A. Implementation Details

A.1. Model

In our implementation, we use a weighted sum strategy to combine the feature maps of the ego vehicle and the surrounding vehicles. The weight of the ego vehicle feature map is set to 1, while the weights of the surrounding vehicles are set to $1/N$, where N is the number of surrounding vehicles.

After combining all feature maps from the ego and surrounding vehicles, we add another 3×3 convolution layer (with normalization and activation layers) to further adjust for any spatial misalignment, following the approach of BEVFusion [59].

We adopt the center-based strategy to predict the locations of objects and employ several regression heads to estimate object size and heading. For further details, please refer to previous studies on 3D object detection [57, 61].

A.2. Training

The voxel size is set to $(0.200, 0.075, 0.200)$ for the x, y, and z directions, respectively. We apply common point cloud data augmentation techniques to prevent overfitting, such as scaling and rotation [58]. The learning rate is initially set to 2×10^{-5}, and we schedule the learning rate with a cosine annealing. For the main results, we trained our models for 20 epochs with a batch size of 4 per GPU (NVIDIA A100).

B. Discussion

Different Fusion Methods We conduct an ablation study to compare the weighted sum method with other feature fusion strategies. These strategies include sum (directly adding up the feature maps), mean (calculating the average of all feature maps), and concat (computing the mean of the feature maps of the surrounding vehicles and then concatenating it with the feature map of the ego vehicle, doubling the number of channels). All the methods are compared with a compression factor of 4, and the models are trained for 5 epochs on the OPV2V dataset [60]. As seen in Tab. 4, the weighted sum approach achieves the top overall performance.

Method	AP@IoU=50/70 (↑)			
	Overall	0-30m	30-50m	50-100m
Mean	89.6/86.2	97.7/96.3	92.2/88.0	77.7/72.5
Sum	74.8/68.9	91.1/87.1	72.0/64.6	54.4/47.7
Concat	91.9/86.2	**98.0**/96.3	**92.2/86.7**	84.5/74.1
Weighted Sum (**MACP**)	**92.4/87.1**	**98.0/96.6**	92.0/86.3	**85.9/76.5**

Table 4. **Ablation Results on Fusion Methods.**

Robustness Taking advantage of an extended field of view from V2V cooperative perception, the proposed MACP model should be able to tackle the occlusion or sudden decreases in visibility, that is, to have higher robustness. We design an experiment to assess the robustness by traversing and masking out part of the ego vehicle's field of view and looking into how it affects performance. Fig. 8 visualizes the results, where with an occlusion range of 40 by 40 meters, the single-agent perception model struggles to maintain a stable performance, revealed by an increasing variance in average precision. On the contrary, the cooperative perception model consistently retains its prediction accuracy, exhibiting an AP standard deviation of 1.08%, demonstrating remarkable robustness.

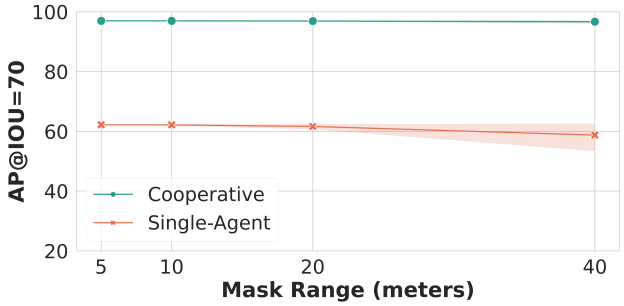

Figure 8. **Results of the robustness analysis.** Visualization reflects that the extra information from cooperative perception significantly helps the model tackle the drop in performance due to occlusions or other factors causing missing observations.

References

[57] 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 *CVPR*, pages 1090–1099, 2022. 1

[58] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *CVPR*, pages 12697–12705, 2019. 1

[59] 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. In *ICRA*, 2023. 1

[60] 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 *ICRA*, pages 2583–2589, 2022. 1

[61] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-Based 3D Object Detection and Tracking. In *CVPR*, pages 11784–11793, 2021. 1