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

Junhao Xu[1]*, Yanan Zhang[2]*, Zhi Cai[1], Di Huang[1]†
[1]State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing, China
[2]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
junhaoxu@buaa.edu.cn, yananzhang@hfut.edu.cn, {caizhi97, dhuang}@buaa.edu.cn

Abstract

Multi-agent collaborative perception enhances perceptual capabilities by utilizing information from multiple agents and is considered a fundamental solution to the problem of weak single-vehicle perception in autonomous driving. However, existing collaborative perception methods face a dilemma between communication efficiency and perception accuracy. To address this issue, we propose a novel communication-efficient collaborative perception framework based on supply-demand awareness and intermediate-late hybridization, dubbed as CoSDH. By modeling the supply-demand relationship between agents, the framework refines the selection of collaboration regions, reducing unnecessary communication cost while maintaining accuracy. In addition, we innovatively introduce the intermediate-late hybrid collaboration mode, where late-stage collaboration compensates for the performance degradation in collaborative perception under low communication bandwidth. Extensive experiments on multiple datasets, including both simulated and real-world scenarios, demonstrate that CoSDH achieves state-of-the-art detection accuracy and optimal bandwidth trade-offs, delivering superior detection precision under real communication bandwidths, thus proving its effectiveness and practical applicability. The code will be released at https://github.com/Xu2729/CoSDH.

1. Introduction

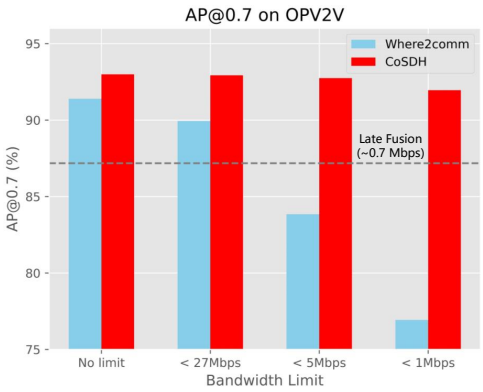
Collaborative perception allows multiple agents to exchange complementary perception information. It fundamentally addresses the issues of limited perception range, sensor blind spots, and occlusion from obstacles inherent in single-agent perception, thereby improving both the range and accuracy of perception. Recent studies have demonstrated that collaborative perception can be applied to various autonomous driving tasks, including 3D object detection [3, 35, 36], semantic segmentation [20, 37], 3D occupancy prediction [31], and trajectory prediction [43], exhibiting enhanced performance and improved robustness to occlusions, thus making it a crucial approach for advancing autonomous driving systems.

Communication efficiency is a key issue in collaborative perception. An early approach to improving communication efficiency was to use autoencoders to compress the intermediate features that need to be transmitted [36]. Later, communication-efficient methods such as Where2comm [11] were proposed, which select important and sparse foreground regions for collaboration. However, these methods still lack sufficient precision in area selection and require more bandwidth than the real-world constraints[1]. As shown in Fig. 1a, when the bandwidth is limited to 27 Mbps, Where2comm [11] experiences a decrease of about 1.5% in average precision (AP@0.7) on the OPV2V [36] dataset with an intersection-over-union (IoU) threshold of 0.7. When the bandwidth is further limited to 5 Mbps and 1 Mbps, AP@0.7 drops by approximately 7.5% and 14.5%, respectively, even falling below the accuracy of late fusion methods at the result level. This makes it difficult for existing collaborative perception methods to meet the demands of practical applications.

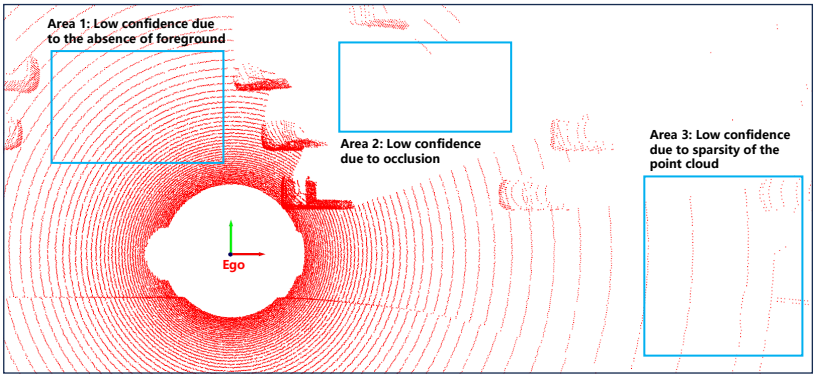
In this paper, we first analyze the problem of selecting collaboration areas and refine this selection based on the supply-demand relationship between agents. As shown in Fig. 1b, for Ego, low-foreground confidence areas can be divided into three types: Area 1 (low conficence due to absence of foreground), Area 2 (low conficence due to occlusion), and Area 3 (low conficence due to sparsity of the point cloud). Although all three areas appear as background from the Ego's perspective, Area 1 is well-observed and does not require collaboration, while Areas 2 and 3,

*Equal Contribution. †Corresponding Author.

[1]To the best of our knowledge, the common V2X communication approach uses IEEE 802.11p-based DSRC (Dedicated Short-Range Communications) technology [1, 24], which provides an information transmission rate of approximately 27 Mbps.



(a) The 3D detection accuracy of Where2comm [11] and CoSDH on OPV2V dataset [36] with different bandwidth limit. Assume the number of collaborative agents is 4 and detection frequency is 10Hz.



(b) Point cloud map and classification of areas with low foreground confidence. Although Area 1 has low confidence due to the absence of foreground, it allows for good observation without collaboration. In contrast, Areas 2 and 3 exhibit poor observation and require collaboration, with Area 2 having low confidence due to occlusion and Area 3 due to sparsity of the point cloud.

