HM-ViT: Hetero-modal Vehicle-to-Vehicle Cooperative Perception with Vision Transformer (supplementary materials)

1. Model details

In this section, we detail the model configurations. The model specifications are depicted in Fig. S1.

BEVFormer: We adopt a variant of BEVFormer with no temporal information for our camera stem. As we need to extract features for multiple agents at once, using a large network can easily consume all the computation resources for our experimented RTX3090 GPUs. Thus, instead of using ResNet101-DCN [3, 2], we leverage ResNet50 as our backbone. Since the OPV2V [10] dataset is smaller than nuScenes [1] dataset and the used image resolution is 512×512 compared with 900×1600 in the original paper, we find that this smaller backbone exhibits faster convergence behavior in our configurations. To further save the computation, we also set the grid size to be 128×128.

PointPillar: In our experiment, the voxel resolution is $0.4, 0.4, 4$ meters for x, y, and z directions, and the final generated feature map has the dimension of $128 \times 128 \times 256$.

Compression and decompression: We adopt a 1×1 convolution followed by a batch normalization layer and ReLU to compress the features along the channel dimension. Similarly, for the decompression, we leverage convolutions to decompress the features to the original size.

HM-ViT: The received features are first concatenated to form a stacked tensor $\mathbf{F}$ of shape $N \times 128 \times 256$. Then, we conduct two iterations of graph feature updates. In each iteration, we spatially warp the features and then pass them to the H^3GAT-L block and H^3GAT-G block. The window sizes for both blocks are 8 and the number of heads is also 8. After the iterative feature fusion learning, we pass the features to the HM-MLP to further refine the features.

Hetero-modal head: The final fused ego vehicle's feature is fed into the hetero-modal decoder and then passed to the classification and regression head.

Hetero-modal MLP: The hetero-modal MLP sequentially feeds features into the linear layer, GELU activation function [4], dropout layer, linear layer, and dropout layer. Different sets of parameters are used as per each agent's modality.

Hetero-modal LN: We adopt classical layer normalization to calculate the statistics for the input data while the learnable affine parameters are learned separately for features of camera agents and LiDAR agents.

2. Design choice of communication

Previous LiDAR-based cooperative perception communication choices [6, 7, 9, 10, 5] can be broadly classified into two categories: 1) ego-centric approach [6, 5, 7] where the LiDAR point clouds of neighboring AVs are first projected to the ego vehicle's coordinate frame and then features are extracted based on the projected point clouds, and 2) agent-centric approach [9, 10] where each agent extracts features based on its own sensing observation in its own coordinate system and then broadcasts the features to all the neighboring agents. The ego-centric approach can preserve more feature points as the points are projected to the ego frame and all the relevant points within the evaluation range are kept but this approach requires more computation and the computation scales linearly as the number of ego agents increases. Moreover, projecting RGB images from the other agents directly to the ego frame is hardly feasible and is ill-posed due to the occlusion and noisy 3D-2D correspondence. To this end, we adopt the agent-centric approach for broadcasting features for its computation efficiency and hospitality for camera feature extraction.

3. Experiment

3.1. Comparison with other dataset

V2V4Real [8] only released the LiDAR data, and the camera data has not yet been released. This makes it unsuitable for conducting hetero-modal experiments, which require data from multiple sensor modalities. DAIR-V2X [11] is an open dataset for Vehicle-to-Infrastructure (V2I) cooperation. The infrastructure-side camera and vehicle-side camera are mounted at different heights with distinct pitch angles, leading to divergent data distribution. Thus hetero-modal V2I cooperation involves another agent type heterogeneity, and it is beyond the scope of this work where we focus on hetero-modal V2V cooperation. OPV2V [10] is a large-scale simulation dataset for V2V cooperative perception. It provides multi-view images and LiDAR data for each agent, which is suitable for investigating

	Output size	HM-ViT framework
PointPillar Encoder	$N_1 \times 128 \times 128 \times 256$	
BevFormer Encoder	$N_2 \times 128 \times 128 \times 256$	
Compression	$128 \times 128 \times 256/k$	Hetero-modal compressor: [Conv1x1, BN, ReLU]
Decompression	$128 \times 128 \times 256$	Hetero-modal decompressor: [Conv1x1, BN, ReLU Conv1x1, BN, ReLU]
HM-ViT Backbone	$N \times 128 \times 128 \times 256$	[ConCat0, 256] [H^3GAT-L, dim 256, head 8 window size, {8, 8} HM-MLP, dim 256 HM-LN, dim 256 H^3GAT-G, dim 256, head 8 window size, {8, 8} HM-MLP, dim 256 HM-LN, dim 256] $\times 2$ [HM-MLP, dim 256]
Hetero-modal Head	$128 \times 128 \times 16$	Hetero-modal decoder: [Conv3x3, BN, ReLU Conv3x3, BN, ReLU] $\times 2$ Hetero-modal class. head: [Conv1x1, 2, stride 1] Hetero-modal reg. head: [Conv1x1, 14, stride 1]

Table S1: Architecture details of HM-ViT. N_1 represents the number of camera agents, N_2 represents the number of LiDAR agents and $N = N_1 + N_2$ represents the total number of agents. k is the compression rate.

hetero-modal V2V cooperative perception. In the future, we plan to further investigate hetero-modal V2X cooperative perception in real-world collected dataset.

