CoGoal3D: Collaborative 3D Object Detection with 3D-Aware Fusion and Refinement

Zhihao Yang[1], Zhiyu Xiang[1,2*], Peng Xu[1], Tianyu Pu[1], Kai Wang[1], Eryun Liu[1], Dongping Zhang[3], and Yong Ding[1]

[1] Zhejiang University, Hangzhou, China
{yangzhihao, xiangzy, xxxupeng, 3190105835, kai-wang, eryunliu, dingyong09}@zju.edu.cn
[2] Zhejiang Provincial Key Laboratory of Multi-Modal Communication Networks and Intelligent Information Processing, Hangzhou, China
[3] China Jiliang University, Hangzhou, China
06a0303103@cjlu.edu.cn

Abstract. V2X collaborative object detection features overcoming the limitations of single-vehicle systems by aggregating environmental features from multiple collaborative agents. However, existing mainstream V2X perception methods mainly focus on 2D BEV object detection. When 3D detection task is concerned, inferior results are obtained because they ignore the 3D spatial misalignment caused by differing height and attitude among the collaborators. In this paper, we propose a novel collaborative 3D object detection framework called CoGoal3D, which extracts and refines the 3D feature gradually in a two-stage pipeline. In the first stage, a multiscale 3D-aware global fusion module is designed to mitigate the 3D spatial misalignment. The resulting proposals are then refined in the second stage with an auxiliary task of 3D point reconstruction. An effective multi-agent collaborative data augmentation strategy is further proposed to enrich the training data while minimizing information loss. Extensive experiments on public real-world datasets demonstrate that our CoGoal3D achieves new state-of-the-art performance, with 3D AP@0.7 improvements of 10.86%, 10.34%, and 10.18% on the DAIR-V2X, V2V4Real, and V2X-Real datasets, respectively.

Keywords: Autonomous driving · Collaborative perception · 3D object detection

1 Introduction

Environmental perception is a fundamental task in autonomous driving. However, single-vehicle perception is inherently limited by the restricted sensing range and occlusions, leading to inferior performance at complex scenarios such as road crossings. In recent years, V2X collaborative perception has emerged as a highly attractive solution to address these limitations by leveraging multi-view information through V2X.

* Corresponding author.

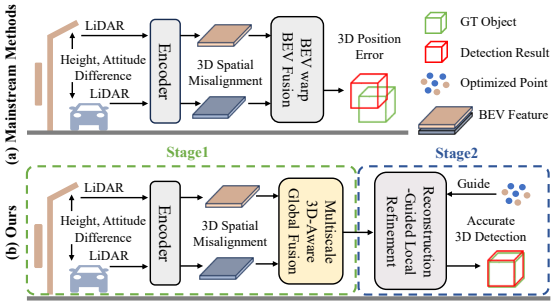

Fig. 1: Difference between the mainstream methods and ours. (a) Mainstream broadcast-based methods perform feature fusion only on 2D BEV space. (b) Ours is a two-stage pipeline, enhancing the 3D alignment by 3D-aware global fusion and reconstruction-guided local refinement respectively.

Currently most of the V2X collaborative perception methods focus on 2D BEV object detection. They usually adopt BEV-based intermediate feature fusion scheme upon a broadcast communication paradigm [2, 7, 9, 16, 20], as shown in Figure 1(a). In this paradigm, each collaborative agent independently extracts its own BEV feature locally and transmits it to the ego-vehicle together with its pose. The ego-vehicle then aligns the collaborative BEV features to its own coordinate system via a 2D warping operation for subsequent feature fusion. This paradigm is efficient in that it requires only one round of communication and the features extracted for self-perception can be directly broadcasted to others without any extra computational cost. However, the alignment with only 2D BEV-based warping assumes all agents observe the scene in the same horizontal plane, regardless of their discrepancies in height and attitude (e.g., pitch angle) caused by different sensor mounting position as well as uneven ground surface. The problem becomes even more pronounced when 3D instead of 2D BEV object detection task is considered.

On the other hand, data augmentation is crucial for deep learning based detection methods. However, currently very few data augmentation methods are specially designed for collaborative perception tasks. The popular method, Dual Point Transformation Projection [19], projects the point cloud of all collaborative agents into a unified coordinate system (e.g., ego-vehicle's coordinate system) for unified global augmentations, and then reprojects the augmented point cloud back to each agent's respective coordinates before using it for training. Although simple, these ego-centric augmentations risk shifting collaborators' point cloud out of their detection range, thereby causing significant information loss. Better data augmentation method tailored for collaborative perception is highly desired.

