attention and heterogeneous self-attention. CoBEVT [46] presents a generic transformer framework for camera-based BEV semantic segmentation.

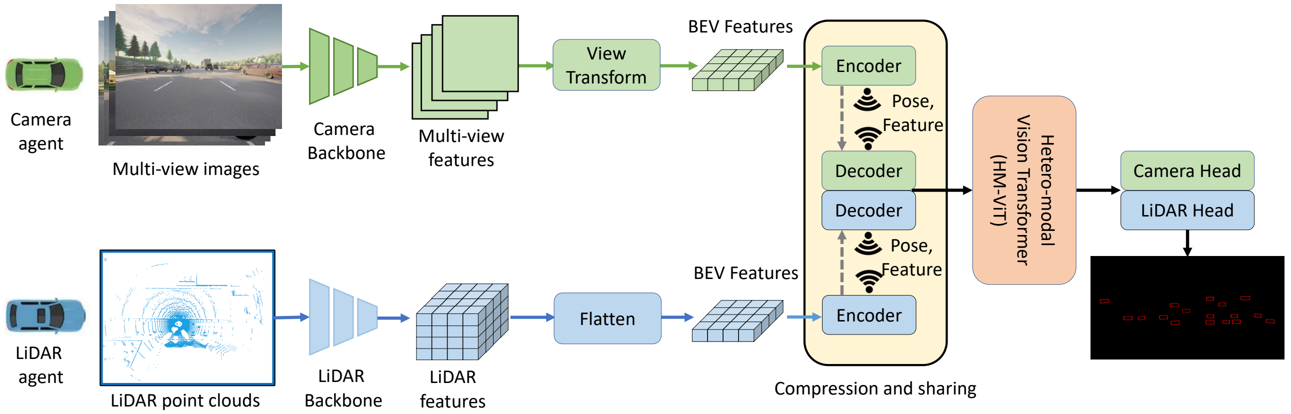


Figure 3: Overview of proposed hetero-modal V2V cooperative perception system. Each agent first produces BEV features through modality-specific feature extractors (Sec. 3.1). The BEV features are then compressed and shared (Sec. 3.3) with neighboring connected agents and the received features are decompressed in the ego agent side and fed into hetero-modal vision transformer to conduct graph-structured feature fusion (Sec. 3.2 and Sec. 3.3). The refined features are finally passed into the hetero-modal detection head to predict 3D bounding boxes (Sec. 3.4).

However, none of the existing works explored multi-agent multi-camera 3D object detection, let alone multi-agent hetero-modal perception. In contrast to existing methods, HM-ViT is the first to employ sparse heterogeneous local and global attentions to capture the 3D inter-agent and intra-agent interactions in a computationally efficient manner.

Camera-based 3D object detection. Early works [4, 8, 33] mainly focus on monocular 3D detection but a single camera can only provide a 2D view of the scene, and inferring 3D from 2D is intrinsically hard. Recent development of self-driving datasets [5, 34, 6] featured with full sensor suits has enabled the research direction of 3D object detection from multiple cameras. DETR3D [41] proposes a 3D-2D query paradigm for extracting 3D features from 2D multi-view images. Graph-DETR3D [10] further utilizes graph structure learning to enhance the representation at the border regions. Alternatively, another stream of works [17, 29, 44, 23] focuses on aggregating BEV features from multi-view cameras for conducting downstream perception tasks. LSS [29] lifts the 2D features to 3D frustum via latent depth and then splats frustums into a BEV grid. M^{2}BEV [44] further extends LSS with less memory consumption and conducts detection and segmentation. BEVFormer [23] constructs BEV queries and explores spatial cross-attention and temporal self-attention to recurrently refine the BEV features, achieving SOTA performance on both NuScenes [5] and Waymo Open Datasets [34].

Multi-modal fusion. Existing multi-sensor fusion methods can be divided into point-level fusion, proposal-level fusion, and BEV-level fusion. Point-level fusion decorates the input from one modality with attributes from the other modality. PointPainting [38] and PaintAugmenting [39] decorate the LiDAR point clouds with semantic segmentation scores and 2D CNN image features respectively while [11, 1] project LiDAR onto the image plane to augment the RGB values with depth information and conduct detection. On the other hand, proposal-level fusion generates proposals from one modality and then indexes features from the other modality for further refinement. MV3D [9] produces object queries from LiDAR BEV and then extracts the features from camera data and LiDAR front view. [30, 42] lift 2D bounding boxes to frustum and conduct 3D object detection from the frustum of point clouds. Conversely, BEV-level fusion converts features from different modalities to unified BEV representations, preserving both geometric and semantic information. BEVFusions [26, 24] concatenate BEV features from camera and LiDAR and fuse it via a fusion module. Hence, existing single-agent multi-modal fusion methods rely on the co-existence of both camera and LiDAR with fixed geometric relationships, which is unsuitable for our multi-agent hetero-modal cooperative perception problem with a dynamic heterogeneous collaboration graph.

