DSRC: Learning Density-Insensitive and Semantic-Aware Collaborative Representation Against Corruptions

Jingyu Zhang[1,2*], **Yilei Wang**[1*], **Lang Qian**[1], **Peng Sun**[3], **Zengwen Li**[4†], **Sudong Jiang**[4], **Maolin Liu**[4], **Liang Song**[1,2†]

[1] Academy for Engineering and Technology, Fudan University, Shanghai, China
[2] Innovation Platform for Academicians of Hainan Province, Haikou, Hainan, China
[3]Duke Kunshan University, Suzhou, China
[4]Chongqing Changan Automobile Co., Ltd. Chongqing, China
{jingyuzhang22, yileiwang23}@m.fudan.edu.cn

Abstract

As a potential application of Vehicle-to-Everything (V2X) communication, multi-agent collaborative perception has achieved significant success in 3D object detection. While these methods have demonstrated impressive results on standard benchmarks, the robustness of such approaches in the face of complex real-world environments requires additional verification. To bridge this gap, we introduce the first comprehensive benchmark designed to evaluate the robustness of collaborative perception methods in the presence of natural corruptions typical of real-world environments. Furthermore, we propose DSRC, a robustness-enhanced collaborative perception method aiming to learn **D**ensity-insensitive and **S**emantic-aware collaborative **R**epresentation against **C**orruptions. DSRC consists of two key designs: i) a semantic-guided sparse-to-dense distillation framework, which constructs multi-view dense objects painted by ground truth bounding boxes to effectively learn density-insensitive and semantic-aware collaborative representation; ii) a feature-to-point cloud reconstruction approach to better fuse critical collaborative representation across agents. To thoroughly evaluate DSRC, we conduct extensive experiments on real-world and simulated datasets. The results demonstrate that our method outperforms state-of-the-art collaborative perception methods in both clean and corrupted conditions.

Code — https://github.com/Terry9a/DSRC

Introduction

Perceiving environment accurately is crucial to ensure the driving safety of autonomous vehicles. With recent advancements in Vehicle-to-Everything (V2X) communication technology and intelligent transportation systems, multi-agent collaborative perception has emerged as a promising solution. By exchanging perceptual information, agents can achieve a comprehensive understanding of their surroundings and overcome limitations inherent in single-agent perception, such as limited perception range and obstructed field of view. Recent researches (Wang et al. 2020; Xu et al. 2022b; Hu et al. 2022; Zhang et al. 2024a) have drawn widespread attention to collaborative perception, particularly in autonomous driving. Notably, LiDAR sensors have emerged as a primary focus of state-of-the-art collaborative perception methods (Xu et al. 2022a; Yang et al. 2023; Li et al. 2021) because of their convenient information fusion advantages.

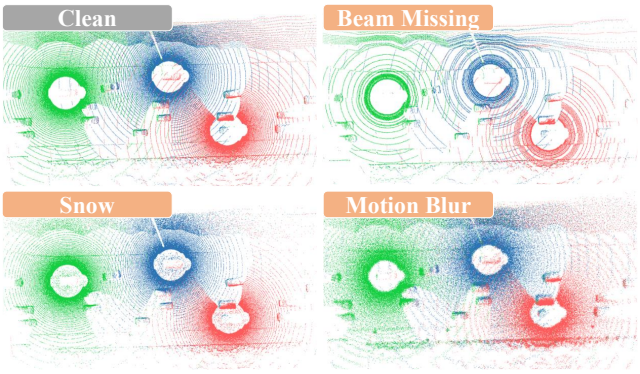


Figure 1: Visualization of typical corruption types in our benchmark. Point cloud from different agents are shown in different colors.

Collaborative perception methods must operate reliably in different geographical locations and different natural environments to ensure the safety of autonomous driving. While some works (Xu et al. 2022a; Wang et al. 2023a,c; Gu et al. 2023) achieve efficient multi-agent feature fusion through well-designed mechanisms to enhance performance within a single domain, the methods often fail in scenarios outside of clean evaluation sets. Natural corruptions like adverse weather and sensor malfunctions cause issues such as occlusion, signal attenuation, and unpredictable reflections in LiDAR data, leading to loss or misinterpretation of perceptual information. Furthermore, LiDAR's inherent limitations in capturing color and texture details hinder its generalization capability. Corruption-induced beam loss or data jitter exacerbates semantic perception difficulties, such as object shape and density. These impairments not only affect each agent's ability to accurately perceive the surrounding environment

[*]Equal contributions.
[†]Corresponding authors.

but also lead to error accumulation in multi-agent perception systems due to the sharing of high-noise perceptual information. Consequently, collaborative perception methods, when exposed to such corrupted scenes, confront risks that are correlated with critical safety concerns.