Figure 1. Issues of existing communication-efficient collaborative perception methods.

with poorer observation, need collaboration. Additionally, to address the issue of significant accuracy degradation under real bandwidth constraints, we propose that compensating intermediate collaboration results with late fusion is a feasible solution. As shown in Fig. 1a, late fusion achieves much higher accuracy than Where2comm [11] even with a bandwidth of approximately 0.7 Mbps, compared to Where2comm's accuracy with bandwidth less than 5 Mbps. So the use of intermediate-late hybrid collaboration can greatly improve the accuracy lower bound under low-bandwidth conditions.

Based on these insights, we propose CoSDH, a novel communication-efficient collaborative perception framework for 3D object detection. As shown in Fig. 2, CoSDH consists of three key components: i) Supply-demand-aware information selection, which chooses sparse yet crucial regions for collaboration; ii) Intermediate feature transmission and fusion, which transmits and effectively aggregates information from multiple agents in a communication-efficient manner; iii) Confidence-aware late fusion, which compensates for the intermediate fusion results at a minimal communication cost to improve accuracy. To evaluate CoSDH, we conducted extensive experiments on the simulation datasets OPV2V [36], V2XSim [20], and the real-world dataset DAIR-V2X [42]. The experimental results show that i) In the baseline case, CoSDH uses less bandwidth while achieving satisfactory accuracy, demonstrating a better accuracy-bandwidth trade-off; ii) Under the simulated real-world communication condition with a total bandwidth limit of 27 Mbps, CoSDH experiences less accuracy degradation and outperforms other methods while using less bandwidth.

In summary, our contributions are as follows:

- We present CoSDH, an innovative communication-efficient collaborative perception framework with better accuracy-bandwidth trade-offs for 3D object detection.
- We propose a novel supply-demand-aware information selection module, further refining the collaboration area selection to achieve more efficient communication.
- We design a novel intermediate-late hybrid collaborative perception paradigm, where confidence-aware late fusion compensates for the intermediate fusion results to maintain high accuracy under low-bandwidth conditions.
- We conduct extensive experiments across multiple datasets with bandwidth constraints set closer to real-world conditions, and CoSDH achieves state-of-the-art detection accuracy along with optimal bandwidth trade-offs, demonstrating its feasibility in real-world collaborative perception scenarios.

2. Related Work

2.1. 3D Object Detection

3D object detection, a key technology in autonomous driving, identifies and localizes objects in 3D scenes using environmental data from onboard sensors and can be categorized into image-based, point-cloud-based, and multimodal-fusion-based methods [26]. Image-based methods can be further classified into monocular [16, 33], stereo [6, 18], and multi-view [13, 44, 45] approaches based on the number of onboard cameras. These methods leverage the rich color and texture information of image, along with its dense data representation, to achieve cost-effective 3D perception for autonomous driving. However, image-based methods are inherently limited by the lack of depth information, which restricts their performance. In contrast, point-cloud-based methods, which adopt voxel-based [7, 39, 48], raw point cloud [27, 28, 46], or point-voxel hybrid [5, 29, 30] representations, can fully leverage the precise 3D geometric information provided by LiDAR, significantly improving perception capabilities. Considering that point-cloud-based methods suffer from sparse characteristic and lack rich semantic information, multimodal-fusion-based meth-

ods [23, 41, 47] further enhance performance by leveraging the complementary advantages of different modalities.

However, these single-agent-based 3D object detection methods are limited by sensor range and susceptible to occlusion, making them unable to detect objects that are further away or completely occluded. This paper focuses on collaborative-perception-based 3D object detection methods, which improve detection performance by supplementing the limitations of single-agent detection with information from other agents.

2.2. Collaborative Perception

Collaborative perception can be categorized into early collaboration, intermediate collaboration, and late collaboration based on the collaboration timing. Early collaboration [2, 4, 42] shares raw perception data, providing good perception accuracy but with high bandwidth. Late collaboration [20, 36] shares perception results, significantly reducing bandwidth, but leads to a decline in perception accuracy. Intermediate collaboration operates at the feature level and can achieve a better trade-off between accuracy and bandwidth by adjusting the intermediate features transmitted [3, 10, 11, 19, 21, 22, 32, 35–37], which is why it has been widely studied. Some of this research has focused on improving perception accuracy. FCooper [3] and CoFF [10] used manual modeling to fuse multi-agent features. Who2com [22] and When2com [21] perform selective communication and use attention-based fusion. V2VNet [32] and DiscoNet [19] employ communication graph-based fusion methods. However, these methods typically transmit complete, uncompressed intermediate BEV features, leading to enormous bandwidth requirements, which makes them challenging to apply in practice.

To address the large bandwidth demand in collaborative perception, AttFuse [36] was the first to use autoencoders to compress intermediate features along the channel dimension, which was later adopted by V2X-ViT [35], CoBEVT [37] and others. However, this method leads to significant accuracy degradation at high compression rates. Where2comm [11] reduces bandwidth requirements by selecting sparse but important foreground regions for collaboration while maintaining perception accuracy. However, as bandwidth is further limited, perception accruacy still rapidly degrades, potentially even falling below the accuracy of simple late collaboration methods. To address this issue, this paper proposes a hybrid collaborative method based on both intermediate and late collaboration, which efficiently compensates for the intermediate collaborative results using late collaboration under bandwidth constraints. Furthermore, we also adopt and improve methods based on auto-encoders and information selection to save bandwidth while maintaining accuracy.

3. Method

3.1. Problem Definition

In this paper, we consider the problem of collaborative perception with N agents. Let X_i and y_i represent the raw observation and the corresponding ground truth supervision of the i-th agent, respectively, and let $P_{j\to i}$ be the message sent from agent j to agent i. In the collaborative perception, agent i aggregates its own observations and the messages $\{P_{j\to i}\}_{j=1}^{N}$ sent from other agents to perform 3D object detection task. Our goal is to maximize the collaborative perception 3D object detection accuracy while ensuring that each agent has a communication budget B:

$$\xi_\Phi(B) = \arg\max_{\theta} \sum_{i=1}^{N} g(\Phi_\theta(X_i, \{P_{j\to i}\}_{j=1}^{N}), y_i), \quad (1)$$

$$\text{s.t.} \sum_{j=1}^{N} |P_{j\to i}| \le B \quad (2)$$

where Φ_θ represents the collaborative perception 3D object detection model, θ denotes the model parameters, $|P_{j\to i}|$ represents the communication volume of the message sent from agent j to agent i, and $g(\cdot,\cdot)$ denotes the 3D object detection evaluation metric.