In this paper, we propose a novel collaborative 3D object detection framework called CoGoal3D, to address the aforementioned problems. The general pipeline of the network is shown in Figure 1(b). In contrast to the mainstream methods, we carefully consider the 3D spatial misalignment problem during BEV fusion, and rely on a two-stage pipeline to gradually align and fuse the collaborative

features. Instead of simple 2D warping of BEV features, we propose a multiscale 3D-Aware Global Fusion (3D-AGF) module in stage 1 to embed the 3D position and aggregate the spatial features more robustly. With the object proposals at hand, we further design an auxiliary 3D point reconstruction task in stage 2 to help optimize the 3D bounding box. We also invent a Multi-Agent Collaborative Data Augmentation (MCDA) strategy which is specifically tailored for the collaborative perception. By combining a special sequence of local and global data transformation, MCDA features little information loss and is highly effective in augmenting the collaborative data. We conduct extensive experiments on widely used real-world collaborative perception datasets: DAIR-V2X [24], V2V4Real [21] and V2X-Real [17]. Experimental results show that our method achieves much higher performance than the SOTA methods.

In summary, the main contributions are summarized as follows:

- We propose CoGoal3D, a novel collaborative 3D object detection framework that well addresses the 3D spatial misalignment problem underestimated in existing mainstream methods.
- We design a multiscale 3D-aware global fusion module and an auxiliary 3D point reconstruction task respectively in the two-stage processing pipeline, which gradually refines the feature for the detection.
- We propose the multi-agent collaborative data augmentation, an effective data augmentation strategy tailored for the collaborative perception task.
- We conduct comprehensive experiments on public real-world datasets. Experimental results demonstrate that our CoGoal3D outperforms previous SOTA methods on both BEV and 3D AP metrics by a large margin.

2 Related Work

2.1 Single-Vehicle 3D Object Detection

Single-vehicle 3D object detection can be categorized into one-stage and two-stage methods based on whether the region proposals are utilized for refinement. In one-stage methods, PointNet [11] and PointNet++ [12] directly encode irregular raw point cloud to extract point-level features and predict 3D bounding boxes. VoxelNet [27] divides the raw point cloud into regular voxels and uses 3D convolution to encode voxel features. SECOND [23] introduces sparse 3D convolution to accelerate voxel feature encoding. PointPillars [5] partitions the point cloud into regular pillars on the X-Y plane, enabling the use of 2D convolution to encode BEV features to reduce computational consumption.

In two-stage methods, Point-RCNN [15] uses PointNet++ [12] as the backbone to generate proposals and introduces point cloud RoI pooling to extract proposal features. PV-RCNN [14] utilizes both point-based and voxel-based representations of point cloud, extracting proposal features through RoI grid pooling. Voxel-RCNN [1] uses voxel-based representations to balance detection accuracy and efficiency. Pillar-RCNN [13] represents point cloud with pillars and extracts proposal features via 2D RoI pooling on the BEV plane.

2.2 Collaborative 3D Object Detection

According to the communication paradigm, current collaborative object detection methods can be classified into handshake-based and broadcast-based ones. Handshake-based paradigm requires the ego-vehicle to send its pose at first, allowing other collaborators to project their point cloud to ego-vehicle's coordinate system before extracting the feature and returning the messages. In this category, V2X-ViT [22] uses a transformer architecture for feature fusion. DI-V2X [6] introduces domain-mixing instance augmentation and follows DiscoNet [7]'s knowledge distillation framework. DSRC [25] utilizes intermediate fusion with augmented point clouds as teacher network for knowledge distillation, enhancing the robustness of the student network. ERMVP [26] improves communication efficiency via feature sampling and handles localization errors with a spatial calibration module. While relatively accurate, handshake-based paradigm suffers from a two-round communication and multiple independent feature extraction for each collaborating agent, resulting in significant communication and computational overhead.

To address these concerns, the broadcast-based paradigm has emerged as the dominant research direction. In this paradigm, the ego-vehicle receives BEV features and poses from other agents and warps the received features to its own coordinate system for subsequent feature fusion. In this line, DiscoNet [7] employs early fusion as teacher network for teacher-student knowledge distillation. CoBEVT [20] introduces fused axial attention to capture both local and global relationships, effectively aggregating features across agents. CoAlign [9] proposes a pose-graph optimization method to improve the robustness of collaborative object detection against pose noise. CoSDH [18] presents a hybrid intermediate-late fusion paradigm that leverages confidence-aware late fusion to improve robustness against low communication bandwidth. However, these existing works mainly focus on 2D BEV detection and ignore the differences in height and attitude among collaborators. When the 3D detection task is considered, they obtain inferior performance since critical 3D spatial information can hardly be compensated by simple 2D BEV warping. In contrast, our method fully accounts for the spatial alignment in 3D, by embedding the 3D pose information and supervising the 3D reconstructed points in the pipeline.

3 Method

3.1 Overview

The overall architecture of the proposed CoGoal3D is illustrated in Figure 2. It consists of two stages, with multiscale 3D-Aware Global Fusion (3D-AGF) as stage 1 to produce object proposals and Reconstruction-Guided Local Refinement (RGLR) as stage 2 to generate final results. During training, Multi-Agent Collaborative Data Augmentation (MCDA) is applied to the point clouds and poses of all collaborative agents.