To address research gaps, we develop a series of common corruption scenarios and establish the first benchmark to comprehensively and rigorously evaluate the corruption robustness of current collaborative 3D object detection methods. Visualization of typical corruption types is shown in Figure 1. These scenarios encompass three distinct corruption sources likely to occur in real-world deployment: 1) Adverse weather conditions, such as fog or snow particles, which obstruct the line of sight of the LiDAR, resulting in sparse object perception and shape degradation (Ren, Pan, and Liu 2022); 2) Internal sensor failures, such as crosstalk between multiple sensors, which often creates noisy points within the mid-range areas; 3) External disturbances, such as dust and insects, which cause the LiDAR beam to be missing. It is essential to proactively address these common corruptions in order to ensure the reliability and robustness of collaborative perception systems. To achieve this, we propose an innovative distillation framework DSRC to learn **D**ensity-insensitive and **S**emantic-aware collaborative **R**epresentation against **C**orruptions for robust collaborative perception. It comprises two key aspects: i) A sparse-to-dense approach is designed to address point sparsity issues in 3D detection under adverse environments. By constructing multi-view dense objects and single-view sparse objects, we establish a sparse-to-dense distillation framework to learn reinforced 3D features in latent space effectively. ii) A semantic-guided approach is designed to tackle the problem of object semantics degradation in 3D detection under adverse environments. Utilizing ground truth bounding boxes to paint teacher point cloud with category semantics, the student model is guided to explore rich semantics effectively, enabling perception of sparse and incomplete objects. Additionally, we design a feature-to-point cloud reconstruction to regularize feature learning and better fuse critical collaborative representation across agents. Our method requires training exclusively on clean data and retains only the student model during inference. Thus, it does not create any additional computational burdens. To validate the effectiveness of DSRC, we conduct extensive experiments on two collaborative 3D object detection datasets OPV2V (Xu et al. 2022b) and DAIR-V2X (Yu et al. 2022). Comprehensive experimental results demonstrate that our method outperforms previous state-of-the-art methods under any corruption setting. The main contributions can be summarized as follows:

- To the best of our knowledge, we conduct the first study on the robustness of multi-agent collaborative perception systems in various corruption scenarios and establish two corruption robustness benchmarks.
- We design a sparse-to-dense and semantic-guided distillation framework to enhance the robustness of collaborative perception methods. Additionally, we devise a point cloud reconstruction module to better fuse critical collaborative representation.
- We conduct extensive experiments on both real-world and simulated datasets. The results demonstrate that DSRC outperforms state-of-the-art collaborative perception methods in both clean and corrupted conditions.

Related Works

Collaborative Perception

Collaborative perception enables multiple agents to share complementary perceptual information, promoting a more holistic perception. Based on information transmission and collaboration stages, collaborative perception modes can mainly be organized into early, intermediate, and late fusion. Among these, intermediate fusion has garnered increasing attention for its optimal balance between performance and transmission bandwidth. Several intermediate fusion methods for collaborative perception have recently been proposed. F-Cooper (Chen et al. 2019) uses an element-wise maximum strategy to aggregate shared features. V2VNet (Wang et al. 2020) introduces a spatially aware message-passing mechanism for collaborative perception. CoBEVT (Xu et al. 2022a) designs a fused axial attention module to capture sparsely local and global spatial interactions across views and agents. CodeFilling (Hu et al. 2024) achieves efficient communication by transmitting integer codes instead of high-dimensional feature maps. UniV2X (Yu et al. 2024) integrates all key driving modules in a multi-agent system into a unified end-to-end network. Although these methods show significant performance in multi-agent perception, they are primarily evaluated using standard benchmarks under clean conditions, neglecting common corruptions. This paper addresses this gap by considering the impact of corruptions on multi-agent perception systems.

Knowledge Distillation

Knowledge distillation was initially proposed in (Hinton, Vinyals, and Dean 2015) for model compression, aiming to transfer the knowledge learned by a complex teacher model to a simpler student model. Knowledge distillation includes not only label knowledge but also intermediate layer knowledge, parameter knowledge, structured knowledge, and graph representation knowledge, among other forms. Due to its effectiveness, knowledge distillation has been extensively studied in various computer vision tasks, such as object detection (Wang et al. 2023b; Huang et al. 2024) and semantic segmentation (Wang et al. 2024; Zhu et al. 2024). Ju et al. (Ju et al. 2022) leverage the semantic information hidden within objects to map semantics onto auxiliary supervisory signals, conveying guiding knowledge to enhance the performance of pure LiDAR models. Wang et al. (Wang et al. 2022) propose a multi-frame to single-frame distillation framework that uses multi-frames to generate dense features as guidance, reinforcing the sparse features of the point cloud in the latent space. RadarDistill (Bang et al. 2024) considers the sparsity and noisy nature of radar data and develops knowledge distillation to improve the representation of radar data by leveraging LiDAR data. This paper