3.2. Overall Architecture

The overall architecture of the proposed CoSDH is shown in Fig. 2. Each agent first processes its locally observed point cloud X_i through a backbone network based on PointPillar [17] and a demand generator to obtain multi-scale BEV features $\{F_i^{(l)}\}_{l=1,2,...,L}$ and a demand mask D_i, respectively. Considering the collaboration between Ego agent i and collaborating agent j, agent j generates a supply mask S_j from its multi-scale features $\{F_j^{(l)}\}_{l=1,2,...,L}$ via a supply generator, and multiplies it element-wise with the received demand matrix to obtain the supply-demand mask $M_{j\to i}$. Agent j then performs supply-demand-aware information selection by multiplying $\{F_j^{(l)}\}_{l=1,2,..,L}$ with $M_{j\to i}$ element-wise to obtain sparse spatial features $\{Z_j^{(l)}\}_{l=1,2,..,L}$. Subsequently, agent j compresses the features through an autoencoder, sending the non-zero parts of the features along with their corresponding coordinates as the message $P_{j\to i}$ to agent i.

Upon receiving $P_{j\to i}$, agent i first decodes the features to restore their dimensions and then fuses them with its local features $\{F_i^{(l)}\}_{l=1,2,...,L}$ across multiple scales to obtain the fused features $\{\tilde{F}_i^{(l)}\}_{l=1,2,...,L}$, which are then passed to the detection head for intermediate collaborative detection results $\tilde{y}_i$. Afterward, we apply confidence-aware late fusion: agent j filters and suppresses its own detection results y_j based on confidence and sends them to agent i for late fusion, yielding the final hybrid collaborative perception detection results $\hat{y}_i$.

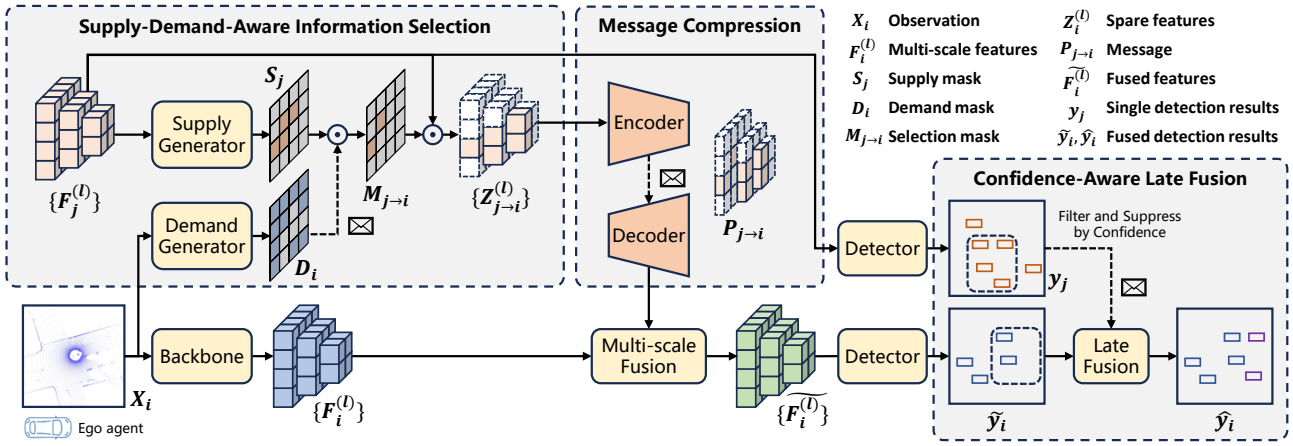


Figure 2. The overall architecture of CoSDH. The Supply-Demand-Aware Information Selection module selects sparse but important information, which is then further compressed by the Message Compression module to achieve efficient communication. Confidence-Aware Late Fusion compensates for the intermediate fusion detection results to improve accuracy.

3.3. Supply-Demand-Aware Information Selection

Previous methods such as Where2comm [11] and How2comm [40] generally use symmetric supply-demand relationships to select sparse features for collaboration, believing that areas with high foreground confidence in the collaborating agent's view should be provided, and conversely, areas with low foreground confidence in the Ego agent's view need collaboration. However, we believe that areas with low foreground confidence from the Ego's perspective can be divided into areas that are hard to observe and those that can be observed but belong to the background. The latter do not require collaboration. Based on this insight, we propose a novel supply-demand-aware information selection method.

The demand mask D_i indicates where agent i needs information from collaborating agents. Intuitively, the agent requires information from areas that are distant or occluded, which have the common characteristic of having low point cloud density or no point cloud at all. For agent i, we consider using the number of point clouds in each pillar to represent point cloud density, and we map it to the range $[0,1]$, i.e., $A_i \in [0,1]^{H\times W}$, where H and W represent the number of Pillars along the height and width dimensions. We then select areas where the point cloud density is below a threshold ε_a to obtain the demand mask for agent i, $D_i = A_i < \varepsilon_a \in \{0,1\}^{H\times W}$. The demand mask indicates where agent i has poor perception and needs collaborative information from other agents. Filtering information from other agents using the demand mask not only helps save bandwidth but also avoids interference from other agents' information in well-perceived areas.

For the object detection task, foreground information is more valuable. Providing sparse foreground features can effectively assist other agents in supplementing undetected and incomplete targets while using less bandwidth. Following previous work [11], we use the spatial confidence map $C_i \in [0,1]^{H\times W}$ output by the detection head to select potential foreground areas. We use a supply threshold ε_c to obtain the supply mask $S_i^{(l)} = C_i > \varepsilon_c \in \{0,1\}^{H\times W}$. By adjusting the threshold, we can dynamically adjust the bandwidth used for collaborative perception to adapt to varying communication conditions.