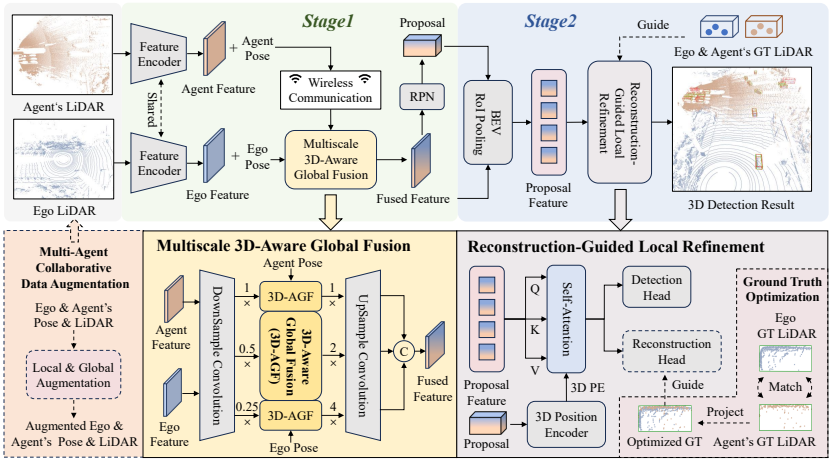

Fig. 2: The overall architecture of the proposed CoGoal3D. The dashed boxes and lines indicate the components used exclusively during training. Further details of these components are illustrated in Section 3.

In the stage 1, the input LiDAR points are passed through shared 3D backbones, from which each agent extracts its individual Bird's-Eye-View (BEV) features. The collaborative agents then broadcast their BEV features and poses to the ego-vehicle. Upon receiving these messages, ego-vehicle utilizes the multiscale 3D-Aware Global Fusion (3D-AGF) module to generate the fused feature. Subsequently, a Region Proposal Network (RPN) is used to obtain a set of initial object proposals for the second stage.

In the second stage, BEV RoI pooling is applied to the fused feature to extract the proposal feature. Then, the obtained feature is fed into the Reconstruction-Guided Local Refinement (RGLR) module for final prediction. It contains two parallel heads: a detection head that produces the refined proposal, and another auxiliary reconstruction head that predicts the 3D point cloud within the proposal. The reconstruction process is supervised by an optimized ground truth point cloud generated by our Ground Truth Optimization (GTO) method, guiding the network to learn fine-grained 3D geometric details of the object, yielding more accurate 3D object results.

Note that MCDA and auxiliary point cloud reconstruction modules are only applied during training.

3.2 Multiscale 3D-Aware Global Fusion

Existing broadcast based methods simply warp the received BEV features on the BEV plane for subsequent feature fusion, which neglects the 3D spatial misalignment caused by different height and attitude among collaborators. To address this, we propose a multiscale 3D-Aware Global Fusion (3D-AGF) module to achieve 3D spatial alignment. We first encode the BEV feature into multiscale

features $F_{k,l}, l \in \{1, \cdots, L\}$, where $F_{k,l}$ denotes k-th agent's BEV feature at the l-th scale, and then performs 3D-AGF at each scale. Figure 3 illustrates the process of 3D-AGF, which consists of the following two steps.

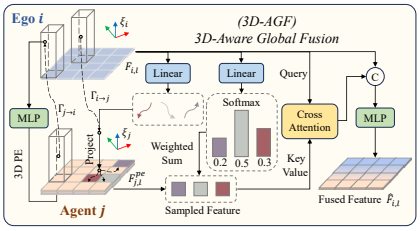

Fig. 3: The architecture of the proposed 3D-Aware Global Fusion (3D-AGF) module. It explicitly incorporates 3D spatial information to align collaborative features via 3D position encoding and 3D-aware deformable cross attention.

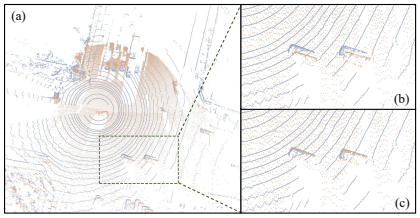

Fig. 4: Illustration of Ground Truth Optimization (GTO). (a) is nominally aligned raw point clouds derived from DAIR-V2X dataset. (b) presents an enlarged visualization of the green region in (a). (c) shows the optimized result of the green region in (a) with our GTO.

3D Position Encoding. To incorporate the crucial 3D spatial information, we introduce 3D position encoding to the BEV features. The relative spatial transformation matrix $\Gamma_{j\to i}$ is first computed based on the poses of ego-vehicle ξ_i and collaborative agent ξ_j. Then, we take the corresponding agent's pillar centers in the l-th scale as its 3D coordinates $p_{j,l} = (x_{j,l}, y_{j,l}, z_{j,l})$, and transform it to the ego-vehicle's coordinate system with $\Gamma_{j\to i}$ as:

$$p_{j\to i,l} = \Gamma_{j\to i} * p_{j,l}. \tag{1}$$

The obtained $p_{j\to i,l}$ is passed through an MLP layer to obtain the 3D position encoding, which is then added to j-th agent's BEV feature as:

$$F^{pe}_{j,l} = F_{j,l} + \mathrm{MLP}(p_{j\to i,l}). \tag{2}$$

3D-Aware Deformable Cross Attention. Unlike previous methods that first warp collaborative agents' features to the ego-vehicle's coordinate system and then perform feature fusion, we jointly handle 3D global spatial alignment and feature fusion using deformable attention [28]. Concretely, we take each grid feature of the ego BEV feature $F_{i,l}$ as a query $q_{i,l}$ and transform its pillar center to j-th agent's coordinate system with $\Gamma_{i\to j}$. The resulting 3D point is then projected to j-th agent's BEV plane to obtain the reference point $r_{q,l}$, which compensates for spatial offsets caused by collaborator attitude differences.