3. Methodology

In this paper, we explore the multi-agent hetero-modal cooperative perception, where each AV is equipped with either a LiDAR or multiple cameras. Our goal is to create a robust and flexible cooperative perception system that allows for efficient collaboration between any number of agents with varying sensor types, ultimately improving the perception capabilities of the vehicle in a unified end-to-end fashion. The pipeline, illustrated in 3, includes modality-specific feature extraction, compression and sharing, HM-ViT for feature fusion, and hetero-modal detection head.

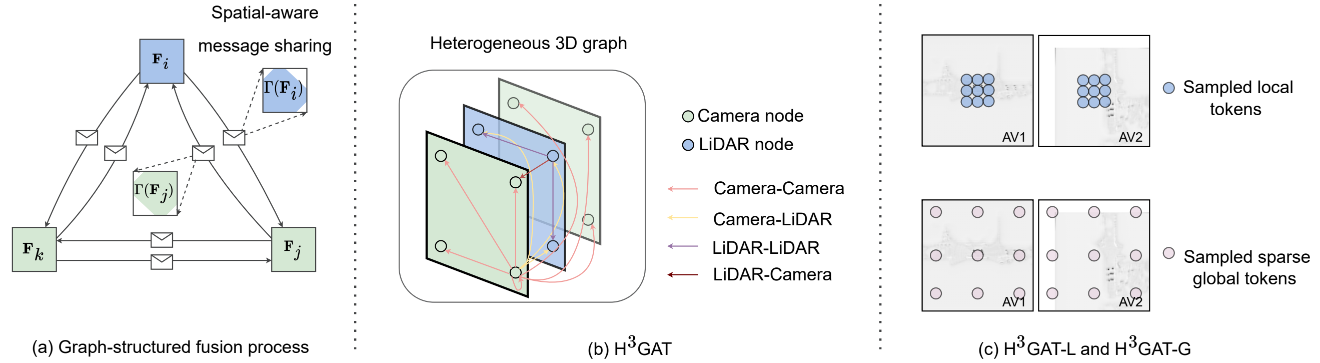


Figure 4: HM-ViT architecture. (a) Graph-structured fusion process. (b) Heterogeneous 3D graph attention (H^{3}GAT). (c) Illustration of sampled tokens for H^{3}GAT-L and H^{3}GAT-G.

3.1. Modality-specific feature extraction

LiDAR stem: We leverage PointPillar [20] to process point clouds for each LiDAR agent. The raw point cloud is converted to a 2D pseudo-image, flattened along the height dimension, and fed into 2D convolutional neural networks to produce the salient feature map $\mathbf{F}_j \in \mathcal{R}^{H\times W\times C}$, which is compressed and shared with all the neighboring agents.

Camera stem: Each camera agent is equipped with m monocular cameras. The sensing observation of i-th agent includes the input images $I_k^i \in \mathcal{R}^{h\times w\times 3}$ and the known projection matrix $P_k^i \in \mathcal{R}^{3\times 4}$ that maps 3D reference points to different image views. Our goal is to generate a BEV feature representation $\mathbf{F}_i \in \mathcal{R}^{H\times W\times C}$ that is amenable for feature fusion with other collaborators. In this work, we adopt similar architecture to BEVFormer [23] with no temporal information for feature extraction. For a faster running time, we adopt ResNet50 to extract 2D image features and then adopt a learnable 2D BEV query to inquire spatial information from the encoded multi-view features via spatial cross attention and projection matrices. The resulting refined BEV feature $\mathbf{F}_i$ is centered around agent i and shared with connected AVs.

3.2. Heterogeneous 3D Graph Attention (H^{3}GAT)

To account for the distinct characteristics of BEV features extracted from different sensor modalities, the learning process of each modality must be distinguished, and the cross-modality interactions between multiple agents should vary. To capture this heterogeneity, we present a novel heterogeneous 3D graph attention (H^{3}GAT), in which nodes and edges are type-dependent to reason spatial interactions and cross-agent relations jointly. We encode both local and global interactions to better capture the 3D ambiguity in BEV feature space. Local attention can help preserve object details, while global attention can provide a better understanding of environmental contexts such as road topology and traffic density.