3.2. Sensor configuration

Each vehicle is equipped with 1 LiDAR and 4 cameras and the cameras are installed on four sides of the vehicles (left/right/front/rear, covering 360 horizontal field-of-the-view as Fig. 1 shows.

3.3. Additional qualitative results

In Fig. 2 (camera ego vehicle) and Fig. 3 (LiDAR ego vehicle), we provide more qualitative comparisons between HM-ViT and other intermediate fusion methods including V2VNet [6], Disconet [5], AttFuse [10], CoBEVT [7], and V2X-ViT [9]. In all the figures, we only plot the LiDAR point clouds of the agent if agent's LiDAR is involved in the collaboration. Our method produces more robust predictions compared with other methods. In particular, as shown in Fig. 2, after collaborating with the LiDAR agents, our HM-ViT can accurately predict all the vehicles, which again demonstrates the superiority of our method and the great potential of multi-agent hetero-modal cooperation.

(a) front (b) right (c) rear (d) left

Figure 1: Visualization of OPV2V camera data where each row is 4 camera data of one vehicle.

References

[1] 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. 1

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1

[3] 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. 1

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[5] 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] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

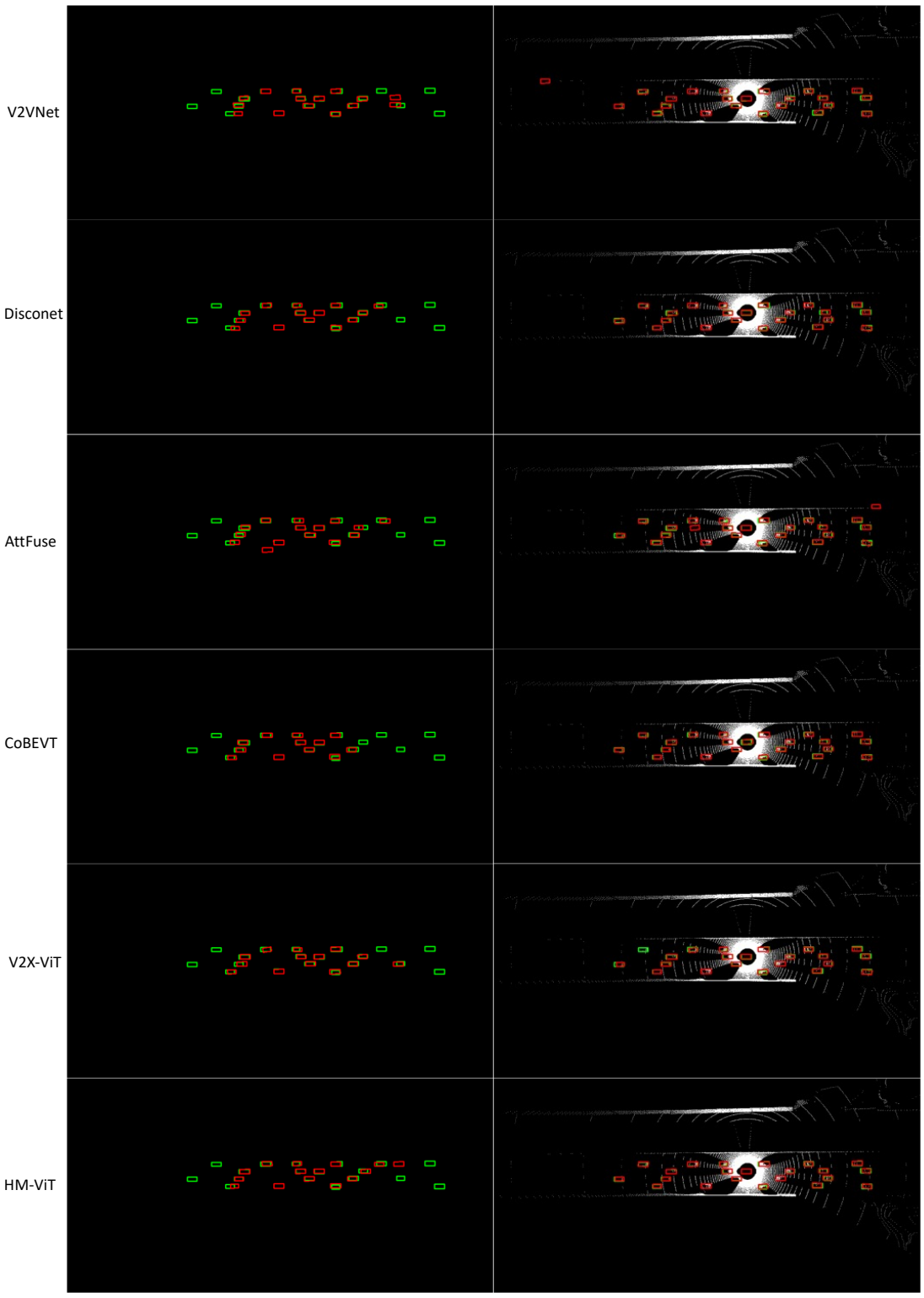

Figure 2: Additional qualitative results for (a) pure camera-based V2V perception and (b) hetero-modal V2V perception with camera ego vehicle and LiDAR collaborators.

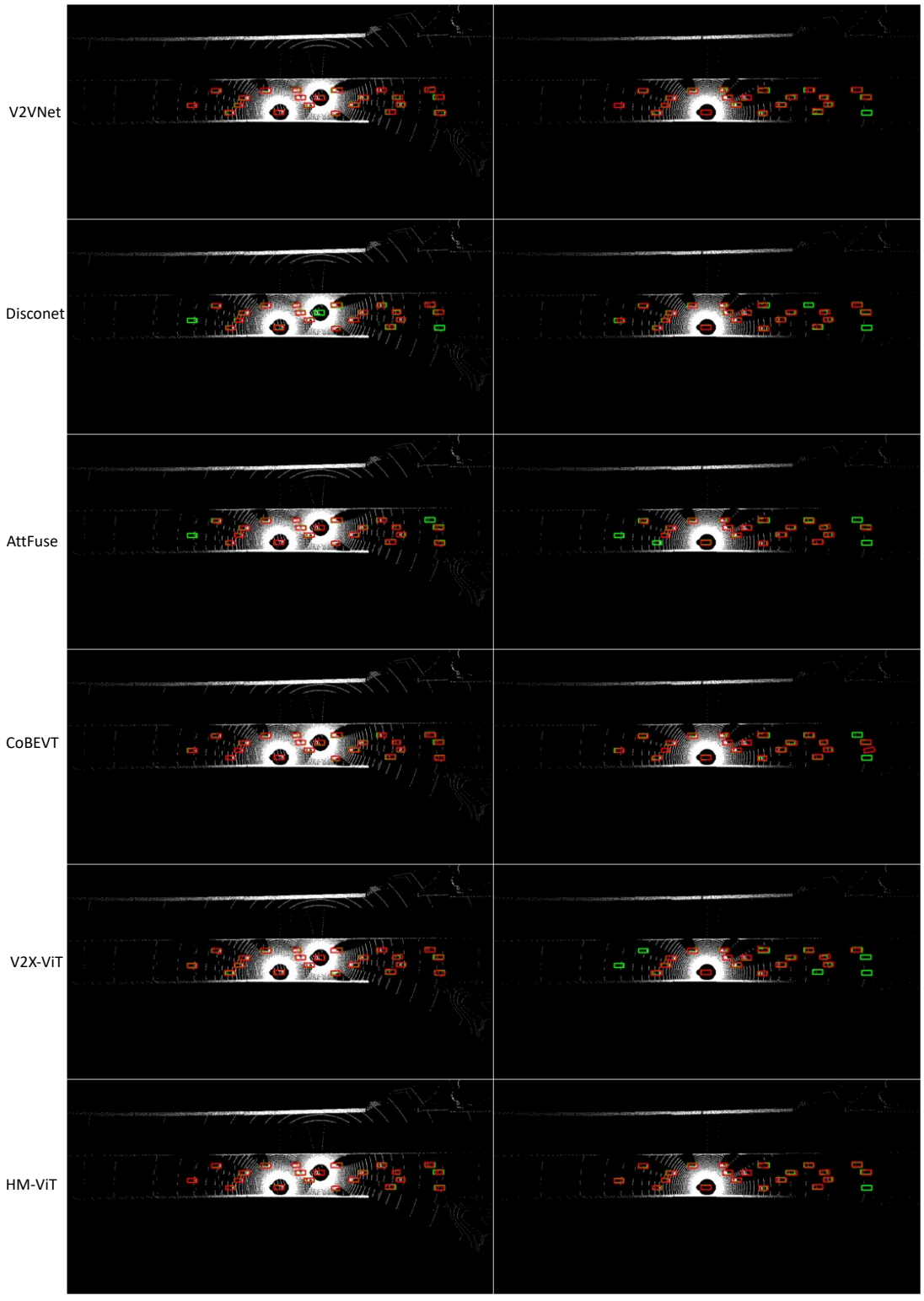

Figure 3: Additional qualitative results for (a) pure LiDAR-based V2V perception and (b) hetero-modal V2V perception with LiDAR ego vehicle and camera collaborators.