With this 3D-aware reference point, deformable attention then learns sampling offsets around $r_{i,l}$ to sample features from $F^{pe}_{j,l}$, yielding the corresponding

aligned collaborator's feature $F_{j\to i,l}(q)$ as:

$$\begin{aligned} F_{j\to i,l}(q) = \text{Deformable-Attention}(q_{i,l}, r_{q,l}, F^{pe}_{j,l}) = \\ \sum_{m=1}^{M} W_m[\sum_{k=1}^{K} A_{mkq} \cdot (W'_m F^{pe}_{j,l}(r_{q,l} + \Delta r_{mkq,l}))], \end{aligned} \tag{3}$$

where M is the number of attention heads and K is the number of the sampling points. A_{mkq} and $\Delta r_{mkq,l}$ denote the attention weight and sampling offset. W_m and W'_m are the learnable matrices, respectively.

Next, we apply cross attention between the aligned feature $F_{j\to i,l}$ and the ego feature $F_{i,l}$. The result is then concatenated with ego feature $F_{i,l}$ and fused by an MLP as:

$$\hat{F}_{i,l} = \text{MLP}[F_{i,l}, \text{Cross-Attention}(F_{i,l}, F_{j\to i,l})]. \tag{4}$$

Here, $\hat{F}_{i,l}$ denotes the fused feature for the ego at the l-th scale, which will be further aggregated into the multiscale fused feature $\hat{F}_i$ by upsample convolution and concatenation, as shown in Figure 2. Finally, a decoder is used to decode the fused feature $\hat{F}_i$ into a set of proposals B.

3.3 Reconstruction-Guided Local Refinement

Building upon the global feature fusion provided by multiscale 3D-AGF, we extend our model to a two-stage fusion framework for fine-grained 3D refinement. As shown in Figure 2, besides a detection decoder for 3D detection task, in this stage we further introduce an auxiliary 3D point reconstruction task to enhance the model's comprehension of 3D spatial information. It consists of the following two key components.

RoI-level 3D Point Reconstruction. Given the fused feature $\hat{F}_i$ and the proposal $b_m \in B$, we perform BEV RoI pooling [13] to obtain the proposal features $\{F^g_{b_m}\}_{g=1,\cdots,G^2}$, where G^2 denotes $G \times G$ regular pillars on the BEV plane that the proposal is divided into. Then, 3D position encoding is generated for the proposal features. Specifically, for the g-th pillar of the proposal, we denote p_g as the pillar center. The relative coordinates of the pillar center with respect to the proposal's center and vertices are encoded by an MLP as:

$$p^{pos}_g = \text{MLP}([p_g - r_c; p_g - r_1; \cdots; p_g - r_8]), \tag{5}$$

where r_c and r_i represent the center and the i-th vertex of the proposal. The proposal features are then added with their corresponding position encoding and passed through a self-attention layer, which enables sufficient interaction between the proposal features. The resulting enhanced proposal feature F_{b_m} is represented as:

$$F_{b_m} = \text{Concat}(\text{SA}(\{F^g_{b_m} + p^{pos}_g\}_{g=1,\cdots,G^2})), \tag{6}$$

where SA and Concat denote the self-attention and concatenating operation. Finally, a decoder is applied to reconstruct a fixed number N_p of points for each pillar of the proposal.

Ground Truth Optimization. Directly using the nominally aligned raw point clouds from collaborators as the ground truth of the reconstruction suffers from slight spatial inconsistencies caused by some artifacts (e.g., calibration errors, temporal asynchrony or LiDAR scanning effects), as shown in Figure 4(b). We propose a Ground Truth Optimization (GTO) method to provide better supervision signal for the stage 2.

Specifically, we match the ground truth bboxes of each collaborator with those of the ego-vehicle. For datasets without ID annotations, we match the gt bboxes based on their IoU using the Hungarian algorithm [4]. Otherwise, we match the gt bboxes directly based on the annotated IDs. For each matched GT pair, we compute the relative transformation between them. This computed relative transformation is subsequently applied to project the collaborator's points within the GT bbox to the matched bbox of the ego-vehicle's. As shown in Figure 4(c), our GTO yields a better aligned ground truth for supervision, which in turn guides the network to learn fine-grained 3D geometric details, leading to more accurate 3D object detection result.

3.4 Multi-Agent Collaborative Data Augmentation

Previous data augmentation methods for collaborative perception, such as Dual Point Transformation Projection [19], are similar to those for single vehicle perception except for introducing an extra coordinate transformation between the ego and the collaborator. They perform global transformation-based augmentations (e.g., flipping, rotation and scaling) in a unified ego-vehicle's coordinate system. However, such ego-centric global transformations are likely to shift the collaborators' point clouds out of their original detection range, leading to significant information loss, as shown in Figure 6(a). To address this issue, we propose Multi-Agent Collaborative Data Augmentation (MCDA), which combines a special sequence of local and global augmentations to maximize data diversity while minimizing the information loss, as shown in Figure 5.