As shown in Fig. 4b, We build a 3D heterogeneous collaboration graph. Each node $v\left(i,x\right) = \mathbf{F}_x^i \in \mathcal{R}^C$ is a feature vector of agent i's feature map at spatial location $x \in \mathcal{R}^2$. 3D heterogeneous graph attention is performed for spatially connected nodes in the BEV feature space. Depending on the definition of spatial connectivity, we will derive local attention and global attention. Here for notation simplicity, we only derive single-head equations but in real implementations, multi-head variants are used. Formally, we first project feature vectors onto different feature spaces to form query, key, and value vectors:

$$\mathbf{Q}_x^j = \text{Dense}_{\tau_j}\mathbf{F}_x^j \tag{1}$$

$$\mathbf{K}_x^j = \text{Dense}_{\tau_j}\mathbf{F}_x^j \tag{2}$$

$$\mathbf{V}_x^j = \text{Dense}_{e_{ij}}\mathbf{F}_x^j \tag{3}$$

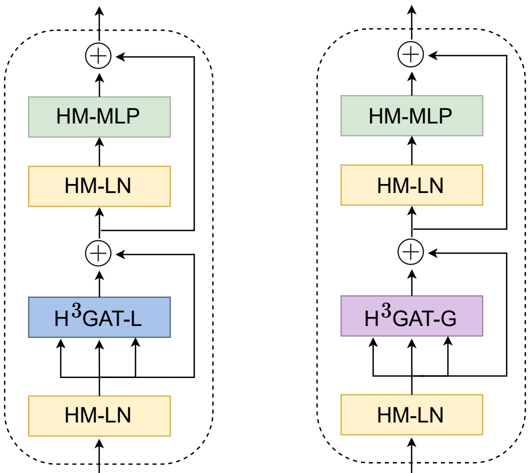
where the $\text{Dense}_{\square}$ is a set of linear layers indexed by subscript $\square$. For the query and key vectors, we use linear projectors Dense_{τ_j} indexed by node type τ_j to extract modality-specific features. For the value vector, we index the projector via edge type $\text{Dense}_{e_{ij}}$ to reflect the heterogeneity of cross-modality multi-agent interactions. The set of connected nodes of $v\left(i,x\right)$ is denoted as $\mathcal{N}\left(i,x\right)$. Then the attention is operated as follows:

$$\mathbf{a}\left(j,y\right) = \underset{(j,y)\in\mathcal{N}(i,x)}{\text{Softmax}}\left(\mathbf{Q}_x^i\mathbf{W}_{e_{ij}}\mathbf{K}_y^j\right) \tag{4}$$

$$\mathbf{F}_x^i = \sum_{(j,y)\in\mathcal{N}(i,x)} \mathbf{a}\left(j,y\right)\mathbf{V}_y^j \tag{5}$$

where $\mathbf{W}_{e_{ij}} \in \mathcal{R}^{C\times C}$ is used to adjust the dot product of $\mathbf{Q}_x^i$ and $\mathbf{K}_y^j$ to further encode the heterogeneity of edges.

Depending on how the nodes are sampled (Fig. 4c), we design two types of attentions: local attention (**H^{3}GAT-L**) which performs local window-based attention and global attention (**H^{3}GAT-G**) which performs sparse global grid-based attention. Fig. 4c visualizes how the local and global



(a) H^{3}GAT-L Block

(b) H^{3}GAT-G Block

Figure 5: Transformer blocks for local and global attentions.

tokens are sampled, where the local tokens are sampled within local windows and the global tokens are sparsely sampled grids scattered across feature maps. The local interactions can help preserve spatial cues and provide reliable estimates while the global reasoning can help understand global semantic context.

Both H^{3}GAT-L and H^{3}GAT-G can be implemented efficiently by decomposing the spatial axes. More specifically, we stack all the agents' features to $\mathbf{F} \in \mathcal{R}^{N \times H \times W \times C}$ where N is the number of agents. For H^{3}GAT-L, we decompose the feature map into 3D non-overlapping windows along the first axis [35, 36], each of size $N \times P \times P$. The partitioned tensor has the shape $(\frac{H}{P} \times \frac{W}{P}, N \times P^2, C)$ where the heterogeneous 3D local graph attention is conducted for NP^2 tokens within the same window. Similarly, for H^{3}GAT-G , we swap the axis and partition the tensor into the shape $(N \times P^2, \frac{H}{P} \times \frac{W}{P}, C)$. Due to the swap operation, the sampled grids will be sparsely scattered and the attention is operated for these sparsely sampled $\frac{H}{P} \times \frac{W}{P}$ grids, which can capture sparse global information.