During collaboration, agent j generates a binary supply-demand selection mask $M_{j\to i} = D_i \odot S_j \in \{0,1\}^{H\times W}$ based on its supply mask S_j and agent i's demand mask D_i, and samples it to multiply element-wise with multi-scale BEV features $\{F_j^{(l)}\}_{l=1,2,\ldots,L}$, obtaining sparse features $\{Z_{j\to i}^{(l)}\}_{l=1,2,\ldots,L}$. During communication, only the non-zero parts and their corresponding coordinates need to be transmitted.

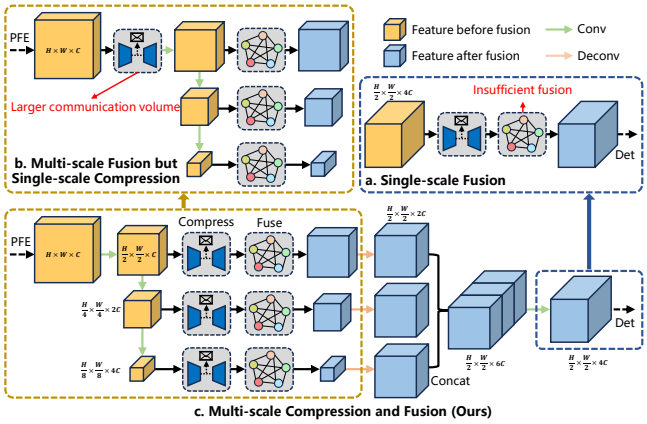


Figure 3. Comparison of our multi-scale compression and fusion with other methods. It can achieve thorough fusion with smaller communication volume.

3.4. Message Compression and Fusion

To further reduce communication bandwidth, we also use autoencoders to compress intermediate features along the channel dimension before communication, while making further improvements to existing compression and fusion schemes. As shown in Fig. 3, traditional single-scale fusion schemes [35, 36] only compress and fuse a single-layer BEV feature map before passing it to the detection head, resulting in insufficient fusion. CoAlign [25] proposed a solution that performs layer-wise fusion on multi-scale features, addressing the issue of insufficient fusion. However, it compresses large-scale single-layer features extracted after Pillar Feature Extraction (PFE), which leads to higher bandwidth. Therefore, we propose a multi-scale compression and fusion approach, where separate autoencoders are designed for each scale of features to perform compression, followed by layer-wise fusion. This approach enables more thorough fusion while using less bandwidth.

For message fusion, we use Max fusion, which has two main advantages: i) it is computationally simple and efficient, with the complexity increasing linearly with the number of agents; ii) max fusion selects the maximum value of features from multiple agents, achieving information complementarity. Considering the collaboration between Ego agent i and collaborating agent j, the process of message compression and fusion can be expressed as:

$$Z'^{(l)}_{j\to i} = f^{(l)}_{encode}(Z^{(l)}_{j\to i}) \in \mathbb{R}^{\frac{C_l}{c_0}\times H_l\times W_l} \tag{3}$$

$$F^{(l)}_{j\to i} = f^{(l)}_{decode}(Z'^{(l)}_{j\to i}) \in \mathbb{R}^{C_l\times H_l\times W_l} \tag{4}$$

$$F'^{(l)}_{j\to i} = f_{transform}(F^{(l)}_{j\to i}, \xi_{j\to i}) \in \mathbb{R}^{C_l\times H_l\times W_l} \tag{5}$$

$$\tilde{F}^{(l)}_{i} = \max(F^{(l)}_{i}, \{F'^{(l)}_{j\to i}\}_{j\neq i}) \in \mathbb{R}^{C_l\times H_l\times W_l} \tag{6}$$

where $Z^{(l)}_{j\to i} \in \mathbb{R}^{C_l\times H_l\times W_l}$ represents the sparse features selected based on supply-demand relationships from the previous stage, and C_l, H_l, W_l denote the channel, height, and width dimensions of the l-th layer feature map. $f^{(l)}_{encode}, f^{(l)}_{decode}$ represent the encoder and decoder of the l-th layer autoencoder, c_0 is the compression ratio in the channel dimension, and $f_{transform}$ represents coordinate transformation. During communication, $\{Z'^{(l)}_{j\to i}\}_{l=1,2,..,L}$ is first converted from float32 to float16, and then only the non-zero parts and corresponding coordinates are transmitted to save bandwidth. Upon receiving the message, agent i decodes the feature dimensions, then aligns the features to its own coordinate system using the coordinate transformation matrix $\xi_{j\to i}$ to obtain $F'^{(l)}_{j\to i}$, and finally fuses its own features with the collaborating agent's features using max fusion to obtain the fused features $\tilde{F}^{(l)}_{i}$.

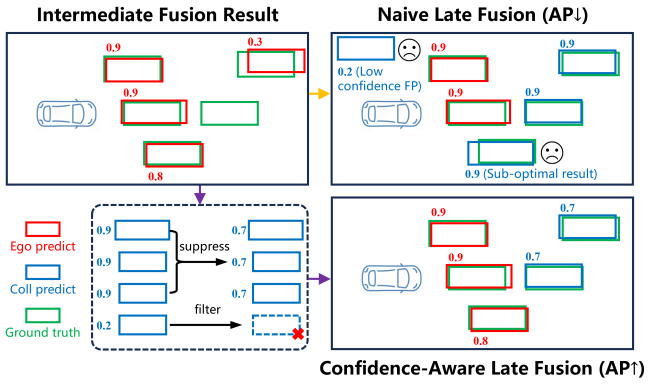


Figure 4. Our confidence-aware late fusion. It filters detection results based on confidence and suppresses suboptimal results from collaborative agents, improving overall detection accuracy.