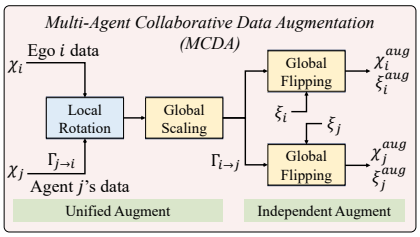

Fig. 5: Illustration of the proposed Multi-Agent Collaborative Data Augmentation (MCDA) strategy. It applies a specific sequence of local and global augmentations, maximizing data diversity while minimizing information loss.

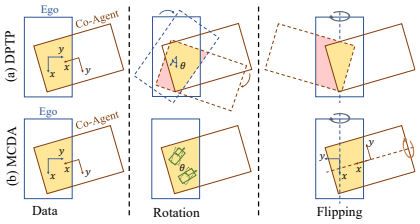

Fig. 6: Comparison between DPTP and MCDA. The detection range for the agents are marked with the solid box. The collaborative overlap region and resulting information loss are highlighted with yellow and pink.

Specifically, MCDA first projects the j-th agent's points χ_j into the ego-vehicle's coordinate system with relative transformation matrix $\Gamma_{j\to i}$, followed by a unified local rotation around the ground truth objects instead of global rotation. The local rotation augments the object's orientation without altering its global position, thus avoiding shifting the GT objects out of the detection area. Next, a unified global scaling is applied to expand entire training data with different scales. Then, the augmented points of the collaborator are reprojected back to its own coordinate system. Finally, global flipping is performed for each agent, where both the point clouds and the poses are flipped to ensure the spatial consistency. Since the detection range is symmetric, this flipping does not lead to any information loss. As shown in Figure 6 (b), our MCDA achieves diversified data augmentation with negligible information loss compared to DPTP, which is valuable for the training of collaborative perception tasks.

3.5 Training Loss

CoGoal3D adopts an end-to-end training strategy, with the training loss consisting of three components: RPN loss L_{RPN}, proposal refinement loss L_{refine}, and 3D point reconstruction loss L_{rec}, as follows:

$$L_{\text{total}} = L_{\text{RPN}} + L_{\text{refine}} + L_{\text{rec}}, \tag{7}$$

L_{RPN} and L_{refine} are composed of classification loss and regression loss, corresponding to the first-stage and second-stage detection loss, respectively. To address the class imbalance in the first-stage, the classification loss of L_{RPN} uses Focal Loss [8], while L_{refine} employs binary cross-entropy loss. Smooth L1 loss is applied as the regression loss of both L_{RPN} and L_{refine}. The 3D point reconstruction loss L_{rec} is computed by calculating the Chamfer Distance between the reconstructed point cloud and the optimized mixed ground truth point cloud.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluated our method on widely used real-world collaborative perception datasets: DAIR-V2X [24], V2V4Real [21] and V2X-Real [17]. DAIR-V2X is the first real-world V2I collaborative dataset, featuring one vehicle and one collaborative infrastructure. The vehicle is equipped with a 40-line LiDAR, while the infrastructure is equipped with a 300-line LiDAR. The dataset consists of 9k frames of collaborative point clouds, split into training, validation, and test sets with a ratio of 5:2:3. We perform experiments using the annotations completed by CoAlign [9], with the detection range set to $x \in [-100.8m, 100.8m], y \in [-40m, 40m], z \in [-3.5m, 1.5m]$. V2V4Real is the first real-world V2V collaborative dataset, containing two collaborative vehicles, each equipped with a 32-line LiDAR. The dataset includes 20k frames of collaborative point clouds, split into training, validation, and test sets with the proportions 14,210/2,000/3,986.

The detection range is set to $x \in [-140.8m, 140.8m]$, $y \in [-38.4m, 38.4m]$, $z \in [-5m, 3m]$. V2X-Real is a large-scale real-world V2X collaborative dataset, featuring a mixture of multiple vehicles and infrastructures equipped with 128 line LiDARs. The dataset contains 33k LiDAR frames, split into training, validation, and test sets with a ratio of 23379/2270/6850. It supports evaluation from different perspectives, including Vehicle-Centric (VC) and Infrastructure-Centric (IC). For our experiments, we focus on the Vehicle-Centric (VC) setting and evaluate the performance on the car class, with the detection range set to $x \in [-102.4m, 102.4m], y \in [-38.4m, 38.4m], z \in [-5m, 3m]$.

Evaluation Metrics. We use both BEV and 3D Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 to evaluate the performance of the collaborative 3D object detection.

4.2 Implementation Details

Our model is implemented in PyTorch [10] and trained on NVIDIA RTX 3090 GPU. We use the Adam optimizer [3] with an initial learning rate of 0.001, which is decayed by a factor of 0.1 at epochs 10, 20, and 40. Our model is trained for a maximum of 60 epochs with a batch size of 6. Early stopping is employed to select the best epoch.

