C2E: Boosting Ego-Only 3D Object Detection via Multi-Teacher Contrastive Knowledge Distillation

Jinlong Wang[1,2⋆], Xun Huang[1,2,3⋆], Qiming Xia[1,2], Shijia Zhao[1,2], and Chenglu Wen[1,2✉]

[1] Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University, China
[2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[3] Zhongguancun Academy, China

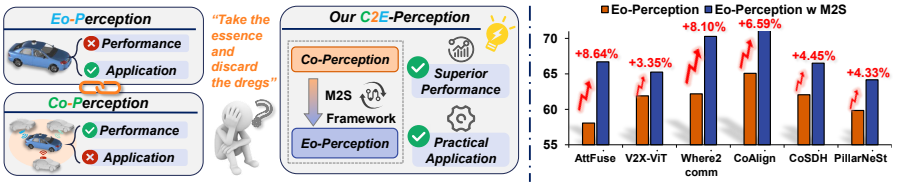
Abstract. LiDAR-based 3D object detection is essential for autonomous driving systems. However, traditional Ego-only Perception (Eo-Perception) suffers from limited perspective and occlusions in a complex outdoor environment, leading to performance bottlenecks. Recently, research on multi-agent Collaborative Perception (Co-Perception) has demonstrated excellent performance, but high communication costs and accumulated pose error hinder its application. To address this, we explore a novel **C2E** (**C**o-Perception **to E**o-Perception) paradigm through the **M**ulti-**to**-**S**ingle (**M2S**) agent contrastive knowledge distillation framework. Our M2S framework first designs Multi-Level Feature Enhancement module to provide more stable features, and introduces Auxiliary Point Cloud Reconstruction and Multi-Teacher Contrastive Distillation mechanisms to mitigate domain gaps in point cloud and feature distributions within the C2E paradigm. Benefiting from this, our M2S can retain the excellent performance of collaborative perception while effectively avoiding the drawbacks, such as communication delays and positioning errors. Extensive experiments on the V2XSet, V2V4Real and DAIR-V2X datasets show the effectiveness and generalizability of our M2S framework when combined with the state-of-the-art CoSDH model and other excellent 3D detectors. Our M2S framework can deliver up to a 8.64% improvement in 3D mAP performance without introducing any communication costs.

Keywords: Autonomous Driving · 3D Object Detection · Collaborative Perception

1 Introduction

3D object detection is a crucial task for autonomous driving and other unmanned systems [1, 10, 27]. Recently, LiDAR-based methods have achieved remarkable progress [18, 34] and have been widely adopted, owing to the precise spatial information provided by LiDAR sensors [14].

⋆ Equal contribution.
✉ Corresponding author: clwen@xmu.edu.cn.



(a) Comparison adv. and disadv. between EoP, CoP, and Our C2E. (b) The performance of models w/o and w M2S.

Fig. 1: (a) Comparison of the advantages and limitations between Eo-perception, Co-perception, and our proposed C2E-perception. (b) The 3D mAP@0.7 performance of SOTA models without and with our M2S framework on the V2XSet.

Outdoor environments, however, present complex and dynamic challenges, including diverse occlusions and extreme long-range detection requirements. These factors significantly impact Ego-only Perception (Eo-Perception), leading to inaccurate perception results [4,23]. While previous research [21,31,40,51,52] has attempted to alleviate these challenges of Eo-Perception by using more information or multi-sensor distillation, a noticeable performance bottleneck remains.

Recently, some efforts have been made to investigate Collaborative Perception (Co-Perception), such as vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-everything (V2X) [8,33,38,39]. Leveraging shared information among agents has demonstrated outstanding performance in the domain of cooperative 3D object detection. Nonetheless, the collaborative challenges that arise in real-world scenarios, such as high communication costs and localization errors, continue to hinder its real-world applications significantly.

Consequently, several existing methods have attempted to address the collaboration challenges by studying how, when, and where to communicate or by compressing communication features [7,17,43,44,49]. Nevertheless, despite some progress, these collaboration challenges still hinder the practical application of collaborative 3D object detection, primarily due to the restrictive nature of the existing collaboration mechanisms [6].

Based on existing research, we have summarized the strengths and weaknesses of these two perception paradigms. As illustrated in Fig. 1(a), Eo-Perception is practical but has performance bottlenecks, whereas Co-Perception delivers superior performance while encountering collaboration challenges. "*Take the essence and discard the dregs.*" Can the advantages of both be combined? This is the core question this paper addresses. To this end, we explore a novel perception paradigm, **C2E-Perception**: transferring the superior performance of Co-Perception to practical Eo-Perception. Specifically, at the training stage, we retain high-density information and collaborative knowledge from the Co-Perception teacher to the Eo-Perception student through the knowledge distillation technique. At the inference stage, we use the Eo-Perception student for individual inference to mitigate collaboration challenges.

However, simple knowledge distillation faces substantial domain gaps in data and feature distributions between single-agent and multi-agent data, which limits

the effective transfer of collaborative perception knowledge. To bridge these gaps, we introduce the Auxiliary Point Cloud Reconstruction (APCR) and Multi-Teacher Contrastive Distillation (MTCD) strategies to mitigate the significant discrepancy in data and feature distributions, respectively. Moreover, to provide more stable features for distillation, we propose Multi-Level Feature Enhancement (MLFE) to enhance the feature extraction capability of the student.

As shown in Fig. 1(b), extensive experiments conducted on both real-world and simulated datasets demonstrate the effectiveness and generalizability of our M2S framework when integrated with the state-of-the-art model [35] and other excellent 3D detectors. In summary, our main contributions are as follows.

- To the best of our knowledge, we are the first to explore the Co-Perception to Eo-Perception (**C2E-Perception**) paradigm. We propose the novel M2S framework based on multi-teacher adaptive contrastive knowledge distillation.
- We effectively mitigate the significant data and feature gaps between multi-agent and single-agent domain through Auxiliary Point Cloud Reconstruction (APCR) and Multi-Teacher Contrastive Distillation (MTCD). Moreover, we use the Multi-Level Feature Enhancement (MLFE) to provide more stable features for distillation.
- Extensive experiments on the V2XSet, V2V4Real, and DAIR-V2X datasets validate the effectiveness of the proposed M2S framework. Our M2S framework can deliver an improvement of up to 8.64 in 3D mAP performance without introducing any communication costs.

2 Related work

3D Object Detection for Single Agent. 3D object detection aims to identify object categories and spatial locations from sparse and irregular point clouds. Existing single-agent methods fall into three categories: point-based [25, 26], grid-based [10, 32, 41, 45], and point-voxel hybrid approaches [3, 24]. However, single-agent systems rely solely on onboard sensors, resulting in limited perception angles and susceptibility to occlusions, blind spots, and complex traffic dynamics [4]. Furthermore, the limited sensing range constrains distant object detection [6], causing performance bottlenecks in high-speed or sparse scenes.