To integrate this local and global attention into transformer architecture, we further present a heterogeneous normalization layer (**HM-LN**) and heterogeneous MLP (**HM-MLP**) which use type-dependent parameters. As shown in Fig. 5, we first pass all the features into the HM-LN where different statistics are calculated and used as per each agent's modality type. Afterward, we feed the normalized features into a heterogeneous 3D graph attention (H^{3}GAT-L/H^{3}GAT-G) to jointly reason heterogeneous inter-agent and intra-agent interactions. Then, we pass the fused features to another HM-LN followed by a Hetero-modal MLP layer where different sets of parameters are used for camera and LiDAR features. By carefully designing these components, we can maintain modality-specific characteristics throughout the fusion process while benefiting from cross-modality multi-agent interactions.

Algorithm 1 **Multi-agent hetero-modal fusion process**

```
Input: decompressed feature $\mathbf{F}_i$, pose $x_i$ for each agent
$\mathbf{F}_i^{(0)} = \mathbf{F}_i$
for $l = 1 \ldots \ldots L$ do
    for each agent $i$ do                                              ▷ Process in parallel
        $\mathbf{F}_{j \to i}^{(l-1)} = \Gamma_{j \to i}\left(\mathbf{F}_j^{(l-1)}\right)$   ▷ Spatially transform neighboring agents' features
        $\mathbf{F}_i^{(l)} = $ H$^3$GAT-L Block$(\{\mathbf{F}_{j \to i}^{(l-1)}\})$   ▷ Update node via local attention
    end for
    for each agent $i$ do                                              ▷ Process in parallel
        $\mathbf{F}_{j \to i}^{(l)} = \Gamma_{j \to i}\left(\mathbf{F}_j^{(l)}\right)$   ▷ Spatially transform neighboring agents' features
        $\mathbf{F}_i^{(l)} = $ H$^3$GAT-G Block$(\{\mathbf{F}_{j \to i}^{(l)}\})$   ▷ Update node via global attention
    end for
end for
$\mathbf{F}_i$ = HM-MLP$\left(\mathbf{F}_i^{(L)}\right)$                ▷ Output updated features
```

3.3. Hetero-modal Vision Transformer

Compression and sharing: To reduce the transmission bandwidth, a series of 1×1 convolutions is applied to reduce the transmitted feature size along the channel dimension. Together with the intermediate features, each agent's pose x_i is also circulated within the collaboration graph. The ego agent will receive these features and decompress them back to the original size via another convolutional network. For processing intermediate features of the camera agent and LiDAR agent, we leverage distinct parameters in the compression and decompression modules to preserve the modality-specific characteristics.

Graph-structured feature fusion: The received BEV features are centered around different spatial locations as each agent perceives the dynamic environment from different view points. To this end, we present a graph-structured fusion process (Fig. 4a): each node maintains a state representation of an agent in its own coordinate frame, and for a fixed number of iterations, spatially warped messages are shared between nodes and the node states are updated based on the aggregated features via the transformer blocks.

The overall process is summarized in Alg. 1. During each iteration, we have two cascaded node updates which capture the local and global heterogeneous interactions respectively. For each node, we first spatially transform [18] neighboring nodes' features to its center $\mathbf{F}_{j \to i} = \Gamma_{j \to i}(\mathbf{F}_j)$. When the transmitting node is the receiving node itself, the transformation matrix is an identity matrix

	V2V-C		V2V-L		V2V-H	
Models	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Fusion	0.094	0.021	0.524	0.363	0.284	0.157
Late Fusion	0.231	0.070	0.770	0.606	0.502	0.308
V2VNet [40]	0.329	0.125	0.820	0.645	0.650	0.366
DiscoNet [22]	0.287	0.115	0.741	0.590	0.624	0.385
AttFuse [49]	0.261	0.095	0.801	0.644	0.647	0.390
CoBEVT [46]	0.317	0.122	0.828	0.637	0.674	0.416
V2X-ViT [48]	0.332	0.125	0.833	0.679	0.671	0.427
HM-ViT	**0.355**	**0.142**	**0.853**	**0.763**	**0.695**	**0.515**

Table 1: Evaluation of V2V perception methods on OPV2V dataset.

and thus $\mathbf{F}_{i \rightarrow i} = \mathbf{F}_i$. These spatial aligned feature maps $\mathbf{F}_{j \rightarrow i}$ are then shared with agent i to update its state representation via the aggregation module. We adopt H^{3}GAT-L Block as our first aggregation module to capture the local heterogeneous interactions and leverage H^{3}GAT-G Block for the second module to further refine the states with global cues. Within each transformer block, we also adopt a mask to mask out non-overlapping areas between the field of views when computing the attention scores. Note that each agent's state update can be processed in parallel for better efficiency. After L such iterations, we pass the features to a hetero-modal MLP to further refine the feature representation. Throughout the whole fusion process, the modality-specific statistics are maintained.