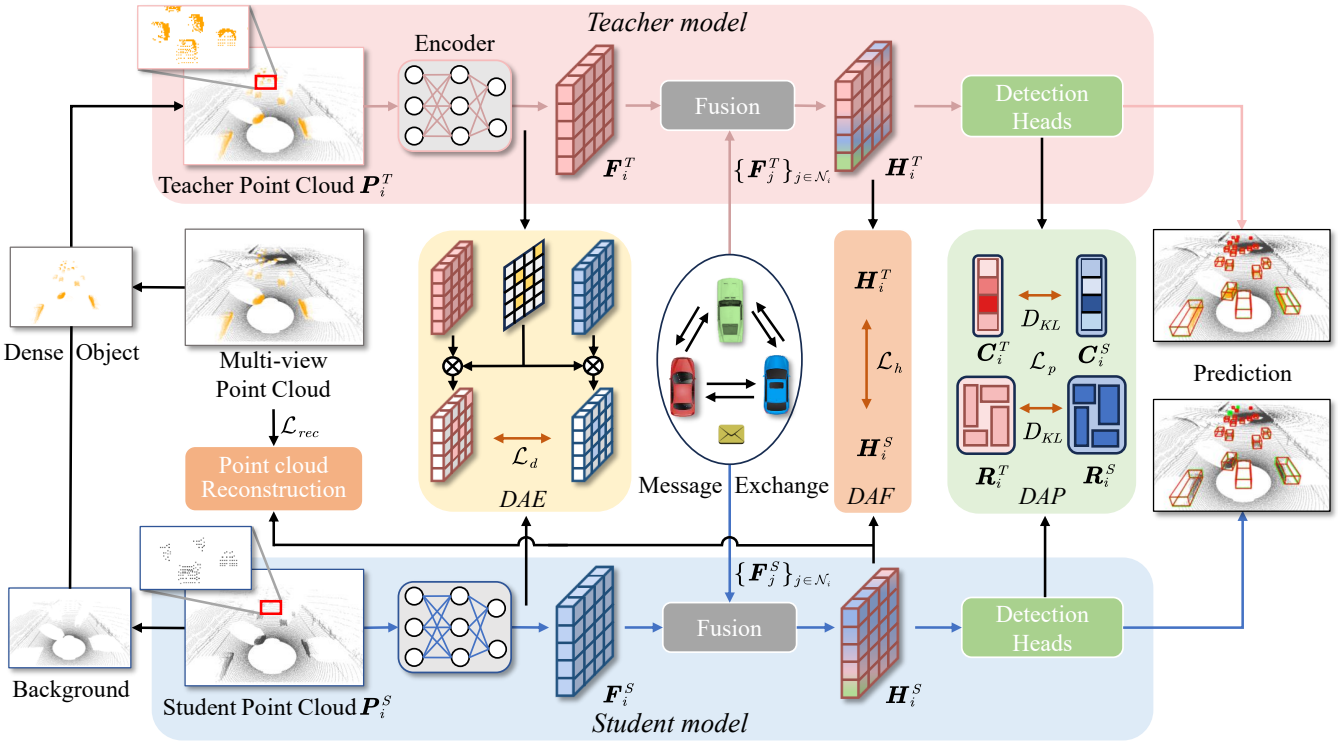


Figure 2: The overall architecture of the proposed framework DSRC. It contains two branches with identical network structures: Student (Bottom) and Teacher (Top). The framework employs a three-stage distillation strategy: distillation after encoding (DAE), distillation after fusion (DAF), and distillation after prediction (DAP) to achieve effective knowledge transfer and a point cloud reconstruction module to better fuse crucial collaborative representation across agents. During inference, the teacher model and point cloud reconstruction are discarded; only the student model (blue data flow) is retained.

proposes a three-stage distillation strategy to learn density-insensitive and semantic-aware collaborative representation against common corruptions.

Methods

In typical driving scenarios involving collaborative perception, corruptions are common. Addressing these corruptions is crucial to ensure the safety of collaborative perception systems. This study aims to achieve robust collaborative 3D object detection through density-insensitive and semantic-aware representation learning. This section introduces the overall architecture, followed by details of each phase.

Overall Architecture

As illustrated in Figure 2, our method employs a teacher-student distillation framework to learn density-insensitive and semantic-aware representation for the student model. Considering M agents in the scenario, we first fuse the multi-view point cloud of all agents to obtain a dense point cloud, representing an overall perspective. Next, we replace object regions in the single-view sparse student point cloud P^S with corresponding multi-view dense object points to obtain the dense teacher point cloud P^T. We then paint the input point cloud with ground truth labels and use these painted point cloud as input to train the teacher model. Specifically, the point cloud are encoded into Bird's Eye View (BEV) features using a 3D detector with PointPillars (Lang et al. 2019) for feature extraction. A multi-scale feature attention fusion method (Hu et al. 2022) is then employed to fuse features of all agents. Lastly, the detection heads predict the classification and regression results. The classification output is the confidence score of the foreground or background, while the regression output is a seven-element tuple $(x, y, z, w, l, h, \theta)$, where (x, y, z) denotes the object center, (w, l, h) defines the 3D box dimension, and θ represents the heading orientation, respectively. The student model shares the same structure as the teacher model, including the feature encoder, multi-scale feature attention fusion, and detection head. During training, we employ a three-stage distillation strategy to facilitate effective knowledge transfer between the teacher and student models. Additionally, point cloud reconstruction is used to enhance feature extraction and fusion. During inference, only the student model is retained.