Collaborative Perception for Multi-Agents. Collaborative perception has been proposed to overcome the limitations of a single viewpoint. Depending on the fusion stage, it can be categorized into early [2, 13, 28], intermediate [8, 20, 35, 38, 39], and late fusion [5, 36]. Despite its advantages, collaborative perception faces challenges such as high bandwidth demands for feature or raw data transmission [16] and accumulated errors from inaccurate relative pose estimation [6, 17]. DiscoNet [13] reduces communication costs via a distilled graph; V2X-INCOP [22] improves fault tolerance under interruptions, yet both remain in the collaborative paradigm with latency and infrastructure constraints. Thus, improving perception performance while reducing communication overhead remains a key research focus.

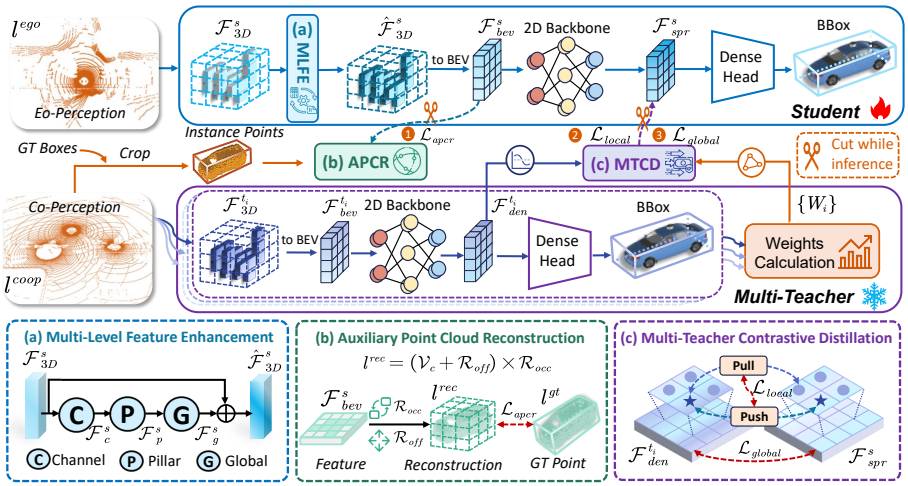


Fig. 2: M2S framework. Teachers (Bottom): Multi-agent point clouds l^{coop} are simultaneously input into three teachers, which are then compressed into BEV features $\{\mathcal{F}^{t_i}_{bev}\}^3_{i=1}$ via pillar transformation. Then a 2D backbone is used to extract dense features $\{\mathcal{F}^{t_i}_{den}\}^3_{i=1}$; Finally, the MTCD module fuses them into a distilled supervision feature map $\hat{\mathcal{F}}^t_{den}$. **Student** (Top): The ego-only point cloud l^{ego} is converted into pillars $\mathcal{F}^s_{3D}$ and enhanced by the MLFE to obtain $\hat{\mathcal{F}}^s_{3D}$. During training, $\hat{\mathcal{F}}^s_{3D}$ is projected to BEV space for voxel-level reconstruction via APCR, and the resulting features $\mathcal{F}^s_{spr}$ are guided by $\hat{\mathcal{F}}^t_{den}$ through multi-teacher adaptive contrastive distillation.

3D Object Detection Based on Knowledge Distillation. Knowledge distillation has been applied to 3D object detection [9,51,52] and recent collaborative perception [12, 13]. Multi-teacher schemes yield richer knowledge than single-teacher ones; MKD-Cooper [15] exploits this for robustness. Unlike [46], which refines other agents' predictions as pseudo-labels, our method distills collaborative knowledge directly into the ego model. We thus propose M2S, a multi-teacher adaptive contrastive distillation framework that transfers dense knowledge from multiple teachers to a sparse student.

3 Methodology

3.1 Problem Statement and Overall Design

Problem Statement. In collaborative perception based on V2X communication, there are three types of agents: Ego, Connected Automated Vehicles (CAVs), and Infrastructure units (Inf). In our proposed framework, the collaborative perception (Co-Perception) teacher detector $\mathcal{D}^t$ receives lidar point cloud input $l^{coop} = \{l^{ego}, l^{cav}, l^{Inf}\}$ from all agents. In contrast, the ego-only perception (Eo-Perception) student detector $\mathcal{D}^s$ only takes point cloud data $\{l^{ego}\}$ from the ego vehicle, thus enabling single-agent 3D object detection. The goal of 3D

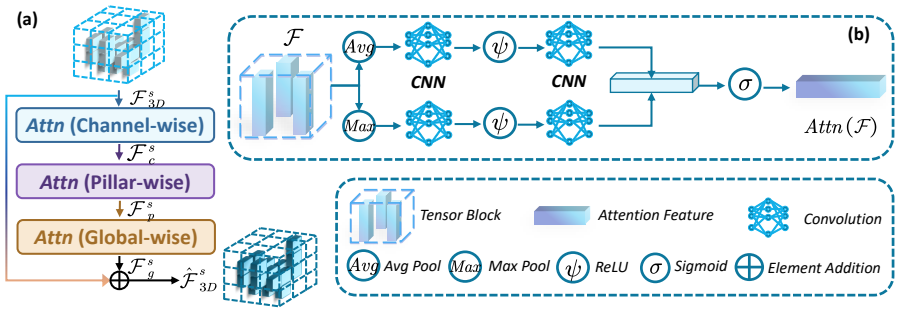


Fig. 3: Multi-Level Feature Enhancement. (a) Serial enhancement strategy. (b) Feature weight calculation ($Attn$). We apply a serial local-to-global enhancement to the input $\mathcal{F}^s_{3D}$ in three stages: channel-wise, pillar-wise, and global-wise.

object detection is to regress a set of 3D bounding boxes $\mathcal{B} = \{b_i\}_{i=1}^{N_b}$, where $\mathcal{B} \in \mathbb{R}^{N_b \times 7}$ and N_b denotes the number of predicted bounding boxes.

Overall Design. Due to the varying number of participating agents, large gaps exist between $\mathcal{D}^t$ and $\mathcal{D}^s$ in both feature representation and point cloud domains. To address this challenge, we propose a **M**ulti-to-**S**ingle (M2S) agent contrastive knowledge distillation framework, as illustrated in Fig. 2. First, the **M**ulti-**L**evel **F**eature **E**nhancement (MLFE) module (Fig. 2(a)) progressively enhances 3D features in $\mathcal{D}^s$ across multiple levels and dimensions. Then, the **A**uxiliary **P**oint **C**loud **R**econstruction (APCR) module (Fig. 2(b)) encourages the student to learn point cloud distributions consistent with $\mathcal{D}^t$. Finally, the **M**ulti-**T**eacher **C**ontrastive **D**istillation (MTCD) strategy (Fig. 2(c)) transfers locally structured and globally salient features from the teacher model to the student via contrastive learning and adaptive weighting, which further mitigates the feature disparity between $\mathcal{D}^t$ and $\mathcal{D}^s$. Note that only the student detector $\mathcal{D}^s$ is used during inference.

