CoSDH: Communication-Efficient Collaborative Perception via Supply-Demand Awareness and Intermediate-Late Hybridization

Supplementary Material

A. Dataset Details

We evaluate the proposed CoSDH against other methods on three different collaborative perception datasets (**OPV2V** [36], **V2XSim** [20], and **DAIR-V2X** [42]) for LiDAR-based 3D object detection. There are more details about these datasets.

OPV2V [36] is a large-scale V2V collaborative perception simulation dataset, obtained through the OpenCDA [34] and CARLA [8] simulators. It contains 11,464 frames of 3D radar point clouds and RGB images, along with 232,913 3D annotated bounding boxes. The training, validation, and test sets consist of 6,374, 2,170, and 1,980 frames, respectively. We set the perception range to $x \in [-140.8m, 140.8m], y \in [-38.4m, 38.4m]$.

V2XSim [20] is a large-scale V2X collaborative perception dataset, obtained using the SUMO [15] and CARLA [8] simulators. It includes 10,000 frames of 3D radar point clouds and RGB images, with annotations for object detection, tracking, and semantic segmentation tasks. The training, validation, and test sets consist of 8,000, 1,000, and 1,000 frames, respectively. We set the perception range to $x \in [-32m, 32m], y \in [-32m, 32m]$.

DAIR-V2X [42] is the first dataset of collaborative perception of V2I in the real world, collected at an intersection in Beijing. It contains 38,845 frames of 3D radar point clouds and RGB images, with each frame containing data from a vehicle and a roadside infrastructure. This dataset contains 4,811 frames for the training set and 1,789 frames for validation set. We set the perception range to $x \in [-140.8m, 140.8m], y \in [-40m, 40m]$. We use the complete 360-degree annotations provided in [25].

B. More Experiments

V2V4Real [38] is the first large-scale real-world V2V collaborative perception dataset, containing ∼10,000 frames of collaborative LiDAR data from two vehicles and providing ∼240,000 annotated 3D bounding boxes across 5 categories. We also conducted additional experiments on the V2V4Real dataset, with the results shown in Table A. The experiments demonstrate that CoSDH achieved the highest accuracy on V2V4Real compared to other methods while using low bandwidth.

C. More Visualization

Fig. A shows the regions selected by the supply-demand masks. From the first row, we can see that the demand mask covers the occluded areas near the Ego agent and regions with sparse point clouds at a distance, which are areas where Ego's perception is poor and require collaboration. The left side of the second and third rows shows the potential foreground regions selected by the supply masks of the two collaborating agents, which correspond closely to the ground truth in the detection results. The right side of the second and third rows shows the supply-demand masks of the two collaborating agents combined with the Ego agent's demand mask. The regions circled in the figures show that the supply-demand mask selects fewer areas around the Ego agent compared to the left side, avoiding the selection of areas where the Ego agent has a good observation, thereby further reducing bandwidth. Due to the large perception range, the well-observed regions of the Ego agent account for only about ∼4%, and the demand mask covers the majority of the area. However, because of occlusions, the demand mask contains fewer foreground regions. By combining supply masks, the bandwidth can be reduced by ∼10% while maintaining detection accuracy.

Fig. B shows the PR (Precision-Recall) curves on the DAIR-V2X [42] dataset for the scenarios without late fusion (intermediate fusion results), with naive late fusion, and with our confidence-aware late fusion. As can be seen, using naive late fusion introduces more suboptimal results, significantly lowering precision under low recall, and these results may overwrite the better detection results in the NMS (Non-Maximum Suppression) stage, causing a drop in overall recall. After using our confidence-aware late fusion, these suboptimal results are avoided, preventing a decrease in precision and improving recall, and it leads to an overall increase in AP (Average Precision).

D. More Ablation

Table B shows the differences in accuracy and bandwidth for various compression and fusion methods discussed in "3.4 Message Compression and Fusion". Since different compression and fusion methods affect the selection ratio of foreground regions, we removed the supply-demand selection module and transmit the complete BEV feature map, in order to more directly demonstrate the advantages of our method in terms of accuracy and bandwidth. From the table, it can be seen that the multi-scale fusion scheme has a significant advantage in accuracy compared to the traditional single-scale fusion scheme. Our multi-scale compression scheme significantly reduces bandwidth by compressing smaller intermediate features, and the fusion performed immediately after compression further improves accuracy.

Method	AP@0.5	AP@0.7	BD
FCooper [3]	71.03%	39.02%	2,640.0 Mbps
AttFuse [36]	68.57%	39.27%	2,640.0 Mbps
V2XViT [35]	72.83%	45.82%	2,640.0 Mbps
Where2comm [11]	71.51%	46.60%	6.4 Mbps
CoAlign [25]	72.08%	45.80%	2,640.0 Mbps
CoSDH	**73.36%**	**47.64%**	6.5 Mbps

Table A. Comparison on V2V4Real [38]. For Where2comm and CoSDH, the bandwidth is constrained within real-world limitation by adjusting the collaboration area.