3.5. Confidence-Aware Late Fusion

Existing collaborative perception methods focus on the singular intermediate collaboration architecture and neglect the advantages of late collaboration. As shown in Table 1, the experimental results indicate that late fusion can demonstrate acceptable perception accuracy with extremely low bandwidth on the OPV2V [36] and V2XSim [20] datasets. Therefore, late fusion can be used to compensate for the results of intermediate fusion, achieving a better accuracy-bandwidth trade-off.

Naive late collaboration methods [20, 36] directly merge results from other agents and then apply NMS (Non-Maximum Suppression) to deduplicate and obtain final results. This approach can effectively improve recall for collaborative perception object detection; however, during the merging process, it may introduce some low-confidence false positives, and suboptimal detection results from collaborating agents may override the Ego agent's detection results, lowering precision and consequently decreasing final AP. The lower AP in the late fusion method in Table 1 on the DAIR-V2X [42] dataset is due to this issue.

To address this problem, we propose a new confidence-aware late fusion method. As shown in Fig. 4, during late fusion, we first filter based on the confidence of the target boxes, discarding target boxes from other agents with confidence lower than ε_l. Furthermore, considering that if both the Ego agent and other agents detect the same target, the detection result from the Ego agent, which has undergone intermediate fusion, is of higher quality. Therefore, we suppress the detection results from other agents to prevent their lower-quality results from degrading the Ego agent's better detection results. Specifically, before merging the detection results, we multiply the confidence of target boxes from other agents by a coefficient $\beta \in (0,1)$.

Setting	Method	OPV2V [36]			V2XSim [20]			DAIR-V2X [42]		
		AP@0.5↑	AP@0.7↑	BD↓	AP@0.5↑	AP@0.7↑	BD↓	AP@0.5↑	AP@0.7↑	BD↓
Basic	No Fusion	79.78%	67.16%	0.0 Mbps	70.31%	58.55%	0.0 Mbps	66.51%	55.46%	0.0 Mbps
	Early Fusion	95.05%	88.98%	83.1 Mbps	95.68%	88.03%	55.5 Mbps	74.52%	59.22%	50.2 Mbps
	Late Fusion	94.70%	87.18%	0.2 Mbps	86.88%	78.13%	0.1 Mbps	67.86%	50.47%	0.2 Mbps
No Limit	When2com [21]	91.75%	81.77%	1,320.0 Mbps	72.65%	62.92%	250.0 Mbps	64.08%	49.14%	984.4 Mbps
	FCooper [3]	90.06%	74.03%	2,640.0 Mbps	72.96%	57.61%	500.0 Mbps	74.58%	56.47%	1,968.8 Mbps
	AttFuse [36]	94.31%	82.03%	2,640.0 Mbps	78.06%	64.84%	500.0 Mbps	73.80%	56.86%	1,968.8 Mbps
	V2VNet [32]	96.66%	92.44%	5,280.0 Mbps	88.97%	85.18%	1,000.0 Mbps	66.63%	47.39%	3,937.5 Mbps
	DiscoNet [19]	90.93%	78.90%	2,640.0 Mbps	77.34%	68.77%	500.0 Mbps	73.58%	58.45%	1,968.8 Mbps
	V2XViT [35]	95.87%	89.88%	2,640.0 Mbps	89.01%	80.26%	500.0 Mbps	76.68%	57.57%	1,920.0 Mbps
	Where2comm [11]	95.59%	91.39%	48.7 Mbps	88.18%	83.66%	27.5 Mbps	76.13%	60.16%	172.3 Mbps
	CoAlign [25]	96.63%	92.63%	2,640.0 Mbps	88.87%	85.23%	500.0 Mbps	**78.06%**	63.09%	1,968.8 Mbps
	CoSDH	**96.83%**	**92.99%**	**13.4 Mbps**	**89.23%**	**86.31%**	**1.1 Mbps**	76.75%	**63.85%**	**7.1 Mbps**
BD $\leq$ 6.75 Mbps	When2com [21]	79.97%	50.88%	5.2 Mbps	62.02%	43.37%	4.0 Mbps	62.37%	44.01%	3.8 Mbps
	FCooper [3]	90.37%	72.92%	5.2 Mbps	76.52%	61.83%	4.0 Mbps	67.71%	47.65%	3.8 Mbps
	AttFuse [36]	93.45%	80.02%	5.2 Mbps	84.58%	71.36%	4.0 Mbps	71.06%	49.57%	3.8 Mbps
	V2VNet [32]	95.71%	87.16%	5.2 Mbps	85.66%	73.72%	4.0 Mbps	66.49%	45.61%	3.8 Mbps
	DiscoNet [19]	90.00%	76.72%	5.2 Mbps	78.04%	68.18%	4.0 Mbps	70.83%	53.40%	3.8 Mbps
	V2XViT [35]	95.85%	87.13%	5.2 Mbps	88.94%	81.47%	4.0 Mbps	71.00%	52.78%	3.8 Mbps
	Where2comm [11]	94.91%	89.86%	5.4 Mbps	87.51%	82.12%	4.7 Mbps	74.98%	59.51%	5.3 Mbps
	CoAlign [25]	94.10%	85.99%	5.2 Mbps	88.01%	83.97%	4.0 Mbps	75.26%	60.19%	3.8 Mbps
	CoSDH	**96.75%**	**92.92%**	**2.0 Mbps**	**89.23%**	**86.31%**	**1.1 Mbps**	**76.47%**	**63.76%**	**1.4 Mbps**

Table 1. Comparison of detection accuracy and bandwidth of different methods on OPV2V [36], V2XSim [20], and DAIR-V2X [42] datasets. "BD" represents the bandwidth required for each collaborative agent, assuming Ego agent collaborates with up to 4 agents and the detection frequency is 10Hz. "BD$\leq$6.75 Mbps" is used to simulate real-world communication limits, assuming a total communication bandwidth of 27 Mbps, with each collaborating agent's bandwidth consumption limited to less than 6.75 Mbps. For intermediate collaboration methods without information selection, we provide a compressed version using autoencoders to meet the bandwidth constraints. For intermediate collaboration methods with information selection, the selection ratio is adjusted to meet the bandwidth constraints.