We provide the formal definitions of the three paradigms in the *Appx.* and the following sections provide detailed descriptions of each module.

3.2 Multi-Level Feature Enhancement

In the teacher $\mathcal{D}^t$, the ego vehicle obtains high-density features through communication with other agents. In contrast, the student $\mathcal{D}^s$ relies solely on its own sensors, producing relatively sparse features that limit 3D detection performance. To enhance the feature extraction capability of $\mathcal{D}^s$ and provide stable features for distillation, we propose the **M**ulti-**L**evel **F**eature **E**nhancement (MLFE) module, which progressively refines 3D features $\mathcal{F}^s_{3D}$ from local to global levels.

As shown in Fig. 3, MLFE enhances the student's 3D features along three dimensions: *Channel*, *Pillar*, and *Global*, through hierarchical attention mechanisms. The student detector $\mathcal{D}^s$ receives a 3D feature tensor $\mathcal{F}^s_{3D} \in \mathbb{R}^{C \times P \times G}$, where C, P and G denote channels, points per pillar, and the numbers of global

pillars, respectively. By progressively re-weighing semantic channels, reinforcing local geometric consistency, and integrating global contextual information, MLFE enhances students' representational capabilities, thereby generating more robust and domain-invariant features for distillation.

Inspired by SCNet3D [11] and CBAM [30], MLFE applies feature weight calculation ($Attn$) to generate three attention maps:

$$\begin{aligned} Attn(\mathcal{F}) = \sigma(\delta(\psi\left(\delta\left(\theta_{max}(\mathcal{F})\right)\right) \\ + \delta\left(\psi\left(\delta\left(\theta_{avg}(\mathcal{F})\right)\right)\right), \end{aligned} \tag{1}$$

where θ_{max} and θ_{avg} denote max_pooling and average_pooling, δ denotes convolution, ψ denotes ReLU activation, and σ denotes sigmoid activation.

Then, we apply $Attn$ to $\mathcal{F}^s_{3D}$ along the channel, pillar, and global dimensions sequentially to compute the attention map and refine features step-by-step as:

$$\begin{aligned} \mathcal{F}^s_c = \mathcal{F}^s_{3D} \otimes Attn(\mathcal{F}^s_{3D}) \oplus \mathcal{F}^s_{3D}, \\ \mathcal{F}^s_p = \mathcal{F}^s_c \otimes Attn(\mathcal{F}^s_c) \oplus \mathcal{F}^s_c, \\ \mathcal{F}^s_g = \mathcal{F}^s_p \otimes Attn(\mathcal{F}^s_p) \oplus \mathcal{F}^s_p, \\ \text{and } \hat{\mathcal{F}}^s_{3D} = \mathcal{F}^s_{3D} \oplus \mathcal{F}^s_g, \end{aligned} \tag{2}$$

where $\otimes$ indicates element-wise multiplication and $\oplus$ denotes element-wise addition. The final output $\hat{\mathcal{F}}^s_{3D}$ integrates attention-weighted 3D feature along all dimensions.

3.3 Auxiliary Point Cloud Reconstruction

The teacher detectors $\mathcal{D}^t$ receives dense point cloud input l^{coop} from multi-agent cooperation, while the student $\mathcal{D}^s$ only processes a sparse ego point cloud l^{ego}. This leads to a clear disparity in the **point cloud distribution domains** between the two networks. To address this gap, we introduce an **A**uxiliary **P**oint **C**loud **R**econstruction module (**APCR**) for the student model during the training phase. Unlike prior works [29, 48, 50], which focus on scene-level reconstruction, our method performs instance-level reconstruction between teacher and student. Specifically, we extract point clouds within ground-truth bounding boxes and re-voxelize them as supervision targets.

Since directly recovering dense point clouds incurs substantial computational and memory overhead, inspired by [29], we decompose the auxiliary point cloud reconstruction task into two sub-tasks: voxel occupancy mask $\mathcal{R}_{occ}$ prediction and point offset $\mathcal{R}_{off}$ estimation. As illustrated in the Fig. 4, APCR first performs dimensionality expansion on the 2D feature $\mathcal{F}^s_{bev}$, projecting it back into 3D space, and then processes it through multiple layers of 1×1 3D convolution with GELU activation to generate $\mathcal{R}_{occ}$ and $\mathcal{R}_{off}$.

Thus, the reconstructed point cloud can be calculated as:

$$l^{rec} = (\mathcal{V}_c + \mathcal{R}_{off}) \times \mathcal{R}_{occ}, \tag{3}$$

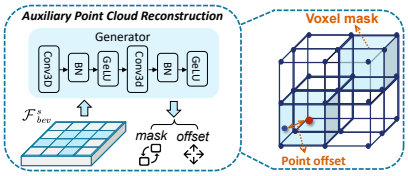


Fig. 4: Auxiliary Point Cloud Reconstruction Task. We project $\mathcal{F}^s_{bev}$ back into 3D space and decouple it into voxel occupancy prediction and point offset estimation to achieve voxel-level point cloud reconstruction.

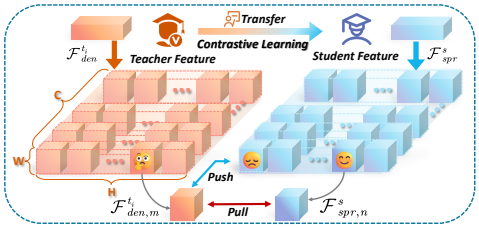


Fig. 5: Multi-Teacher Contrastive Learning. We first partition the feature map into non-overlapping patches of equal size, then perform spatial and channel comparisons within each local patch.

where l^{rec} is the reconstructed dense point cloud and $\mathcal{V}_c$ denotes to the center of voxel.

We supervise the voxel occupancy prediction mask using a binary cross-entropy H loss, and the point offset estimation using an L1 regression loss. They are defined as:

$$\begin{aligned} \mathcal{L}_{occ} &= -\frac{N_{bg}}{N_{fg}} H(\mathcal{G}_{occ}, \mathcal{P}_{occ}), \\ \mathcal{L}_{off} &= \frac{1}{|N_{fg}|} \sum_{i=1}^{N_{fg}} \|(\mathcal{V}_{c_i} + \mathcal{R}_{off_i}) - O_{gt_i}\|, \end{aligned} \tag{4}$$

where $\mathcal{G}_{occ}$ denotes the ground-truth occupancy annotation and $\mathcal{P}_{occ}$ denotes the predicted occupancy probability; N_{fg} and N_{bg} are the numbers of foreground and background voxels; O_{gt_i} represents the ground-truth point cloud of the i-th foreground voxel.