3.4. Hetero-modal Head

As camera and LiDAR contain distinct characteristics, we design a hetero-modal head where a different set of parameters are applied for camera and LiDAR ego vehicles to generate the final predictions. More specifically, the final fused feature maps are passed to a series of 3 × 3 convolutions with batch normalization and ReLU for feature refinement. Then, we adopt a 1×1 convolution layer to generate the regression and classification predictions. Smooth ℓ_1 loss is utilized for regression and a focal loss [25] is used for classification.

4. Experiments

4.1. Datasets and Evaluation

OPV2V. OP2V [49] is a large-scale multi-modal cooperative V2V perception dataset collected in CARLA [12] and OpenCDA [45]. It contains over 70 driving scenarios of around 25 seconds duration each. Each scenario contains multiple connected AVs (2 to 7) and each AV is equipped with 1 LiDAR and 4 monocular cameras covering 360° horizontal field of view (FoV). In our hetero-modal cooperative perception setting, we only use one type of sensor modality for each AV, leading to two types of agents: vehicles only equipped with multiple cameras (camera agent), and vehicles only equipped with LiDARs (LiDAR agent).

Evaluations. We adopt Average precision (AP) at Intersection-over-Union (IoU) 0.5 and 0.7 to measure the perception performance. As each scenario consists of multiple AVs, a fixed agent is selected as the ego agent and the evaluation is conducted in the range of 100 m × 100 m around it. Following [49, 48], the train/validation/test splits are 6764/1981/2719. We evaluate models mainly under three configurations: 1) V2V Camera-based 3D detection (**V2V-C**) where AVs are only equipped with 4 cameras with 360 horizontal FoV, 2) V2V LiDAR-based detection (**V2V-L**) where all the agents only have LiDAR sensors, and 3) V2V Hetero-modal detection (**V2V-H**) where half of the agents only has cameras while the other half only has LiDARs. To further assess models' capability with dynamic sensor configurations, we also assess models with fixed ego sensor modality and varying collaborator sensor types.

4.2. Experimental Setups

Implementation details. We use the PointPillar [20] as 3D backbones and a modified BEVFormer [23] for our camera stem. For BEVFormer, we adopt its variant with no temporal information and ResNet50 [15] as image backbones for better computation efficiency and use a smaller grid resolution (0.4 m) to preserve fine-trained spatial details. The intermediate BEV feature map has a dimension of 128 × 128 × 256. Following prior works [22, 49, 48, 46], we only change the fusion module for different intermediate fusion methods while keeping the other components such as hetero-modal header, compression, and lidar/camera feature extractors the same. For the heterogeneous methods, we provide the network with additional feature types so that they can distinguish whether the feature is from LiDAR or camera while for the homogeneous methods, the fusion networks would treat the features from LiDAR and camera equally. For HM-ViT, we conduct two iterations of graph-structured feature fusion and employ a window size of 8 for both local and global attentions. We adopt AdamW

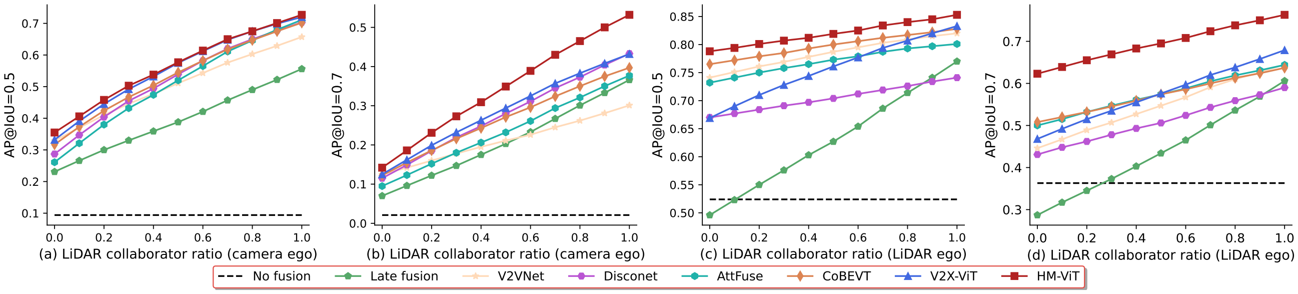


Figure 6: Agent modality ratio experiment. The x-axis is the ratio of LiDAR collaborators among all the collaborators. In (a) and (b), ego vehicles are equipped with cameras. In (c) and (d), ego vehicles are equipped with LiDARs.