4. Experiments

4.1. Datasets and Experimental Settings

Datasets. 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. The datasets include both simulated and real-world scenarios, and cover two types of collaboration: V2V (Vehicle to Vehicle) and V2I (Vehicle to Infrastructure).

Evaluation Metrics We use the average precision (AP) with intersection-over-union (IoU) thresholds of 0.5 and 0.7 to evaluate the performance of different methods on 3D object detection. We assume the target detection frequency is 10Hz and calculate the communication bandwidth based on the average data transmitted by each collaborative agent to the Ego agent, in order to evaluate the communication cost of different methods. Specifically, we consider the bandwidth limitations in real-world collaborative perception scenarios, setting the vehicle's communication rate to 27 Mbps [1, 24]. Considering the typical case where the Ego agent collaborates with up to 4 other agents [36], the bandwidth limit for each collaborative agent is $27/4 = 6.75$ Mbps.

Implementation Our experiments are based on the OpenCOOD [36] framework, using PointPillar [17] as the encoder with a grid size of $(0.4m, 0.4m)$, and a maximum of 32 points per Pillar. For our method, we set the number of intermediate feature layers to $L = 3$, the demand threshold $\varepsilon_a = 4/32 = 0.125$, and the supply threshold $\varepsilon_c = 0.01$. For the OPV2V [36] and DAIR-V2X [42] datasets, the compression rate is $c_0 = 16$, and for the V2XSim [20] dataset, the compression rate is $c_0 = 8$ due to its smaller perception range and lower inherent bandwidth requirement. The late fusion threshold is set to $\varepsilon_l = 0.3$, with a suppression coefficient $\beta = 0.9$ for OPV2V and V2XSim dataset and $\beta = 0.8$ for DAIR-V2X dataset because of its difficulty. In late fusion, the dense prediction results before NMS are transmitted, as this reduces the computational burden on the collaborating agents. The experiments use the Adam [14] optimizer, with an initial learning rate set between 0.0001 and 0.002 based on the model's testing complexity to ensure proper training. The maximum number of collaborating agents is set to 5. The number of training epochs is set to 40 to ensure model convergence. Other experimental parameters are kept consistent with the OpenCOOD framework. All methods are trained on four NVIDIA GeForce RTX 3090 GPUs.

4.2. Quantitative Evaluation

Benchmark Comparison. Table 1 presents the collaborative 3D object detection accuracy and required bandwidth of the proposed `CoSDH` compared to previous methods across different datasets. Experimental results show that, with the default uncompressed settings, `CoSDH` achieves the highest accuracy on the OPV2V [36] and V2XSim [20] datasets while requiring only about 1/100 to 1/1000 of the bandwidth compared to other non-communication-efficient

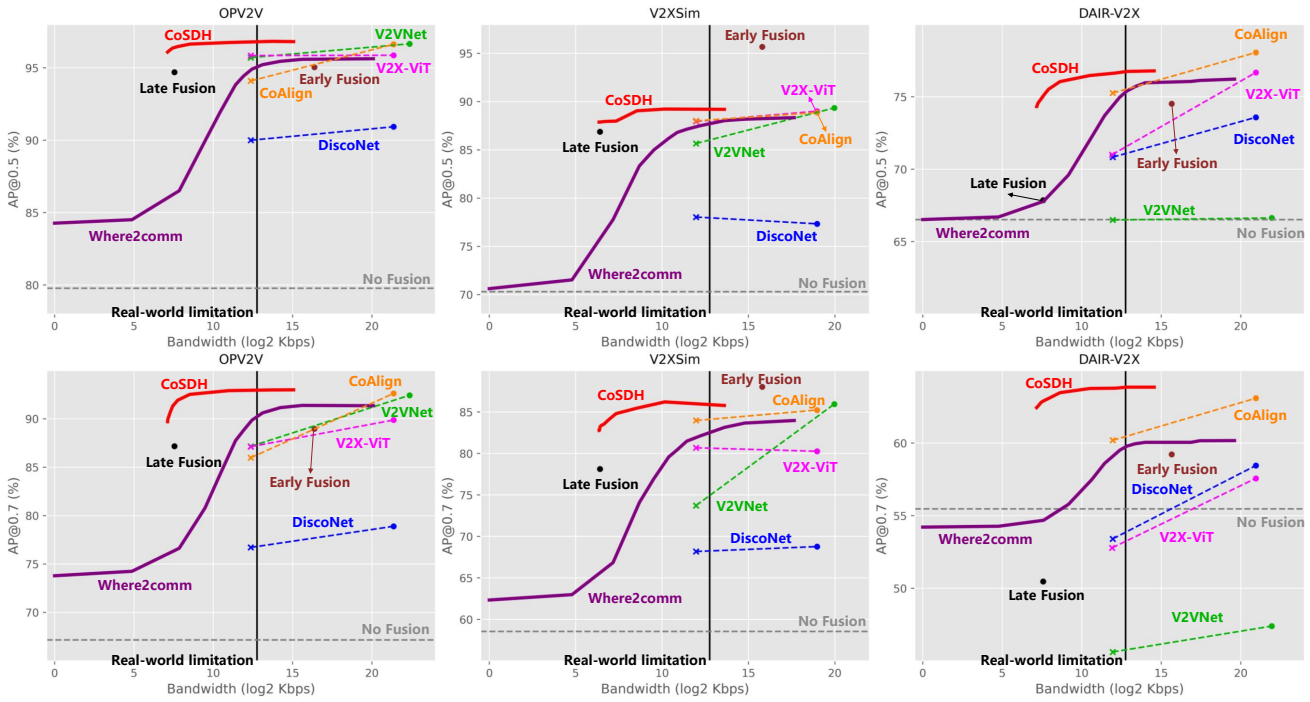


Figure 5. Comparison of the trade-off between detection accuracy and bandwidth of different methods on OPV2V [36], V2XSim [20] and DAIR-V2X [42] datasets, CoSDH achieves the best accuracy-bandwidth trade-off. The real-world limitation refers to the total bandwidth limit of 27 Mbps, which means that each collaborative agent does not exceed 6.75 Mbps.