The final reconstruction loss is the sum of occupancy loss and offset loss:

$$\mathcal{L}_{apcr} = \mathcal{L}_{occ} + \mathcal{L}_{off}. \tag{5}$$

3.4 Multi-Teacher Contrastive Distillation

Multi-agent teachers contain richer contextual and structural knowledge, while the student's single-view limitation creates a **feature gap** that hinders the direct acquisition of such dense features. To bridge this gap and effectively transfer locally structured and globally salient features from teacher detectors $\mathcal{D}^t$ to the target student model $\mathcal{D}^s$, we propose the Multi-Teacher Contrastive Distillation (MTCD) strategy.

As shown in Fig. 5, we partition each BEV feature map into multiple non-overlapping regions, which are treated as contrastive samples. Unlike label-based contrastive frameworks [42], the positive and negative pairs are constructed

within teacher–student feature correspondences $(\mathcal{F}^t_{den}, \mathcal{F}^s_{spr})$, based on their spatial alignment. Formally, a positive pair is defined for identical spatial indices $(m = n)$ between $(\mathcal{F}^s_{spr,m}, \mathcal{F}^t_{den,n})$, where m and n denote the spatial locations on the BEV feature map. Conversely, negative pairs correspond to index mismatches $(m \neq n)$.

The contrastive loss $\mathcal{L}^{t_i}_{CL}$ for the i-th teacher is:

$$\mathcal{L}^{t_i}_{CL} = \frac{1}{N} \sum_{m=1}^{N} -\log \frac{\exp(-d^{t_i}_{m,m}/\tau)}{\sum_{n=1}^{N} \mathbb{1}_{m \neq n} \exp(-d^{t_i}_{m,n}/\tau)}, \tag{6}$$

Where $\tau = 0.07$ is the temperature parameter, N denotes the number of contrastive pairs, and $\mathbb{1}_{m \neq n}$ is an indicator function that equals 1 only when $m \neq n$. The feature similarity is measured using the squared Euclidean distance: $d^{t_i}_{m,m} = \|\mathcal{F}^s_{spr,m} - \mathcal{F}^{t_i}_{den,m}\|_2^2$ and $d^{t_i}_{m,n} = \|\mathcal{F}^s_{spr,m} - \mathcal{F}^{t_i}_{den,n}\|_2^2$.

To enable the student to capture the teacher's locally dense and structured knowledge by focusing on contextual cues and position–channel group information, we combine two complementary contrastive modes: Spatial Contrasting and Channel Contrasting. The detailed formulation is provided in the *Appx.*.

Accordingly, we use $\mathcal{L}_{local}$ to emphasize fine-grained local structural alignment, which is defined as:

$$\mathcal{L}_{local} = \sum_{i=1}^{T} W_i \cdot \mathcal{L}^{t_i}_{CL}, \tag{7}$$

where W_i is the confidence weight for the i-th teacher, based on multi-teacher adaptive distillation mechanism.

Multi-Teacher Adaptive Distillation. Due to variations in perceptual capabilities across different teacher models under diverse scenarios or samples, direct averaging fusion may introduce noise or even misguide training. To achieve high-quality knowledge transfer, we further designed a Multi-Teacher Adaptive Distillation mechanism. This design dynamically assigns weights based on the teacher model's performance on the current sample.

Specifically, we use the localization loss of 3D bounding box prediction as the performance indicator. For the i-th teacher, the localization loss $\mathcal{L}^{t_i}_{loc}$ is computed by Smooth L_1 loss between the predicted box $\hat{\mathcal{B}}_{t_i}$ and the ground truth $\mathcal{B}_{gt}$: $\mathcal{L}^{t_i}_{loc} = \text{Smooth}_{L_1}\left(\hat{\mathcal{B}}_{t_i}, \mathcal{B}_{gt}\right)$.

Then, the confidence weight W_i is obtained by the localization loss across all teachers:

$$W_i = \frac{\exp(-\alpha \cdot \mathcal{L}^{t_i}_{loc})}{\sum_{j=1}^{T} \exp(-\alpha \cdot \mathcal{L}^{t_j}_{loc})}, \tag{8}$$

where $\alpha = 0.8$ is a scaling factor that controls the sensitivity of the confidence to the loss. This weighting scheme prioritizes teachers with higher confidence.

Moreover, to further guide the student $\mathcal{D}^s$ to approximate the comprehensive dense BEV feature $\hat{\mathcal{F}}^t_{den,k}$ of teachers $\mathcal{D}^t$ in global level, we employ a

multi-teacher adaptive distillation approach. Specially, we consider T pre-trained teacher models $\mathcal{D}^{t_i}$, which generate dense features. For the k-th layer, let $\mathcal{F}^{t_i}_{den,k} = \mathcal{D}^{t_i}(l^{coop})$ denote the dense BEV features of the i-th teacher. We fused these features into $\hat{\mathcal{F}}^{t}_{den,k}$ by:

$$\hat{\mathcal{F}}^{t}_{den,k} = \sum_{i=1}^{T} W_i \cdot \mathcal{F}^{t_i}_{den,k}, \tag{9}$$

where W_i is the confidence weight for the i-th teacher, which is computed based on the detection performance of the teacher.

Finally, we employ KL-divergence feature distillation as:

$$\mathcal{L}_{global} = \sum_{k=1}^{K} \beta^{k}_{kd} \cdot \mathcal{D}_{KL}\left(\rho(\mathcal{F}^{s}_{spr,k}) \parallel \rho(\hat{\mathcal{F}}^{t}_{den,k})\right), \tag{10}$$

where ρ denotes softmax function, and $\mathcal{D}_{\mathrm{KL}}(\pi \| p) = \mathbb{E}_{\pi(x)}[\log \frac{\pi(x)}{p(x)}]$ is the Kullback-Leibler divergence, and the hyperparameter $\beta^{k}_{kd} \in \{0.2, 0.3, 0.5\}$ to balance weights across different feature layers, which set same to [15].