Teacher Point Cloud Generation

Due to variations in the locations and viewpoints of multiple agents, significant diversity exists in the observation density and quality within the same spatial region. For a specific location in world coordinate, we represent its feature distribution $\mathcal{D}$ as containing all possible features of this specific location as observed from different angles, with $\boldsymbol{F}_o$ denoting

the optimal observable feature. Our objective is to achieve an optimized feature distribution $\mathcal{D}$, defined as follows:

$$\min \sum_{\boldsymbol{F}_j \in \mathcal{D}} \|\boldsymbol{F}_j - \boldsymbol{F}_o\|^2 . \tag{1}$$

In this context, optimal feature is defined as those derived from a multi-view high-density point cloud painted using ground truth bounding boxes. This process ensures that features extracted at low point densities exhibit similarities to those extracted at high point densities, while also guiding towards better semantic information capture and facilitating the convergence of features towards improved representation. Subsequently, we will introduce the multi-view construction of object point cloud and point cloud painting.

Multi-view Construction of Object Point Cloud. To obtain the teacher point cloud supervision, the raw point clouds of individual agent are aggregated to generate a comprehensive multi-view point cloud. Then, the object regions in the single-view sparse point cloud are replaced with dense object points derived from multiple views to generate the teacher point cloud. Specifically, given the raw 3D point cloud P_i collected from the ego agent i, and the 3D point cloud of all collaborative agents $\{\boldsymbol{P}_j\}_{j\in\mathcal{N}_i}$ where $\mathcal{N}_i$ denotes the neighbors of the i-th agent, we initially transform the 3D point cloud from collaborative agents to the coordinate system of the ego agent $\{\boldsymbol{P}_{j\to i}\}_{j\in\mathcal{N}_i} = \Gamma_{j\to i}\{\boldsymbol{P}_j\}_{j\in\mathcal{N}_i}$, where the transformation $\Gamma_{j\to i}$ is based on the poses ξ_i^t and ξ_j^t of two agents. Then, we aggregate each individual point cloud to construct a multi-view 3D scene: $\widetilde{\boldsymbol{P}} = \|(\{\boldsymbol{P}_{j\to i}\}_{j\in\mathcal{N}_i}, \boldsymbol{P}_i)$, where $\|$ represents the aggregate operator. Finally, we replace the object regions in the sparse point cloud $\boldsymbol{P}^S$ with the corresponding dense object points from the multi-view 3D scene to obtain the dense point cloud $\boldsymbol{P}^T$.

Point Cloud Painting. Although LiDAR excels in determining the 3D positions of objects, its monochromatic nature limits the acquisition of color and texture information. Additionally, adverse environmental conditions lead to sparse object perception and semantic information degradation, exacerbating the difficulty of object detection. To address this challenge, we propose a semantic-guided method to mitigate object semantic degradation in 3D detection under adverse environmental conditions. The method guides the student model in extracting richer semantics for the effective perception of sparse and incomplete objects by painting teacher point cloud using ground truth bounding boxes to assign category semantics. Specifically, considering the original teacher point cloud $\boldsymbol{P}^T \in \mathbb{R}^{N\times 4}$, where N denotes the number of points and the 4-tuple (x, y, z, r) represents the coordinates and the reflectance intensity. We enhance the teacher point cloud by incorporating a semantic indicator s. Given a point p_i, if it lies within the ground truth bounding box, the semantic indicator is set to 1, indicating it as an object point; conversely, the semantic indicator is set to 0, representing a background point. Consequently, we obtain the painted teacher point cloud $\boldsymbol{P}^T \in \mathbb{R}^{N\times 5}$, where the 5-tuple is (x, y, z, r, s).

Density-insensitive and Semantic-aware Distillation

We employ a three-stage distillation strategy to achieve effective knowledge transfer: distillation after encoding, distillation after fusion, and distillation after prediction. This strategy forces the student model to match the output representation from the teacher model at various granularities during the encoding, fusion, and prediction processes, thereby reducing the disparities between the two models.