For the network architecture, we use PointPillars [5] as the 3D backbone and extract BEV features with a pillar size of 0.4m $\times$ 0.4m. The multiscale 3D-AGF module employs a 3-layer multiscale deformable attention with 8 attention heads and 9 sampling points. In the second stage, both the BEV RoI pooling size G and the number of reconstructed points for each pillar of the proposal N_p are set to 6.

During training, our model is augmented by the proposed Multi-Agent Collaborative Data Augmentation (MCDA). Specifically, this includes a global random flipping along the x-axis with a 50% probability, a local random rotation with an angle sampled uniformly from $[-\pi/20, +\pi/20]$, and a global random scaling with a factor sampled uniformly from $[0.95, 1.05]$. The models used for comparison are augmented with Dual Point Transformation Projection (DPTP) [19], which consists of a global random flipping along the x-axis with a 50% probability, a global random rotation with an angle sampled uniformly from $[-\pi/4, +\pi/4]$, and a global random scaling with a factor sampled uniformly from $[0.95, 1.05]$.

4.3 Quantitative Evaluation

Comparison of Detection Performance. We compare our method with the existing state-of-the-art (SOTA) collaborative object detection methods.

The results on the DAIR-V2X dataset are shown in Table 1. Consistent with existing methods, we report the algorithm's evaluation results on the validation set. Experimental results show that our method ranks the first and outperforms the existing best DI-V2X [6] by a large margin, i.e., 11.31% and 6.77% improvements on 3D and BEV AP@0.7, respectively. It is worth noting that DI-V2X is

Table 1: Performance comparison with state-of-the-art methods on DAIR-V2X *val* set. The best results are presented in **bold**, while the second-best results are underlined. B denotes broadcast communication and H denotes handshake communication.

Method	Publication	Comm	BEV AP@0.5/0.7	3D AP@0.5/0.7	FPS
No Fusion [5]	CVPR 2019	-	65.56/53.89	59.65/29.44	**32.4**
DiscoNet [7]	NeurIPS 2021	B	73.52/58.15	64.01/32.34	29.7
V2X-ViT [22]	ECCV 2022	H	76.23/58.76	68.68/33.17	16.1
CoBEVT [20]	CoRL 2022	B	72.77/57.91	64.90/35.55	13.6
CoAlign [9]	ICRA 2023	B	78.14/64.81	68.80/39.69	29.1
DI-V2X [6]	AAAI 2024	H	79.39/65.39	72.54/39.24	22.3
ERMVP [26]	CVPR 2024	H	75.37/61.49	68.61/37.51	12.9
DSRC [25]	AAAI 2025	H	74.96/60.23	67.95/36.08	26.1
CoSDH [18]	CVPR 2025	B	78.38/64.84	67.95/36.78	6.7
CoGoal3D(Stage1)	-	B	79.49/67.33	73.24/42.66	24.8
CoGoal3D(Ours)	-	B	**81.75/72.16**	**76.59/50.55**	16.8

Table 2: Performance comparison with state-of-the-art methods on V2V4Real and V2X-Real *test* sets.

Method	BEV AP@0.5/0.7		3D AP@0.5/0.7	
	V2V4Real	V2X-Real	V2V4Real	V2X-Real
No Fusion [5]	55.16/40.65	64.58/54.02	49.32/21.38	61.54/32.40
DiscoNet [7]	75.74/44.96	76.60/62.66	42.66/14.31	69.67/36.95
V2X-ViT [22]	71.98/48.32	79.13/63.23	59.61/19.60	75.27/40.66
CoBEVT [20]	71.00/40.46	79.31/65.16	46.55/13.22	75.90/43.21
CoAlign [9]	76.32/50.90	79.64/66.44	52.73/17.83	75.25/44.30
ERMVP [26]	70.94/41.58	79.94/65.79	48.38/12.55	76.76/44.46
DSRC [25]	75.53/53.07	78.64/64.64	61.20/21.16	75.30/42.67
CoSDH [18]	78.98/51.25	84.35/70.02	50.87/16.67	63.47/25.19
CoGoal3D(Ours)	**82.68/59.72**	**86.85/76.98**	**71.48/31.50**	**81.72/54.64**

a handshake-based method, which removes the 3D feature misalignment by performing prior 3D coordinate transformation at the collaborator side. However, it cannot well tackle the residual spatial inconsistencies contained in real-world data, resulting in inferior detection results. In contrast, being a broadcast-based method, our CoGoal3D effectively handles the spatial feature misalignment by the two-stage gradual refinement paradigm, and achieve much better performance in both BEV and 3D AP metrics. This is thanks to the multiscale 3D-aware feature fusion module and the auxiliary 3D point reconstruction task, as well as the effective collaborative data augmentation strategy. Meanwhile, we see that our method with stage1-only also performs better than the existing methods, demonstrating the effectiveness of the multiscale 3D-aware global fusion module.