3.5 Loss Function

We train our M2S using the following loss function:

$$\begin{aligned} \mathcal{L}_{M2S} = \gamma_{cls}\mathcal{L}_{cls} + \gamma_{loc}\mathcal{L}_{loc} + \gamma_{apcr}\mathcal{L}_{apcr} \\ + \gamma_{global}\mathcal{L}_{global} + \gamma_{local}\mathcal{L}_{local}, \end{aligned} \tag{11}$$

where the hyperparameters γ_{cls}, γ_{loc}, γ_{apcr}, γ_{global} and γ_{local} balance the contributions of different losses, and $\mathcal{L}_{cls}$ denotes the classification loss implemented with focal loss, $\mathcal{L}_{loc}$ represents the bbox regression loss using Smooth L_1.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We conduct comprehensive experiments on two large-scale real-world datasets, V2V4Real [37] and DAIR-V2X [47], as well as a simulated dataset, V2XSet [38]. We provide more datasets information in the *Appx.*.

Evaluation Metrics. We adopt the average precision (AP) as the main indicators for evaluating 3D object detection performance. Additionally, we record the delay and frames per second (FPS) to measure the communication latency and inference efficiency, respectively. And we report the number of floating-point operations (FLOPs) to assess the computational complexity during inference. Furthermore, we use single NVIDIA RTX 3090 GPU to report training time (hours) and GPU memory usage (MB) and discuss the training cost.

Table 1: Results on the V2XSet. We report the performance of different models on the V2XSet with **single-agent input**. Improvements are highlighted in **bold**.

Models	Publication	3D mAP@Validation			3D mAP@Testing		
		IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
1.AttFuse	ICRA 2022	78.89	76.72	68.10	75.67	73.72	58.06
2.AttFuse w/ M2S		85.20	82.83	73.73	82.15	79.26	66.70
Improvement (2-1)		**+6.31**	**+6.11**	**+5.63**	**+6.48**	**+5.54**	**+8.64**
1.V2X-ViT	ECCV 2022	78.20	76.84	67.51	78.86	76.17	61.90
2.V2X-ViT w/ M2S		83.97	82.45	72.49	83.91	81.02	65.25
Improvement (2-1)		**+5.77**	**+5.61**	**+4.98**	**+5.05**	**+4.85**	**+3.35**
1.Where2comm	NeurIPS 2022	80.22	78.23	68.44	79.21	76.24	62.18
2.Where2comm w/ M2S		84.99	83.26	75.15	84.42	82.32	70.28
Improvement (2-1)		**+4.77**	**+5.03**	**+6.71**	**+5.21**	**+6.08**	**+8.10**
1.CoAlign	ICRA 2023	80.60	78.63	69.74	80.62	78.02	65.07
2.CoAlign w/ M2S		88.11	86.27	77.15	86.49	83.88	71.66
Improvement (2-1)		**+7.51**	**+7.64**	**+7.41**	**+5.87**	**+5.86**	**+6.59**
1.CoSDH	CVPR 2025	80.74	78.53	67.91	79.52	76.75	62.05
2.CoSDH w/ M2S		87.00	84.18	75.13	84.81	81.39	66.50
Improvement (2-1)		**+6.26**	**+5.65**	**+7.22**	**+5.29**	**+4.64**	**+4.45**
1.PillarNeSt	TIV 2024	79.14	77.66	67.85	77.05	74.29	59.83
2.PillarNeSt w/ M2S		85.94	84.09	71.82	84.14	81.53	64.16
Improvement (2-1)		**+6.80**	**+6.43**	**+3.97**	**+7.09**	**+7.24**	**+4.33**

4.2 Baselines Selection and Experimental Details

Baselines Selection. We use traditional ego-only methods, for instance, PillarNeSt [19] to validate our M2S framework and also run the collaborative baselines, including AttFuse [39], V2X-ViT [38], Where2comm [7], CoAlign [17], and CoSDH [35], with ego-only input to examine performance in realistic ego-only perception scenarios. Overall, our experimental setup was designed to fully validate the performance of our M2S across different perception methods.

Experimental Details. We use the ego agent for evaluation, following the settings of [39]. We employ three pre-trained models as teachers: AttFuse [39], V2X-ViT [38], and CoSDH [35]. We implement a teacher network that receives only the multi-agent input and a student network that receives only the ego-agent input. During training, the student network is constructed following the full collaborative architecture, including the fusion modules, but receives only the ego-agent input. During inference, only the student detector, which takes only the ego-agent input, is used for prediction to ensure a fair comparison. We provide more experimental details in the *Appx.*.

4.3 Performance Evaluation

Results on the V2XSet dataset. Table 1 compares several baselines with our M2S framework. Both the collaborative perception methods and ego-only perception method [19] baselines exhibit clear performance gains with our M2S. Notably, on the test set, M2S improves 3D mAP (IoU=0.7) by 8.64%, 3.35%, 8.10%, 6.59%, 4.45%, and 4.33% for AttFuse [39], V2X-ViT [38], Where2comm [7], CoAlign [17], CoSDH [35], and PillarNeSt [19], respectively. Most importantly,

Table 2: Results on the V2V4Real and DAIR-V2X. We show the performance of different models with **single-agent input**. Improvements are highlighted in **bold**.

Models	V2V4Real Val			V2V4Real Test			DAIR-V2X Val		
	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7
1.AttFuse	49.22	44.80	25.69	44.64	40.66	21.49	58.97	49.70	33.25
2.AttFuse w/ M2S	57.16	47.44	26.57	53.85	45.76	26.88	61.02	52.16	39.86
Improvement (2-1)	**+7.94**	**+2.64**	**+0.88**	**+9.21**	**+5.10**	**+5.39**	**+2.05**	**+2.46**	**+6.61**
1.V2X-ViT	52.00	46.98	25.30	46.59	41.40	19.56	60.71	51.06	35.08
2.V2X-ViT w/ M2S	58.78	47.44	26.77	54.66	44.49	20.85	60.65	52.77	39.77
Improvement (2-1)	**+6.78**	**+0.46**	**+1.47**	**+8.07**	**+3.09**	**+1.29**	-0.06	**+1.71**	**+4.69**
1.Where2comm	46.80	43.57	27.03	47.33	42.08	25.17	59.42	51.10	38.54
2.Where2comm w/ M2S	57.04	44.72	25.55	55.50	47.07	26.79	60.84	52.24	39.84
Improvement (2-1)	**+10.24**	**+1.15**	-1.48	**+8.17**	**+4.99**	**+1.62**	**+1.42**	**+1.14**	**+1.30**
1.CoAlign	46.67	41.75	23.77	48.54	40.94	21.48	61.26	51.26	37.86
2.CoAlign w/ M2S	55.16	45.45	24.13	54.27	46.84	23.74	62.77	53.14	38.75
Improvement (2-1)	**+8.49**	**+3.70**	**+0.36**	**+5.73**	**+5.90**	**+2.26**	**+1.51**	**+1.94**	**+0.95**
1.CoSDH	46.04	42.85	25.07	46.79	43.20	23.11	57.50	48.57	34.37
2.CoSDH w/ M2S	55.98	46.25	24.42	56.54	47.91	25.53	60.19	51.67	37.33
Improvement (2-1)	**+9.94**	**+3.40**	-0.65	**+9.75**	**+4.71**	**+2.42**	**+2.69**	**+3.10**	**+2.96**
1.PillarNeSt	50.61	45.36	21.70	47.90	40.20	16.18	57.30	48.90	33.46
2.PillarNeSt w/ M2S	52.19	45.28	26.03	53.99	43.55	25.24	60.36	51.85	38.17
Improvement (2-1)	**+1.58**	-0.08	**+4.33**	**+6.09**	**+3.35**	**+9.06**	**+3.06**	**+2.95**	**+4.71**

these gains are achieved with no extra communication overhead, validating our M2S's effectiveness and practicality.