Distillation After Encoding. Effective feature extraction is pivotal for accurate perception. Thus, we conduct the first stage distillation to align spatial features generated by the teacher and student models, facilitating the student model to produce dense high-quality features. Given the i-th vehicle local observations $\boldsymbol{P}_i^S$, the extracted features are represented as $\boldsymbol{F}_i^S = \Phi_{\text{enc}}(\boldsymbol{P}_i^S) \in \mathbb{R}^{H\times W\times C}$, where $\Phi_{\text{enc}}(\cdot)$ denotes the feature encoder and H, W, and C stand for the height, width, and channel of the feature map, respectively. Similarly, the features extracted from the dense teacher point cloud $\boldsymbol{P}_i^T$ are $\boldsymbol{F}_i^T$. Then, we perform feature constraints by minimizing the $l2$ distance between the two feature maps. To mitigate the effect of background noise, we use labels to create a foreground binary feature mask $\boldsymbol{M}_i \in \mathbb{R}^{H\times W}$, ensuring that the loss computation focuses only on the foreground region. The loss function can be formulated as:

$$\mathcal{L}_d = \sum_{j\in\mathcal{N}_i\cup\{i\}} \boldsymbol{M}_j \cdot \left\| \boldsymbol{F}_j^T - \boldsymbol{F}_j^S \right\|_2 . \tag{2}$$

Distillation After Fusion. By fusing perception features from all agents, we aim to achieve a highly capable fused representation. High-quality feature fusion is the first step towards holistic perception. Therefore, we perform the second-stage distillation to align the intermediate fused features $\boldsymbol{H}_i^S$ with $\boldsymbol{H}_i^T$, effectively ensuring consistent integration of each agent's perception throughout the learning process. The distillation loss is formulated as:

$$\mathcal{L}_h = \left\| \boldsymbol{H}_i^T - \boldsymbol{H}_i^S \right\|_2 . \tag{3}$$

Distillation After Prediction. Prediction discrepancies intuitively reflect significant information that distinguishes the student model from the teacher model, and our final goal is to decode the classifications and 3D bounding boxes from the fusion features. Thus, ensuring alignment at the prediction level further contributes to the consistency and accuracy of results. To this end, we employ Kullback-Leibler (KL) divergence loss to compute prediction discrepancies between the teacher and student, aiming to transfer deep knowledge by minimizing prediction differences between them. Given the class decoding outputs $\{\boldsymbol{C}_i^T, \boldsymbol{C}_i^S\}$ and regression decoding outputs $\{\boldsymbol{R}_i^T, \boldsymbol{R}_i^S\}$ for the teacher model and the student model. The prediction level distillation can be formulated as:

$$\mathcal{L}_p = D_{KL}\left(\boldsymbol{C}_i^T, \boldsymbol{C}_i^S\right) + D_{KL}\left(\boldsymbol{R}_i^T, \boldsymbol{R}_i^S\right). \tag{4}$$

In summary, the total loss of distillation is formulated as:

$$\mathcal{L}_{kd} = \alpha\cdot\mathcal{L}_d + \beta\cdot\mathcal{L}_h + \gamma\cdot\mathcal{L}_p, \tag{5}$$

where α, β and γ are balance hyperparameters.

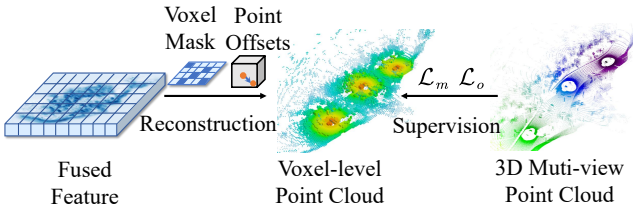


Figure 3: Illustration of the proposed point cloud reconstruction module. It provides additional supervision to better fuse critical collaborative representation across agents.

Point Cloud Reconstruction

While distillation between the teacher and student models facilitates knowledge transfer, it achieves suboptimal performance because the model relies solely on task-specific losses to indirectly learn collaborative representation fusion. Therefore, we supervise the student model on a feature-to-point cloud reconstruction to better fuse critical collaborative representation across agents.

As illustrated in Figure 3, our primary insight is that if the agent successfully acquires a reliable fused feature after message exchange and fusion, the fused feature should be able to reconstruct the complete point cloud scene in reverse. This reconstruction idea is task-agnostic and provides an explicit and sensible supervision of collaboration. However, directly reconstructing large-scale dense point clouds is challenging. Inspired by (Wang et al. 2022), we implement a voxel-level reconstruction strategy, which decouples the point cloud reconstruction task into two subtasks: occupancy mask prediction and point offsets prediction for non-empty voxels. The occupancy mask prediction generates a soft voxel occupancy mask V_m, representing the probability of a voxel being non-empty. For non-empty voxels, the point offsets prediction estimates the point offset set O_p for each voxel, representing the offsets from the voxel center V_c to the average input points of the voxel. Therefore, the reconstructed point can be expressed as follows:

$$P_c = (O_p + V_c) \times V_m. \tag{6}$$

The occupancy mask prediction loss $\mathcal{L}_m$ and point offsets prediction loss $\mathcal{L}_o$ can be expressed as follows:

$$\mathcal{L}_m = -\frac{N_b}{N_f} \sum_{j=1}^{H \times W} \left(y_j \log(p_j) + (1 - y_j) \log(1 - p_j) \right), \tag{7}$$

$$\mathcal{L}_o = \frac{1}{|N_f|} \sum_{i}^{N_f} \left| (O_{p_i} + V_{m_i}) - O_{\text{gt}_i} \right|, \tag{8}$$

where N_b and N_f represent the numbers of background and foreground voxels, p_j and y_j stand for the predicted value and ground truth of the voxel mask, H and W stand for the height and width of the mask, and j indexes the voxels in V_m. The reconstruction loss $\mathcal{L}_{rec}$ is defined as $\mathcal{L}_m$ + $\mathcal{L}_o$. During training, the overall loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{detect}} + \mathcal{L}_{kd} + \mathcal{L}_{rec}, \tag{9}$$

where $\mathcal{L}_{\text{detect}}$ is the detection loss.

Experiment

Datasets and Evaluation Metrics

Datasets. We validate the proposed DSRC in the LiDAR-based 3D object detection task using two main datasets: OPV2V (Xu et al., 2022c) and DAIR-V2X (Yu et al., 2022). **OPV2V** is a large vehicle-to-vehicle collaborative perception dataset collected by Carla (Dosovitskiy et al. 2017) and OpenCDA (Xu et al. 2021). This dataset comprises 11,464 frames of annotated 64-line point cloud and RGB images with 3D annotations. The training/validation/testing splits include 6,764, 1,981, and 2,719 frames. **DAIR-V2X** is a real-world dataset for vehicle-to-infrastructure perception. It samples 9K synchronized vehicle and infrastructure LiDAR frames from 100 representative scenes at a frequency of 10Hz. The RSU LiDAR is 300 lines, while the vehicle's LiDAR is 40 lines. The ratio of training/validation/testing sets is 5:2:3.

Evaluation Metrics. We adopt the average precision (AP) at the intersection-over-union (IoU) thresholds of 0.5 and 0.7 to evaluate the detection performance. Meanwhile, we also use the mean Corruption Error (mCE) as the primary metric to compare the model robustness following (Dong et al. 2023; Kong et al. 2023). The mCE represents the percentage of performance drop as:

$$CE_i = \frac{AP_{clean} - AP_i}{AP_{clean}}, \quad mCE = \frac{1}{N} \sum_{i=1}^{N} CE_i, \tag{10}$$

where AP_{clean} denotes the average precision on the *clean* evaluation set and N is the total number of corruption types.

Experiment Setup

Due to lacking a suitable robustness evaluation benchmark, existing 3D perception models tend to overfit clean data distributions rather than realistic scenarios. This paper aims to enhance the perception performance of multi-agent systems under unknown corruptions and establish a benchmark for evaluating the robustness of collaborative perception. Assuming a point in a LiDAR point cloud, with coordinates and intensity, we simulate a corrupted point through a mapping, where the mapping rules are constrained by physical principles or engineering experience. We simulate six common types of corruption, including beam missing, motion blur, fog, snow, crosstalk, and cross sensor.

Implementation Details

We implement the proposed and comparative models using the PyTorch framework (Paszke et al. 2019) and train them on a single RTX 3090 24G GPU using the Adam optimizer (Kingma and Ba 2015). The cosine annealing learning rate scheduler is used with an initial learning rate of 2e-3. All models are trained for 40 epochs with a batch size of 2 on the original dataset, employing early stopping to identify the optimal epoch. All detection models utilize PointPillars (Lang et al. 2019) as the backbone to extract 2D features from the point cloud, and 0.4 m width/length is used for each voxel. We assume that all agents have a communication range of 70 m following (Xu et al. 2022b). All the agents out of this