In terms of efficiency, our first-stage model is remarkably fast at 24.8 FPS as a broadcast-based method. The full model CoGoal3D, while slower, achieves the highest AP and still runs at a real-time level of 16.8 FPS.

The experimental results on V2V4Real and V2X-Real datasets are presented in Table 2, where similar phenomena can be observed. Our method achieves great improvements on these datasets, outperforming DSRC [25] in BEV and

3D AP@0.7 by 6.65% and 10.34% on V2V4Real, and by 12.34% and 11.97% on V2X-Real, respectively, demonstrating its effectiveness under complex and multi-agent scenarios.

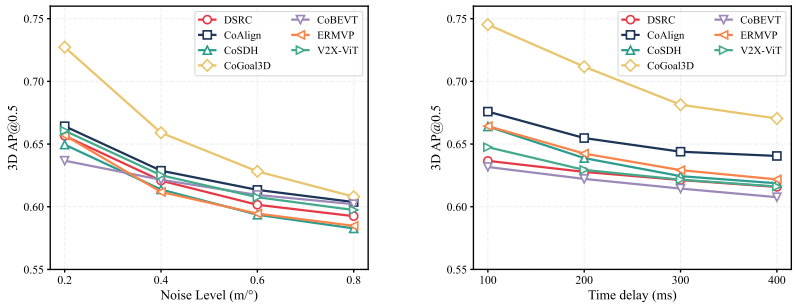

Fig. 7: Robustness evaluation against localization error (left) and transmission latency (right) on the DAIR-V2X *val* set.

Robust Evaluation. We further evaluate the robustness of CoGoal3D against localization error and transmission latency. As shown in Figure 7, we assess the robustness to localization errors by injecting Gaussian noise with varying standard deviations, and simulate transmission latency by introducing different time delays. Thanks to the multi-scale deformable attention that dynamically compensates for spatial misalignment and the optimized point-guided refinement that recovers fine-grained geometric details, CoGoal3D consistently outperforms other methods under these noisy conditions, demonstrating its robustness for practical V2X scenarios.

4.4 Ablation Study

We conduct our ablation study on the DAIR-V2X *val* set.

Table 3: Ablation studies of core components on DAIR-V2X *val* set. MCDA: Multi-Agent Collaborative Data Augmentation; 3D-AGF: multiscale 3D-Aware Global Fusion; RCNN: Second-Stage Refinement; RGLR: Reconstruction-Guided Local Refinement.

MCDA	3D-AGF	RCNN	RGLR	3D AP@0.5/0.7
				65.98/33.27
✓				71.04/40.85
✓	✓			73.24/42.66
✓	✓	✓		75.10/48.28
✓	✓	✓	✓	**76.59/50.55**

Table 4: Ablation studies of 3D PE and GTO on DAIR-V2X *val* set. 3D PE: 3D Position Encoding in 3D-AGF; GTO: Ground Truth Optimization in RGLR.

3D PE	GTO	3D AP@0.5/0.7
		76.59/50.55
×		73.12/46.70
	×	75.87/49.57
×	×	73.29/46.99

Ablation of Core Components. Table 3 presents the ablation results of our CoGoal3D. Firstly, introducing the MCDA strategy for network training can greatly improve the baseline by more than 5.06% on 3D AP@0.5, revealing the superiority of this new data augmentation method. Adding the 3D-AGF module (row 3 vs. row 2) boosts the 3D AP@0.5 by 2.20%, validating its capability in handling 3D spatial misalignment of the real data. On the other hand, adding RGLR (row 5 vs. row 4) in the second stage improves 3D AP@0.7 by 2.27%, confirming the benefit of this module. Finally, the full CoGoal3D model outperforms the baseline by 10.61%/17.28% on 3D AP@0.5/0.7, demonstrating the effectiveness of our design. More detailed ablation for the core components within the modules will be shown in the following.

Ablation of 3D PE and GTO. Table 4 ablates the role of 3D Position Encoding (3D-PE) in 3D-AGF module and Ground Truth Optimization (GTO) in the RGLR module. The results show that removing either 3D-PE or GTO from CoGoal3D leads to performance drops (3.85% and 0.98% in 3D AP@0.7, respectively), confirming their individual contributions in 3D spatial alignment and high quality 3D point supervision. Meanwhile, removing the 3D-PE causes larger performance degradation than the GTO, indicating the importance of the 3D-PE. Interestingly, further remove GTO on the basis of eliminating 3D-PE (row 4) leads to performance improvement (comparing with row 2), which hints that the functionality of 3D reconstruction guidance is dependent on the high-quality 3D-aware fused feature.

Table 5: Ablation and comparison between DPTP and MCDA on DAIR-V2X *val* set. F: Flipping. R: Rotation. S: Scaling.

Method			3D AP@0.5/0.7	
F	R	S	DPTP [19]	MCDA
			70.52/33.99	70.52/33.99
✓			71.26/37.97	72.57/40.14
✓	✓		69.76/37.32	**73.29**/41.50
✓	✓	✓	70.02/38.48	73.24/**42.66**

Table 6: Generalization performance comparison between DPTP and MCDA on the DAIR-V2X *val* set.