Results on the real-world datasets. To further verify the effectiveness and generalization of our M2S in real-world scenarios, we conduct experiments on V2V4Real [37] and DAIR-V2X [47]. As with V2XSet [38], both collaborative [7, 17, 35, 38, 39] and ego-only [19] baselines gain consistently. Specifically, On V2V4Real test set at AP@0.7, all baselines consistently improve with M2S, with PillarNeSt achieving the largest gain of 9.06, followed by AttFuse (+5.39), CoSDH (+2.42), and CoAlign (+2.26). A similar trend is observed on the DAIR-V2X, where M2S consistently enhances performance across multiple models, confirming its robustness across diverse environments.

4.4 Ablation Study

Table 3 ablates each module in M2S. We adopt AttFuse [39], V2X-ViT [38], and CoSDH [35] as teacher models, with AttFuse that receives single-agent input as the student. Introducing single-teacher distillation (ST) alone yields a slight improvement in AP@0.7 from 58.06% to 58.29%. Replacing it with the multi-teacher adaptive distillation module (row 3) significantly boosts performance to 62.76%, demonstrating the benefit of leveraging multiple teachers. Adding the multi-level feature enhancement (MLFE) and auxiliary point cloud reconstruction (APCR) modules further increases AP@0.7 to 66.50%. Finally, incorporating multi-teacher contrastive learning (row 6) achieves the best result of 66.70%, confirming the complementary effectiveness of all components in M2S.

Table 3: Ablation study of M2S components. Results report 3D mAP (IoU = 0.3/0.5/0.7) on the V2XSet. Here, ST denotes single-teacher distillation (AttFuse only), MTCD† indicates multi-teacher adaptive distillation, which refers to global-level feature distillation. MLFE is multi-level feature enhancement, APCR refers to auxiliary point cloud reconstruction, and MTCD‡ is multi-teacher contrastive learning, which refers to local-level feature distillation. Best are highlighted in **bold**.

Row	Module					Validation			Testing		
	ST	MTCD†	MLFE	APCR	MTCD‡	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7
1	-	-	-	-	-	78.89	76.72	68.10	75.67	73.72	58.06
2	✓	-	-	-	-	82.83	80.17	68.54	78.10	74.77	58.29
3	-	✓	-	-	-	84.10	81.91	71.64	81.98	78.55	62.76
4	-	✓	✓	-	-	84.31	82.27	72.85	82.15	**79.44**	64.58
5	-	✓	✓	✓	-	84.91	82.75	73.08	81.80	79.32	66.50
6	-	✓	✓	✓	✓	**85.20**	**82.83**	**73.73**	**82.15**	79.26	**66.70**

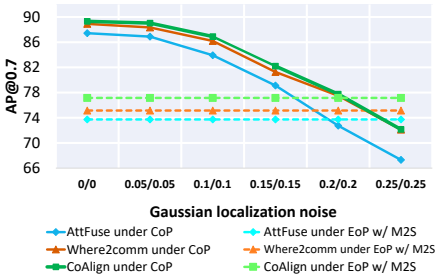


Fig. 6: Comparison performance of noise robustness under Co-Perception (CoP) and Eo-perception (EoP) w/ M2S modes. We report the 3D mAP (IoU=0.7) of AttFuse, Where2comm and CoAlign under diffent gaussian noise on the V2XSet.

Table 4: Comparison of computational complexity, inference speed, and communication latency under CoP and EoP w/ M2S. FLOPs (G) denotes model complexity, FPS (fps) is the inference speed, and Delay (ms) is communication latency.

Methods	Type	FLOPs ↓	FPS ↑	Delay ↓
AttFuse	CoP	143.8	10.3	266.6
	EoP w/M2S	92.8	20.7	0
V2X-ViT	CoP	315.9	6.6	261.6
	EoP w/M2S	205.3	9.2	0
Where2comm	CoP	415.9	12.1	281.9
	EoP w/M2S	185.5	16.0	0
CoAlign	CoP	138.1	15.7	276.7
	EoP w/M2S	92.9	17.7	0
CoSDH	CoP	428.6	12.9	260.5
	EoP w/M2S	189.7	14.3	0
PillarNeSt	CoP	893.8	10.6	274.5
	EoP w/M2S	391.8	13.9	0

4.5 Analysis Experiments

Comparison of noise robustness. Following the noise injection settings from [7] and [38], we applied Gaussian noise ζ from $(0.0\,\mathrm{m}, 0.0°)$ to $(0.25\,\mathrm{m}, 0.25°)$ to simulate localization errors under complex scenarios. As shown in Fig. 6, collaborative perception methods suffer severe performance degradation as errors accumulate. In contrast, when the error is within $(0.25\,\mathrm{m}, 0.25°)$, our M2S maintains consistent performance by relying solely on ego-agent inference, showing that M2S can inherently mitigate the negative impact of localization errors.

Comparison of computational complexity, inference speed, and communication latency. Table 4 compares Co-Perception (CoP) and Ego-Perception (EoP) with our M2S **during inference**. The CoP receives multi-agent input, while the EoP only receives the ego. Overall, M2S notably reduces computational complexity, accelerates inference, and eliminates communication latency.

Table 5: Comparison of different teacher combinations. We compared five teacher combinations and uniformly used [39] as student. Best in **bold** and second in underline.

Teacher Combinations	Performance			Train	GPU
	AP@0.3	AP@0.5	AP@0.7	Time (h)	Mem.(M)
AttFuse (Single teacher)	84.34	82.35	73.26	7.54	7640
AttFuse+CoSDH	83.95	82.20	73.69	8.05	7724
AttFuse+V2X-ViT	83.74	81.45	73.41	12.13	10730
CoSDH+V2X-ViT	84.25	82.74	73.30	12.10	10348
AttFuse+V2X-ViT+CoSDH	**85.20**	**82.83**	**73.73**	12.69	10992

Table 6: Performance comparison of different methods under various range settings. We report the reults of V2Xset with **single-agent input**. Best in **bold**.