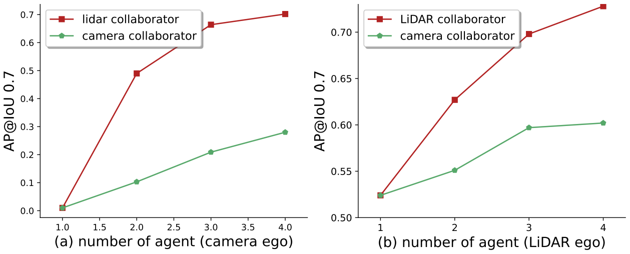


Figure 7: Ablation study on number of agent. The LiDAR/camera collaborator refers to the case where the collaborators are equipped with LiDAR and Camera respectively. (a) ego vehicle is equipped with camera. (b) ego vehicle is equipped with LiDAR.

with a decay rate of 10^{-2} and cosine annealing learning rate scheduler to optimize the models.

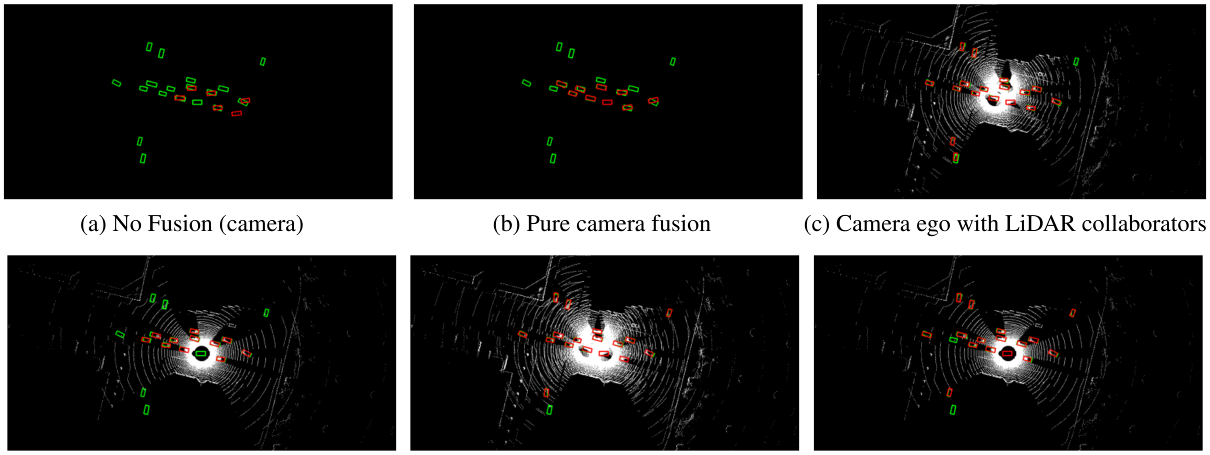
Training strategy. We find that the intermediate fusion models won't converge if directly trained end-to-end under V2V-H and the resulting models usually only exhibit good performance for either camera perception or LiDAR perception but hardly for both. Instead, we first train the model on single modality configurations (*i.e.*, V2V-C and V2V-L) until convergence for 40 epochs and then fine-tune the models under V2V-H for 10 epochs with fixed parameters of modality-specific backbones on 4 RTX3090 GPUs. This training strategy can help models converge well with reliable performance. For a fair comparison, we leverage this training strategy for all the methods.

Compared methods. We regard No Fusion as the baseline method. We also evaluate the Late Fusion, which transmits the detection proposals and leverages Non-maximum suppression to generate the final predictions. For the intermediate cooperation methods, we benchmark five approaches: V2VNet [40], DiscoNet [22], AttFuse [49], CoBEVT [46], and V2X-ViT [48]. For a fair comparison, hetero-modal head is used for all the models.

4.3. Quantitative evaluation

Main performance comparison. Tab. 1 demonstrates the performance comparisons on V2V-C, V2V-L and V2V-H. Under all three settings, all the cooperative methods outperform the No Fusion baseline and the intermediate fusion beats the classical Late Fusion, showing the great benefit of end-to-end V2V hetero-modal cooperative perception. The HM-ViT outperforms all the other SOTA intermediate fusion methods by at least 1.7%, 8.4%, 8.8% in AP@0.7 under V2V-C, V2V-L, V2V-H settings respectively.

Agent modality ratio experiment. As shown in Fig. 6, we fix the ego vehicle sensor modality and vary the ratio of collaborators' sensor modalities to evaluate the models' performance under various heterogeneous traffic scenarios. The larger LiDAR collaborator ratio corresponds to more vehicles only equipped with LiDARs while a smaller ratio means more vehicles only equipped with multiple cameras. The left two figures are the evaluation results for camera ego vehicles while the right two figures are the performance for LiDAR ego vehicles. Under most ratios, late fusion outperforms No Fusion however for LiDAR ego vehicle when most collaborators are camera agents, the Late Fusion performs poorer than No Fusion. We argue this is due to the fact that the camera predictions are usually noisy and merging proposals from different modalities equally could lead to ambiguity and thus deteriorate the performance. On the other hand, all the intermediate fusion methods outperform No Fusion by a large margin especially for the camera ego vehicles, demonstrating the great value of V2V cooperation between agents with different modalities. Among all the compared methods, HM-ViT ranks first for both camera ego vehicles and LiDAR ego vehicles under all the ratios, illustrating the great capability of HM-ViT for capturing modality-specific characteristics and cross-modality multi-agent interactions. In contrast, other intermediate fusion methods only show good performance for a certain ratio range, which demonstrates the importance of heterogeneity for hetero-modal cooperative perception.