Corruptions	Clean	Beam Missing	Motion Blur	Fog	Snow	Crosstalk	Cross Sensor
Model/Dataset				OPV2V			
No Collaboration	78.85/65.05	63.50/48.01	61.99/36.79	59.88/49.38	57.56/45.48	73.73/58.47	60.64/44.00
Late Fusion	87.48/80.41	80.40/69.36	76.68/51.84	82.91/69.48	69.60/58.87	82.47/71.98	77.43/63.75
F-Cooper (Chen et al. 2019)	87.40/79.40	75.65/65.91	73.44/48.30	63.76/53.52	59.60/53.21	63.76/53.52	75.65/65.91
V2VNet (Wang et al. 2020)	92.29/83.20	83.30/70.35	83.87/65.17	68.22/57.06	73.90/66.93	90.33/77.68	81.40/67.53
V2X-ViT (Xu et al. 2022b)	91.49/83.27	82.31/70.88	80.97/61.18	70.97/60.97	64.96/56.87	78.07/66.42	80.47/68.02
CoAlign (Lu et al. 2023)	91.08/84.61	83.00/74.31	77.46/58.48	70.29/64.49	66.10/59.95	82.28/74.68	80.87/71.36
ERMVP (Zhang et al. 2024b)	92.03/85.41	81.85/69.77	84.02/67.42	73.62/65.21	68.95/63.63	85.85/80.12	79.52/72.78
Mrcnet (Hong et al. 2024)	91.73/83.28	83.42/71.88	85.71/68.20	70.17/59.86	67.54/62.85	81.25/76.47	74.29/62.06
CoBEVT (Xu et al. 2022a)	91.39/86.18	82.07/74.53	84.78/67.21	75.04/68.63	68.83/63.00	87.35/80.55	80.23/73.04
DSRC (Ours)	**92.58/88.45**	**85.82/79.59**	**86.20/69.41**	**83.54/69.84**	**74.14/67.25**	**90.76/84.57**	**85.77/77.64**
Model/Dataset				DAIR-V2X			
No Collaboration	54.19/42.43	43.11/33.62	38.74/24.53	26.19/19.75	35.75/26.82	47.83/35.05	30.01/22.70
Late Fusion	56.33/43.47	41.35/30.37	43.41/23.04	37.14/25.09	45.04/28.01	52.18/33.80	33.32/19.95
F-Cooper (Chen et al. 2019)	58.23/41.74	40.11/27.76	33.71/17.36	27.03/18.64	20.88/11.69	50.78/33.97	26.29/17.48
V2VNet (Wang et al. 2020)	61.89/44.63	46.69/31.05	54.68/33.77	36.69/25.80	48.25/32.40	60.80/42.07	31.48/19.67
V2X-ViT (Xu et al. 2022b)	59.70/43.67	41.73/29.67	44.15/29.53	31.19/23.32	37.43/29.05	56.29/39.72	26.30/17.86
CoAlign (Lu et al. 2023)	68.23/55.08	54.31/41.33	57.46/41.39	41.40/32.15	43.31/31.97	61.64/49.54	39.31/28.10
ERMVP (Zhang et al. 2024b)	61.02/46.39	46.59/34.64	49.40/31.80	35.09/26.94	42.36/31.37	56.64/41.56	28.10/20.00
Mrcnet (Hong et al. 2024)	58.15/43.93	44.51/31.82	49.67/30.78	35.81/26.65	37.72/24.63	54.29/39.64	30.67/20.82
CoBEVT (Xu et al. 2022a)	61.77/45.18	46.38/31.93	43.16/24.32	35.39/24.15	36.72/24.00	58.31/40.31	32.35/20.73
DSRC (Ours)	**69.51/56.54**	**55.62/42.62**	**59.76/42.81**	**42.83/33.65**	**45.41/33.52**	**63.99/51.26**	**39.81/28.50**

Table 1: Overall performance on OPV2V and DAIR-V2X datasets under clean and corrupted conditions. The results are reported in AP@0.5/0.7.

broadcasting radius of ego agent will not have any collaboration. The balance hyperparameters α, β, and γ are 1, 1, and 0.5. Following existing work (Lang et al. 2019), we adopt the smooth $L1$ loss for regression and focal loss (Lin et al. 2017) for classification.

Quantitative Evaluation

Comparison of Detection Performance. Table 1 presents the 3D detection performance comparison results based on two datasets. We use the No Collaboration method as a baseline, which relies solely on the LiDAR point cloud from the ego agent. Additionally, we compare this with the Late Fusion approach, where detection outputs from all agents are merged, and non-maximum suppression is applied to generate the final results. Furthermore, we evaluate several state-of-the-art methods for the intermediate fusion strategy: FCooper (Chen et al. 2019), V2VNet (Wang et al. 2020), V2X-ViT (Xu et al. 2022b), CoAlign (Lu et al. 2023), and CoBEVT (Xu et al. 2022a). We see that DSRC: i) under clean conditions, all methods achieve acceptable performance, with our method outperforming state-of-the-art collaborative perception methods on both datasets, thus showcasing the superiority of the DSRC collaboration paradigm; ii) our method exhibits strong robustness under adverse conditions, surpassing state-of-the-art collaborative perception methods across all types of corruption. For instance, it shows a performance improvement of 4.9% in cross sensor corruption at AP@0.5 and 4.02% in crosstalk corruption at AP@0.7 on OPV2V dataset.

Comparison of corruption types. Table 1 and Figure 4 indicate that all types of corruption lead to a decline in model performance, with weather-related corruption having the most significant impact. For example, snow and fog conditions result in an average drop in AP@0.7 of over 20% across all models, highlighting the threat of adverse weather to collaborative perception methods. Notably, under fog conditions, our model achieves 8.5% higher AP@0.5 than the state-of-the-art on the OPV2V dataset, demonstrating its robustness. Additionally, motion blur poses a substantial challenge to all models, likely due to noise offsets exceeding the grid size. In contrast, most models show minor performance degradation under crosstalk corruption. This is mainly because the multi-agent environment of collaborative perception makes such corruption ubiquitous in the training dataset, increasing the models' resilience to it.