methods. Although the AP@0.5 of CoSDH on the DAIR-V2X [42] dataset is slightly lower than that of CoAlign [25], it achieves a higher AP@0.7 with only about 1/300 of the bandwidth. Both our CoSDH and Where2comm [11] are communication-efficient methods based on information selection, which enable dynamic accuracy-bandwidth trade-offs by adjusting the selection ratio. The table shows accuracy at a specific bandwidth, and a more detailed comparison of the accuracy-bandwidth curves is provided in the "Accuracy-Bandwidth Trade-Off Comparison" part.

Furthermore, we simulate real-world communication rate limitations. Result shows that CoSDH achieves improvements in AP@0.7 of 3.06%/2.34%/3.75% on OPV2V/V2XSim/DAIR-V2X compared to the previous best methods, while using less bandwidth. When comparing the scenarios with and without bandwidth limitations on the OPV2V and DAIR-V2X datasets, we found that CoSDH achieves less than a 0.3% decrease in AP under conditions where bandwidth is reduced by 80% to 85%, exhibiting less accuracy degradation than Where2comm. Most other methods also show varying degrees of accuracy degradation under bandwidth constraints. Interestingly, under bandwidth limitations, some methods show an improvement in accuracy on certain datasets after using autoencoders to compress intermediate features. For example, FCooper [3] shows improved accuracy on the V2XSim dataset, which may be due to the model's initially poor performance. The compressed version, with the added autoencoder, increases the model's parameter count, thereby enriching its expressive capability.

Accuracy-Bandwidth Trade-Off Comparison. Fig. 5 shows the accuracy-bandwidth trade-off of the proposed CoSDH and previous advanced methods across different datasets. When the bandwidth grater than 1 Mbps, Where2comm [11] demonstrates a good enough accuracy-bandwidth trade-off, maintaining high detection accuracy as bandwidth decreases. However, as bandwidth further decreases, its accuracy quickly declines, even falling below that of late fusion method. CoSDH can leverage late fusion at low bandwidths to maintain high accuracy, achieving a better performance-bandwidth trade-off. Notably, on the DAIR-V2X dataset, the late fusion method performs worse than the no-fusion case because the naive late fusion method unselectively merges detection results from other agents, introducing a large number of low-quality detections. CoSDH uses a confidence-aware late fusion method to select higher-quality detection results, improving this situation and enhancing perception accuracy. We also observe that the AP@0.7 of CoSDH on the V2XSim dataset initially increases slightly as bandwidth decreases before subsequently declining, indicating that selecting key collaboration areas can help reduce interference from other regions to some extent, thereby slightly improving detection accuracy.

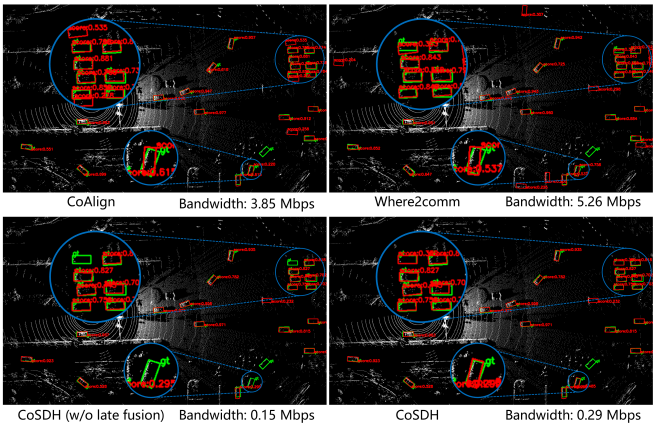


Figure 6. Visualization of detection results under real-world communication rate limitations on the DAIR-V2X [42] dataset. Green represents ground truth box, and red represents predicted box.

4.3. Qualitative Evaluation

Fig. 6 shows the visualization of detection results for our method compared to other methods under simulated real-world communication rate limitations on the DAIR-V2X dataset. A comparison of the detection results reveals that, while using less bandwidth, our method achieves the same recall rate as CoAlign [25] and Where2comm [11], with fewer false positive predictions and more accurate object localization, demonstrating that our method performs better under low-bandwidth conditions. Comparing the two images at bottom, it can be seen that under low bandwidth conditions, late fusion effectively compensates for the results of intermediate fusion, improving the recall rate of objects.

4.4. Ablation Studies

Table 2 presents the results of the ablation study on various modules of the proposed method using the OPV2V [36] dataset. The results show that after using the autoencoder for compression, the bandwidth is reduced by a factor of c_0, and there is no significant loss in perception accuracy, even with a slight improvement in AP@0.7. Converting intermediate features to float16 can nearly halve the bandwidth without significant loss, demonstrating the effectiveness of our proposed message compression module. After performing information selection based on the supply mask, perception accuracy slightly decreases, but the bandwidth is reduced by about 95%, which is a worthwhile trade-off. Further using the demand mask results in no significant decrease in accuracy but reduces bandwidth by about 10%. This demonstrates that our supply-demand-aware information selection can reduce bandwidth while maintaining accuracy.

After adding late fusion, accuracy improves at a relatively small bandwidth cost. However, since this phase uses more bandwidth, the improvement in accuracy is not significant. Table 3 shows the accuracy improvement due to late collaboration under different bandwidth conditions. It can be observed that late collaboration provides more accuracy improvements under lower bandwidth conditions. For example, when the bandwidth used for intermediate fusion is 0.05 Mbps, late collaboration with a bandwidth cost of 0.14 Mbps leads to a 3% improvement in AP@0.5 and a 4% improvement in AP@0.7. This is one of the key reasons why our method achieves significantly higher detection accuracy under lower bandwidth compared to Where2comm [11].