Methods	All Range		Range [0,30) (m)		Range[30,50) (m)		Range[50,+inf) (m)	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
AttFuse	76.72	68.10	84.34	81.08	76.31	67.41	58.43	39.63
AttFuse w/M2S	**82.83**	**73.73**	**95.29**	**92.49**	**78.13**	**68.76**	**60.46**	**39.97**
Where2comm	78.23	68.44	84.55	81.70	**78.65**	68.04	61.93	39.32
Where2comm w/M2S	**83.26**	**75.15**	**95.59**	**93.23**	78.10	**69.66**	**62.36**	**42.60**
CoAlign	78.63	69.74	85.08	82.15	78.31	68.95	62.93	42.31
CoAlign w/M2S	**86.27**	**77.15**	**96.10**	**93.76**	**80.88**	**72.20**	**71.06**	**48.54**

Specifically, M2S reduces FLOPs by 35.43% and doubles FPS for AttFuse [39]. Since our M2S does not require external communication, which eliminates transmission delays, whereas collaboration methods suffer from 260 to 282 ms latency. In short, these results confirm M2S's practicality and deployment potential.

Comparison of teacher combinations and discussion of trade-offs. As shown in Table 5, the combination AttFuse+V2X-ViT+CoSDH achieves the highest performance, which comes with substantial training overhead, requiring more training time and higher GPU memory compared to the single-teacher. Notably, other high-cost combinations, like AttFuse+V2X-ViT, fail to deliver comparable accuracy, while the CoSDH+V2X-ViT offers competitive but still inferior results with similar resource consumption. Given its superior accuracy and acceptable training cost, we select AttFuse+V2X-ViT+CoSDH as the final teacher configuration for subsequent experiments.

Comparison of performance across distance ranges. Table 6 shows that integrating M2S consistently improves all methods across all distance ranges. CoAlign achieves the largest overall gain, with AP@0.7 rising from 69.74 to 77.15. In the short range [0,30), M2S boosts AP@0.7 by over 10 points for AttFuse, CoAlign, and Where2comm, indicating enhanced local feature modeling. Gains persist in the mid-range [30,50) and extend to the challenging far-range [50,+inf), where CoAlign's AP@0.5 increases from 62.93 to 71.06. These results demonstrate M2S's effectiveness across all distance conditions.

4.6 Qualitative Analysis

To demonstrate M2S's efficacy in transferring dense multi-teacher knowledge to sparse students, we show FOV and BEV results on V2XSet with and without

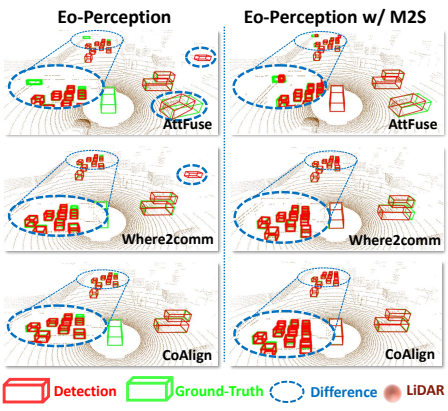


Fig. 7: Visualization of detection results in the Field of view (FOV) on the V2XSet.

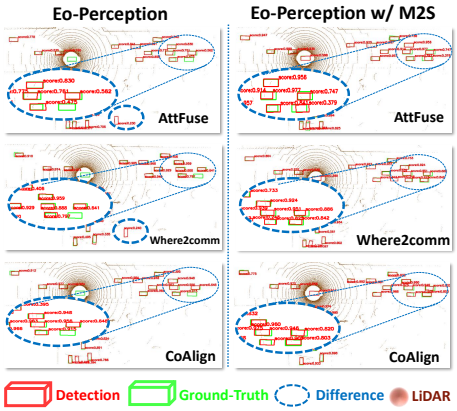


Fig. 8: Visualization of detection results in the Bird's Eye View (BEV) on the V2XSet.

M2S in Figs. 7 and 8. M2S achieves more comprehensive and precise detection. Without M2S (Eo-Perception alone), students frequently miss or misclassify objects, especially distant or in dense areas. M2S significantly mitigates these issues, effectively distilling collaborative knowledge and enhancing student performance.

5 Conclusion

In this paper, we analyze the inherent advantages and limitations of Co-Perception and Eo-Perception, and explore the new **C2E** (**C**o-Perception **to** **E**o-Perception) paradigm. Through the **M**ulti-**to**-**S**ingle (**M2S**) agent contrastive knowledge distillation framework, we achieve state-of-the-art performance on 3D object detection without increasing any communication and inference overhead. Extensive experiments on V2XSet, V2V4Real, and DAIR-V2X datasets demonstrate the effectiveness and generalizability of our M2S framework. In conclusion, the C2E paradigm and M2S framework can provide a new technical path for the reliable deployment of autonomous driving perception systems in real-world scenarios.

Limitation and Future Work. Our current M2S framework is designed for LiDAR-based detection, and extending it to multi-modal sensing (e.g., LiDAR-camera fusion) remains a promising direction. Additionally, the adaptive teacher weighting currently relies on ground-truth labels during training; future work could explore a GT-free multi-teacher voting mechanism to identify and suppress outlier teachers without annotation-based supervision.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.42571514).

References

1. Chae, Y., Kim, H., Yoon, K.J.: Towards robust 3d object detection with lidar and 4d radar fusion in various weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15162–15172 (June 2024)
2. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524. IEEE (2019)
3. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. AAAI **35** (2021)
4. Gao, X., Zhang, X., Lu, Y., Huang, Y., Yang, L., Xiong, Y., Liu, P.: A survey of collaborative perception in intelligent vehicles at intersections. IEEE Transactions on Intelligent Vehicles (2024)
5. Glaser, N., Liu, Y.C., Tian, J., Kira, Z.: Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2406–2413. IEEE (2021)
6. Han, Y., Zhang, H., Li, H., Jin, Y., Lang, C., Li, Y.: Collaborative perception in autonomous driving: Methods, datasets, and challenges. IEEE Intelligent Transportation Systems Magazine **15**(6), 131–151 (2023)
7. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
8. Huang, X., Wang, J., Xia, Q., Chen, S., Yang, B., Li, X., Wang, C., Wen, C.: V2x-r: Cooperative lidar-4d radar fusion with denoising diffusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27390–27400 (June 2025)
9. Huang, X., Wu, H., Li, X., Fan, X., Wen, C., Wang, C.: Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2409–2416 (2024)
10. Huang, X., Xu, Z., Wu, H., Wang, J., Xia, Q., Xia, Y., Li, J., Gao, K., Wen, C., Wang, C.: L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 3806–3814 (2025)
11. Li, J., Wang, Z., Gong, D., Wang, C.: Scnet3d: Rethinking the feature extraction process of pillar-based 3d object detection. IEEE Transactions on Intelligent Transportation Systems (2024)
12. Li, X., Yin, J., Li, W., Xu, C., Yang, R., Shen, J.: Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3208–3215 (2024)
13. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021)
14. Li, Y., Ibanez-Guzman, J.: Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. IEEE Signal Processing Magazine **37**(4), 50–61 (2020)
15. Li, Z., Liang, H., Wang, H., Zhao, M., Wang, J., Zheng, X.: Mkd-cooper: Cooperative 3d object detection for autonomous driving via multi-teacher knowledge distillation. IEEE Transactions on Intelligent Vehicles **9**(1), 1490–1500 (2023)