(a) No Fusion (camera)
(b) Pure camera fusion
(c) Camera ego with LiDAR collaborators
(d) No Fusion (LiDAR)
(e) Pure LiDAR fusion
(f) LiDAR ego with camera collaborators

Figure 8: Qualitative visualizations for (a) No Fusion with camera ego vehicle, (b) pure camera-based V2V perception, (c) hetero-modal V2V perception with camera ego vehicle and LiDAR collaborators, (d) No Fusion with LiDAR ego vehicle, (e) pure LiDAR-based V2V perception, and (f) hetero-modal V2V perception with LiDAR ego vehicle and camera collaborators. The red and green boxes represent the detection outputs and ground truth respectively. More visualizations can be found in the supplementary materials.

HM-MLP&LN	H^3GAT-L	H^3GAT-G	AP@0.7
			0.404
✓			0.420
✓	✓		0.460
✓	✓	✓	**0.515**

Table 2: Component Ablation study on the V2V-H setting

Compression Rate	AP@0.7
0x	0.515
8x	0.513
16x	0.470
32x	0.455

Table 3: Compression rate effects for HM-ViT on V2V-H.

Number of agent. In this experiment, we investigate the effect of the number of agents on the perception performance of HM-ViT. As Fig. 7 depicts, for both camera and LiDAR ego vehicles, the perception performance increases as more agents are involved in the cooperative perception and both LiDAR and camera collaborators can contribute to the performance gain for ego vehicles with different modalities, which again shows the benefit of hetero-modal V2V cooperation. Additionally, the increase rate generally decreases when increasing the number of agents and the LiDAR collaborators can bring more AP gains over the camera collaborators for both camera and LiDAR ego vehicles. Notably, similar as Fig. 6a-b display, the camera ego vehicles' performance can be greatly improved with only a small number of LiDAR collaborators, demonstrating the great potential of reducing the cost for each vehicle when the V2V system is deployed at scale as we may only need to install expansive LiDARs for a small number of agents (*e.g.*, infrastructure) while all the other agents only require relatively cheap camera sensors.

Component ablation study. Here we investigate the key components of the proposed HM-ViT. As the layer normalization and MLP are usually combined together in typical transformer designs, thus we jointly evaluate the combined effect of HM-MLP and HM-LN (HM-MLP&LN). As Tab. 2 shows, all the proposed components improve the performance and local and global attentions can largely improve the AP@0.7 by 4% and 5.5%, proving the great benefit brought by jointly reasoning inter-agent and intra-agent heterogeneous interactions both locally and globally.

Compression rate. Tab. 3 describes the influence of compression rate. It demonstrates that HM-ViT can still outperform other methods even under large compression rates.

4.4. Qualitative results

Fig. 8 depicts the qualitative visualizations for HM-ViT and No Fusion baselines. In fig. 8a-c, we plot the detection results for camera ego vehicles with no collaborator, camera collaborators, and LiDAR collaborators respectively while for Fig. 8d-f, we plot the results for LiDAR ego vehicles. As shown in these figures, the collaborations with both homogeneous and heterogeneous agents are beneficial for camera ego vehicles and LiDAR ego vehicles with enhanced detection results. In particular, the collaboration between camera ego vehicles and LiDAR collaborators can dramatically enhance the perception performance.

5. Conclusion

In this paper, we present HM-ViT, a hetero-modal vision transformer, for the hetero-modal multi-agent cooperative perception problem which is an important but underexplored research direction. We propose a generic heterogeneous 3D graph attention to jointly reason heterogeneous inter-agent and cross-agent interactions. Our extensive experiments demonstrate the outstanding performance of the proposed method and the great potential of hetero-modal multi-agent collaborations for increasing the scalability and robustness of V2V systems. We hope our findings and open-source efforts will inspire more research on this new problem.