Compression		Selection		Late Fusion	AP@0.5↑	AP@0.7↑	BD↓
Autoencoder	FP16	Supply	Demand				
					96.62%	92.62%	1,155.00 Mbps
✓					96.59%	92.99%	72.19 Mbps
✓	✓				96.59%	92.99%	36.09 Mbps
✓	✓	✓			96.30%	92.62%	1.99 Mbps
✓	✓	✓	✓		96.31%	92.60%	1.82 Mbps
✓	✓	✓	✓	✓	96.75%	92.92%	1.97 Mbps

Table 2. Ablation study of the modules in CoSDH on the OPV2V dataset. "Autoencoder" refers to the use of autoencoders to compress features along the channel dimension, with a compression ratio of $c_0 = 16$. "FP16" refers to converting features from float32 to float16 for transmission. "Supply" and "Demand" refer to using supply and demand masks for information selection, and "Late Fusion" refers to applying confidence-aware late fusion.

Late Fusion	ε_c	0.01	0.02	0.03	0.05	0.07
	AP@0.5↑	96.59%	96.31%	96.19%	95.18%	93.35%
	AP@0.7↑	92.95%	92.60%	92.27%	90.67%	87.60%
	BD↓	13.26 Mbps	1.82 Mbps	0.54 Mbps	0.12 Mbps	0.05 Mbps
	AP@0.5↑	96.83%	96.75%	96.72%	96.54%	96.41%
✓	AP@0.7↑	92.99%	92.92%	92.73%	92.23%	91.68%
	BD↓	13.40 Mbps	1.97 Mbps	0.68 Mbps	0.26 Mbps	0.19 Mbps

Table 3. Ablation study of confidence-aware late fusion under different bandwidths on the OPV2V [36] dataset. The table shows the impact of late fusion on accuracy and bandwidth with different values of ε_c.

5. Conclusion

In this paper, we propose CoSDH, a novel communication-efficient collaborative perception framework for 3D object detection. By finely modeling the supply-demand relationship between agents, it selects key and sparse regions for collaboration. Additionally, we innovatively incorporate confidence-aware late fusion on top of intermediate collaboration to form an intermediate-late hybrid collaborative perception method. Experiments on multiple datasets show that our method offers a better accuracy-bandwidth trade-off, demonstrating outstanding accuracy under bandwidth constraints close to real-world communication limits, making it highly valuable for practical applications.

Acknowledgment

This work is supported by the National Key Research and Development Plan (2024YFB3309302).

References

[1] Fabio Arena and Giovanni Pau. An overview of vehicular communications. *Future internet*, 11(2):27, 2019. 1, 6

[2] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 1852–1864, 2020. 3

[3] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019. 1, 3, 6, 7, 2

[4] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019. 3

[5] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9775–9784, 2019. 2

[6] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12536–12545, 2020. 2

[7] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1201–1209, 2021. 2

[8] 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. 1

[9] Sebastian Gräfling, Petri Mähönen, and Janne Riihijärvi. Performance evaluation of ieee 1609 wave and ieee 802.11 p for vehicular communications. In *2010 second international conference on ubiquitous and future networks (ICUFN)*, pages 344–348. IEEE, 2010. 2

[10] Jingda Guo, Dominic Carrillo, Sihai Tang, Qi Chen, Qing Yang, Song Fu, Xi Wang, Nannan Wang, and Paparao Palacharla. Coff: Cooperative spatial feature fusion for 3-d object detection on autonomous vehicles. *IEEE Internet of Things Journal*, 8(14):11078–11087, 2021. 3

[11] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022. 1, 2, 3, 4, 6, 7, 8

[12] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15481–15490, 2024. 2

[13] Zheng Jiang, Jinqing Zhang, Yanan Zhang, Qingjie Liu, Zhenghui Hu, Baohui Wang, and Yunhong Wang. Fsd-bev: Foreground self-distillation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 2

[14] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012. 1

[16] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11867–11876, 2019. 2

[17] 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. 3, 6

[18] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 2

[19] 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. 3, 6

[20] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 1, 2, 3, 5, 6, 7

[21] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4106–4115, 2020. 3, 6

[22] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883. IEEE, 2020. 3

[23] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 3

[24] Jose Manuel Lozano Dominguez and Tomas Jesus Mateo Sanguino. Review on v2x, i2x, and p2x communications and their applications: a comprehensive analysis over time. *Sensors*, 19(12):2756, 2019. 1, 6

[25] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4812–4818. IEEE, 2023. 5, 6, 7, 8, 1, 2

[26] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A

comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023. 2

[27] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7463–7472, 2021. 2

[28] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[29] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 2

[30] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020. 2

[31] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17996–18006, 2024. 1

[32] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 605–621. Springer, 2020. 3, 6

[33] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 2

[34] 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. 1

[35] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 1, 3, 5, 6, 2

[36] 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, 3, 5, 6, 7, 8

[37] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning*, pages 989–1000. PMLR, 2023. 1, 3

[38] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 1, 2

[39] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2

[40] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[41] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024. 3

[42] 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. 2, 3, 5, 6, 7, 8, 1

[43] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 1

[44] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3348–3357, 2023. 2

[45] Jinqing Zhang, Yanan Zhang, Yunlong Qi, Zehua Fu, Qingjie Liu, and Yunhong Wang. Geobev: Learning geometric bev representation for multi-view 3d object detection. *arXiv preprint arXiv:2409.01816*, 2024. 2

[46] Yanan Zhang, Di Huang, and Yunhong Wang. Pc-rgnn: Point cloud completion and graph neural network for 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3430–3437, 2021. 2

[47] Yanan Zhang, Jiaxin Chen, and Di Huang. Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2022. 3

[48] Chao Zhou, Yanan Zhang, Jiaxin Chen, and Di Huang. Octr: Octree-based transformer for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5166–5175, 2023. 2