Method	3D AP@0.5/0.7	
	DPTP [19]	MCDA
DI-V2X [6]	72.54/39.24	73.54/42.54
ERMVP [26]	68.61/37.51	70.82/38.86
DSRC [25]	67.95/36.08	71.63/40.66
CoSDH [18]	67.95/36.78	71.28/39.98
CoGoal3D	**74.26/48.52**	**76.59/50.55**

Ablation of Data Augmentation. Table 5 presents the ablation and comparison of our data augmentation strategy (MCDA) against DPTP, using our first-stage model as the baseline. The results show that our MCDA yields consistent performance gains as more transformations for data augmentation are applied, boosting 3D AP@0.7 by 8.67% in total. In contrast, much less gains are obtained for the DPTP, with final 3D AP@0.5 even becomes 0.5% lower than the baseline. Comparing the two methods, the key turning point lies in the introduction of different rotation. With the local rotation instead of the global unified one, as well as different flipping operation, our MCDA is able to increase data diversity without introducing information loss.

Table 6 further illustrates the generalization performance of our MCDA by applying it to existing SOTA methods and comparing against DPTP. The results show that MCDA consistently improves the performance on all methods over DPTP, exhibiting its generality and superiority for the collaborative perception task. As expected, our method CoGoal3D maintains the best performance among all compared approaches under the same augmentation strategy.

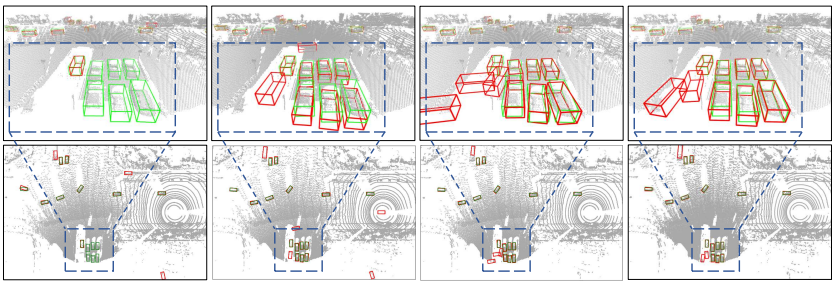

Fig. 8: Qualitative results on DAIR-V2X *val* set. The first row shows the 3D zoom-in views of the blue windows in the second row of BEV views. Green and red bounding boxes denote the 3D object ground truths and the detection results, respectively.

4.5 Qualitative Evaluation

A representative qualitative comparison between our CoGoal3D and other methods on the DAIR-V2X dataset is shown in Figure 8. CoGoal3D obviously achieves better results than others, in that higher consistency of the detected 3D bounding box with the ground truth and fewer false positives. The broadcast-based CoSDH [18] exhibits more position errors in the bounding box due to its weakness in correcting 3D spatial misalignment. DSRC [25] has better alignment, but more false positives are predicted due to the inconsistency contained in the real data. This comparison further demonstrates the superiority of our proposed method.

5 Conclusion

In this paper, we propose CoGoal3D, a novel collaborative 3D object detection framework, to address the critical issues of 3D spatial misalignment and collaborative data augmentation that are underestimated in existing methods. CoGoal3D performs multiscale 3D-Aware Global Fusion and Reconstruction-Guided Local Refinement in a two-stage manner, gradually refining the 3D feature for the collaborative detection task. A special Multi-Agent Collaborative Data Augmentation strategy tailored for the collaborative task, which features

diversifying the training data with little information loss, is also presented. Experimental results on real-world datasets demonstrate that our method achieves superior performance than the existing approaches, validating its effectiveness and great potential in practical application.

Acknowledgements

This work was supported by the Key Project of Natural Science Foundation of Zhejiang Province under Grant LZ26F010003, the Key Research & Development Plan of Zhejiang Province under Grant No.2024C01010, 2024C01017, the Joint R&D Program of the Yangtze River Delta Community of Sci-Tech Innovation with grant number 2024CSJGG01000, and National Key Laboratory of Collective Intelligence & Collaboration (Open Fund Project No. QXZ25017101).

References

1. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1201–1209 (2021)
2. 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)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
5. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
6. 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)
7. 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)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
9. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023)
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

11. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
12. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
13. Shi, G., Li, R., Ma, C.: Pillar r-cnn for point cloud 3d object detection. arXiv preprint arXiv:2302.13301 (2023)
14. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
15. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
16. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: 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. pp. 605–621. Springer (2020)
17. Xiang, H., Zheng, Z., Xia, X., Xu, R., Gao, L., Zhou, Z., Han, X., Ji, X., Li, M., Meng, Z., et al.: V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In: European Conference on Computer Vision. pp. 455–470. Springer (2024)
18. 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)
19. Xu, R., Chen, C.J., Tu, Z., Yang, M.H.: V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. IEEE transactions on pattern analysis and machine intelligence (2024)
20. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022)
21. 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)
22. 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)
23. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
24. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., 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. pp. 21361–21370 (2022)
25. 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)

26. Zhang, J., Yang, K., Wang, Y., Wang, H., Sun, P., Song, L.: Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12575–12584 (2024)
27. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)