16. Liu, S., Gao, C., Chen, Y., Peng, X., Kong, X., Wang, K., Xu, R., Jiang, W., Xiang, H., Ma, J., et al.: Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. arXiv preprint arXiv:2308.16714 (2023)
17. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. arXiv preprint arXiv:2211.07214 (2022)
18. Mao, J., Shi, S., Wang, X., Li, H.: 3d object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision **131**(8), 1909–1963 (2023)
19. Mao, W., Wang, T., Zhang, D., Yan, J., Yoshie, O.: Pillarnest: Embracing backbone scaling and pretraining for pillar-based 3d object detection. IEEE Transactions on Intelligent Vehicles pp. 1–10 (2024). https://doi.org/10.1109/TIV.2024.3386576
20. Qiao, D., Zulkernine, F.: Adaptive feature fusion for cooperative perception using lidar point clouds. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1186–1195 (2023)
21. Qin, Y., Wang, C., Kang, Z., Ma, N., Li, Z., Zhang, R.: Supfusion: Supervised lidar-camera fusion for 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22014–22024 (2023)
22. Ren, S., Lei, Z., Wang, Z., Dianati, M., Wang, Y., Chen, S., Zhang, W.: Interruption-aware cooperative perception for v2x communication-aided autonomous driving. IEEE Transactions on Intelligent Vehicles **9**(4), 4698–4714 (2024)
23. Ruan, J., Cui, H., Huang, Y., Li, T., Wu, C., Zhang, K.: A review of occluded objects detection in real complex scenarios for autonomous driving. Green energy and intelligent transportation **2**(3), 100092 (2023)
24. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In: CVPR (2020)
25. Shi, S., Wang, X., Li, H.: PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In: CVPR (2019)
26. Shi, W., Rajkumar, R.: Point-gnn: Graph neural network for 3d object detection in a point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1711–1719 (2020)
27. Song, Z., Liu, L., Jia, F., Luo, Y., Jia, C., Zhang, G., Yang, L., Wang, L.: Robustness-aware 3d object detection in autonomous driving: A review and outlook. IEEE Transactions on Intelligent Transportation Systems **25**(11), 15407–15436 (2024). https://doi.org/10.1109/TITS.2024.3439557
28. Su, S., Li, Y., He, S., Han, S., Feng, C., Ding, C., Miao, F.: Uncertainty quantification of collaborative detection for self-driving. arXiv preprint arXiv:2209.08162 (2022)
29. Wang, T., Hu, X., Liu, Z., Fu, C.W.: Sparse2dense: Learning to densify 3d features for 3d object detection. Advances in Neural Information Processing Systems **35**, 38533–38545 (2022)
30. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
31. Wu, H., Wen, C., Shi, S., Li, X., Wang, C.: Virtual Sparse Convolution for Multimodal 3D Object Detection. In: CVPR (2023)
32. Xia, Q., Deng, J., Wen, C., Wu, H., Shi, S., Li, X., Wang, C.: Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited anno-

tations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6254–6263 (October 2023)
33. Xia, Q., Lin, W., Xiang, H., Huang, X., Chen, S., Dong, Z., Wang, C., Wen, C.: Learning to detect objects from multi-agent lidar scans without manual labels. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1418–1428 (2025)
34. Xia, Q., Ye, W., Wu, H., Zhao, S., Xing, L., Huang, X., Deng, J., Li, X., Wen, C., Wang, C.: Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In: CVPR. pp. 15321–15330 (2024)
35. Xu, J., Zhang, Y., Cai, Z., Huang, D.: Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6834–6843 (2025)
36. Xu, R., Chen, W., Xiang, H., Liu, L., Ma, J.: Model-agnostic multi-agent perception framework. arXiv preprint arXiv:2203.13168 (2022)
37. Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., Meng, Z., Xiang, H., Dong, X., Song, R., 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. pp. 13712–13722 (2023)
38. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
39. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
40. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: European conference on computer vision. pp. 677–695. Springer (2022)
41. Yan, Y., Mao, Y., Li, B.: SECOND: Sparsely Embedded Convolutional Detection. Sensors **18** (2018)
42. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12319–12328 (2022)
43. Yang, D., Yang, K., Wang, Y., Liu, J., Xu, Z., Yin, R., Zhai, P., Zhang, L.: How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. Advances in Neural Information Processing Systems **36**, 25151–25164 (2023)
44. Yang, K., Yang, D., Zhang, J., Wang, H., Sun, P., Song, L.: What2comm: Towards communication-efficient collaborative perception via feature decoupling. In: Proceedings of the 31st ACM international conference on multimedia. pp. 7686–7695 (2023)
45. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11784–11793 (June 2021)
46. Yoo, J., Feng, Z., Pan, T.Y., Sun, Y., Phoo, C.P., Chen, X., Campbell, M., Weinberger, K.Q., Hariharan, B., Chao, W.L.: Learning 3d perception from others' predictions. arXiv preprint arXiv:2410.02646 (2024)

47. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., Nie, Z.: 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. pp. 21361–21370 (2022)
48. Zhang, J., Wang, Y., Qian, L., Sun, P., Li, Z., Jiang, S., Liu, M., Song, L.: Dsrc: Learning density-insensitive and semantic-aware collaborative representation against corruptions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9942–9950 (2025)
49. Zhao, B., Zhang, W., Zou, Z.: Bm2cp: Efficient collaborative perception with lidar-camera modalities. arXiv preprint arXiv:2310.14702 (2023)
50. Zheng, B., Lu, S., Huang, R., Huang, M., Lu, F., Tian, W., Zhuo, G., Xiong, L.: R2ldm: An efficient 4d radar super-resolution framework leveraging diffusion model. arXiv preprint arXiv:2503.17097 (2025)
51. Zheng, W., Hong, M., Jiang, L., Fu, C.W.: Boosting 3D Object Detection by Simulating Multimodality on Point Clouds. In: CVPR (2022)
52. Zheng, W., Jiang, L., Lu, F., Ye, Y., Fu, C.W.: Boosting Single-Frame 3D Object Detection by Simulating Multi-Frame Point Clouds. In: ACMMM (2022)