E. Discussion About the Latency

Our CoSDH requires the transmission of three messages between collaborative agents: demand masks, intermediate features, and detection results. For late fusion, we transmit the detection results of single cars rather than the detection results from intermediate fusion, as this can significantly reduce latency, allowing the detection results to be transmitted together with the intermediate features. In this way, CoSDH only requires two rounds of inter-agent communication, which is still more than the common single-communication methods. In the following, we discuss the latency issues of CoSDH.

1. **The total latency of CoSDH is not necessarily higher.** (1) All collaborative methods require communication to transmit features. Since the communication volume is large, the transmission latency primarily depends on it (the larger the volume, the higher the latency). CoSDH selects key regions according to the supply-demand relationship, which reduces the communication volume by over 60% while maintaining the highest accuracy, thereby significantly lowering the latency. (2) Another two communications with smaller volumes marginally contribute to the total latency as the system supports parallel processing. We record the latency of main components in Fig 2 on OPV2V dataset: $t_B = 15$ ms (backbone), $t_{DG} = 1$ ms (demand generator), $t_{SG} = 3$ ms (supply generator). We roughly use $t_C = 20$ ms as the latency for the two communications [9]. Since $t_B + t_{SG} = 18$ ms $< t_{DG} + t_C = 21$ ms, the transmission of demand masks only adds 3 ms to the total latency. Similarly, transmitting detection results does not greatly increase the total latency.

2. **CoSDH proves robust to the latency because of multi-scale fusion and collaborative region selection.** As shown in Fig. C, CoSDH maintains the highest accuracy under a 100 ms latency (the V2X communication based on IEEE 802.11p features a low latency, capable of keeping the latency within 100 ms). With 200 ms, almost all collaborative methods perform worse than "No Fusion".

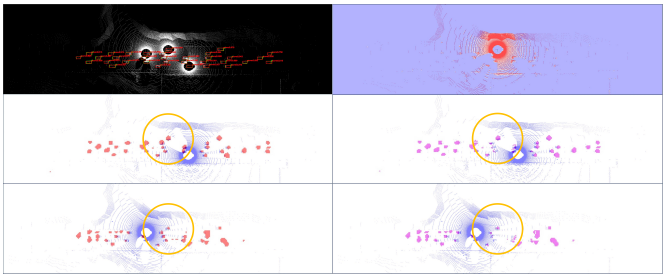

Figure A. Visualization of supply and demand mask on the OPV2V [36] dataset. The left side of the first row shows the collaborative perception detection results, while the right side shows the Ego agent's point cloud and its own demand mask (blue). The left side of the second and third rows shows the point clouds of the two collaborating agents and their corresponding supply masks (red), and the right side showing the supply-demand masks (pink) combined with the Ego's demand mask.

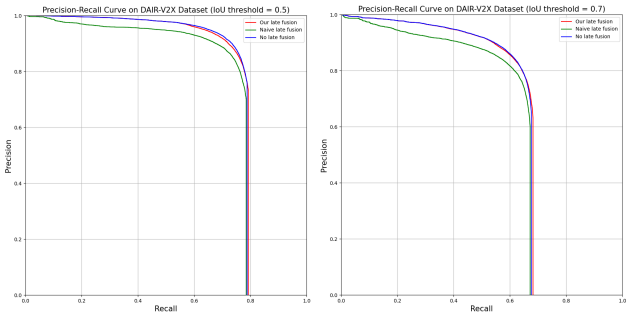

Figure B. PR curves for different fusion schemes on DAIR-V2X [42] dataset.

	Autoencoder			Codebook [12]		
	SF	MF/SC	MF/MC	SF	MF/SC	MF/MC
AP@0.5↑	95.72%	96.39%	96.81%	90.48%	92.52%	92.96%
AP@0.7↑	91.18%	92.31%	93.00%	84.32%	87.40%	88.17%
BD↓	76.2 Mbps	76.2 Mbps	33.4 Mbps	0.4 Mbps	1.0 Mbps	0.4 Mbps

Table B. Accuracy and bandwidth for different fusion and compression methods on OPV2V [36] dataset. "SF" represents single-scale fusion, "MF/SC" represents multi-scale fusion and single-scale compression, "MF/MC" represents multi-scale fusion and multi-scale compression. "BD" represents the bandwidth required for each collaborative agent with the detection frequency of 10Hz.

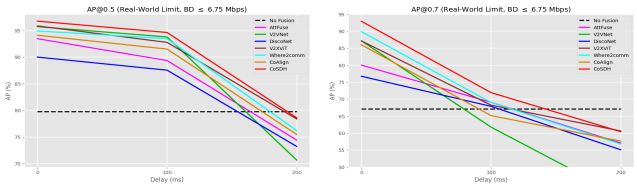

Figure C. Robustness to latency on the OPV2V [36] dataset.