References

[1] Luís A Alexandre. 3d object recognition using convolutional neural networks with transfer learning between input channels. In *Intelligent Autonomous Systems 13*, pages 889–898. Springer, 2016. 3

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan 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*, pages 1090–1099, 2022. 2

[3] Johannes Betz, Hongrui Zheng, Alexander Liniger, Ugo Rosolia, Phillip Karle, Madhur Behl, Venkat Krovi, and Rahul Mangharam. Autonomous vehicles on the edge: A survey on autonomous vehicle racing. *IEEE Open Journal of Intelligent Transportation Systems*, 3:458–488, 2022. 1

[4] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 3

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[6] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 3

[7] Hung-Hsun Chen, Yi-Bing Lin, I-Hau Yeh, Hsun-Jung Cho, and Yi-Jung Wu. Prediction of queue dissipation time for mixed traffic flows with deep learning. *IEEE Open Journal of Intelligent Transportation Systems*, 3:267–277, 2022. 1

[8] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2156, 2016. 3

[9] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 3

[10] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Graph-detr3d: Rethinking overlapping regions for multi-view 3d object detection. *arXiv preprint arXiv:2204.11582*, 2022. 3

[11] Zhuo Deng and Longin Jan Latecki. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5762–5770, 2017. 3

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 6

[13] Yi Guo and Jiaqi Ma. Leveraging existing high-occupancy vehicle lanes for mixed-autonomy traffic management with emerging connected automated vehicle applications. *Transportmetrica A: Transport Science*, 16(3):1375–1399, 2020. 1

[14] Yi Guo, Jiaqi Ma, Edward Leslie, and Zhitong Huang. Evaluating the effectiveness of integrated connected automated vehicle applications applied to freeway managed lanes. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):522–536, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *arXiv preprint arXiv:2209.12836*, 2022. 2

[17] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

[19] Sou Kitajima, Hanna Chouchane, Jacobo Antona-Makoshi, Nobuyuki Uchida, and Jun Tajima. A nationwide impact assessment of automated driving systems on traffic safety using multiagent traffic simulations. *IEEE Open Journal of Intelligent Transportation Systems*, 3:302–312, 2022. 1

[20] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 4, 6

[21] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *European Conference on Computer Vision*, pages 316–332. Springer, 2022. 2

[22] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 1, 2, 6, 7

[23] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 3, 4, 6

[24] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022. 3

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[26] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 3

[27] Jiaqi Ma, Edward Leslie, Amir Ghiasi, Zhitong Huang, and Yi Guo. Empirical analysis of a freeway bundled connected-and-automated vehicle application using experimental data. *Journal of Transportation Engineering, Part A: Systems*, 146(6):04020034, 2020. 1

[28] Vasileia Papathanasopoulou, Ioanna Spyropoulou, Harris Perakis, Vassilis Gikas, and Eleni Andrikopoulou. A data-driven model for pedestrian behavior classification and trajectory prediction. *IEEE Open Journal of Intelligent Transportation Systems*, 3:328–339, 2022. 1

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

[30] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 3

[31] Kelli Raboy, Jiaqi Ma, Edward Leslie, and Fang Zhou. A proof-of-concept field experiment on cooperative lane change maneuvers using a prototype connected automated vehicle testing platform. *Journal of Intelligent Transportation Systems*, 25(1):77–92, 2021. 1

[32] Steven E Shladover. Opportunities and challenges in cooperative road vehicle automation. *IEEE Open Journal of Intelligent Transportation Systems*, 2:216–224, 2021. 1

[33] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 3

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[35] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 5

[36] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 5

[37] Rodolfo Valiente, Behrad Toghi, Ramtin Pedarsani, and Yaser P Fallah. Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic. *IEEE Open Journal of Intelligent Transportation Systems*, 3:397–410, 2022. 1

[38] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2, 3

[39] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021. 2, 3

[40] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*, pages 605–621. Springer, 2020. 1, 2, 6, 7

[41] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 3

[42] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1742–1749. IEEE, 2019. 3

[43] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception. *arXiv preprint arXiv:2209.13679*, 2022. 2

[44] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. Mˆ 2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 3

[45] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162. IEEE, 2021. 6

[46] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022. 1, 2, 6, 7

[47] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. *arXiv preprint arXiv:2303.07601*, 2023. 1

[48] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *arXiv preprint arXiv:2203.10638*, 2022. 1, 2, 6, 7

[49] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi 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)*, pages 2583–2589. IEEE, 2022. 1, 2, 6, 7

[50] Xinli Xu, Shaocong Dong, Lihe Ding, Jie Wang, Tingfa Xu, and Jianan Li. Fusionrcnn: Lidar-camera fusion for two-stage 3d object detection. *arXiv preprint arXiv:2209.10733*, 2022. 2

[51] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 1

[52] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22552–22562, 2023. 1

[53] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. *Advances in neural information processing systems*, 35:25739–25753, 2022. 1

[54] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 1

[55] Lin Zhao, Hui Zhou, Xinge Zhu, Xiao Song, Hongsheng Li, and Wenbing Tao. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *arXiv preprint arXiv:2108.07511*, 2021. 2