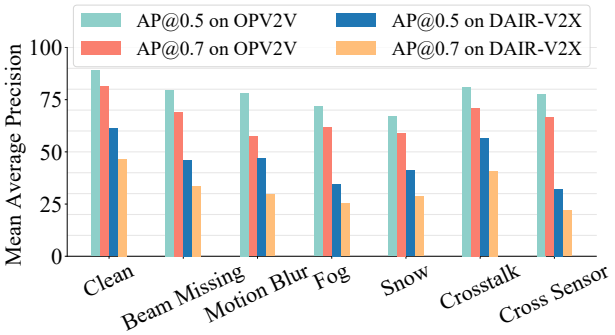


Figure 4: The average performance of all models under different corruption types.

Comparison of corruption robustness. Evaluation of model robustness under various corruption scenarios is crucial for achieving practical perception. To this end, we employ the mean corruption error (mCE) and the mean average precision (mAP) to assess the model's robustness.

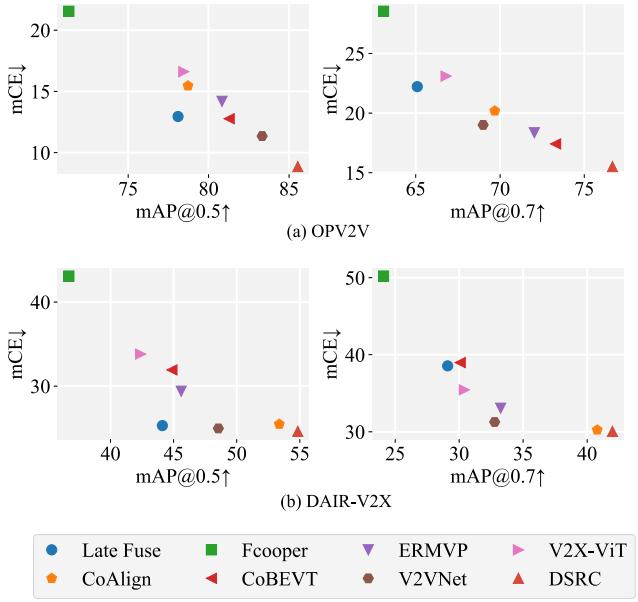


Figure 5: Benchmarking results of all models on the six robustness sets. The figure shows the mean corruption error (mCE) vs. the mean average precision (mAP).

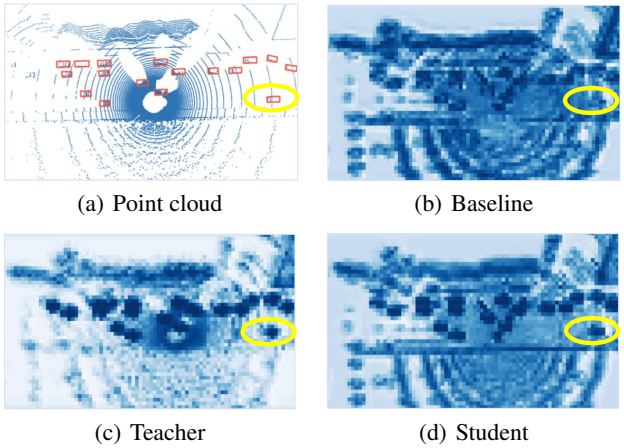
The mCE represents the average percentage of performance drop across different types of corruption, while the mAP indicates the average precision under these corruptions. As shown in Figure 5, F-cooper exhibits the poorest model performance due to its simple maximum element selection method. In contrast, the Late Fusion approach demonstrates better model robustness by avoiding the accumulation of errors from multi-agent perception features. Notably, DSRC achieves the highest accuracy and the least performance drop across both datasets, underscoring its superior robustness. The reasonable explanations are: (i) the proposed distillation framework effectively learns enhanced collaborative representation in the latent space; (ii) the point cloud reconstruction module achieves a better fusion of crucial collaborative representation across agents.

Ablation Studies

SDD	PCP	REC	OPV2V	DAIR-V2X
			79.75/70.54	53.11/40.12
✓			83.42/73.92	53.84/40.98
	✓		84.04/74.33	53.97/41.02
✓	✓		84.83/75.99	54.28/41.37
✓	✓	✓	**85.54/76.67**	**54.84/41.98**

Table 2: Ablation study results of the proposed core designs on the both datasets. SDD: Sparse to Dense Distillation; PCP: Point Cloud Painting; REC: Point Cloud Reconstruction. The results are reported with average precision under all corruptions at IoU thresholds of 0.5 and 0.7.

Effect of Core Designs. Table 2 details the contribution of each core design in our DSRC framework. The base model is a student model supervised solely by detection loss. We then assess the impact of each design by sequentially introducing: i) Sparse to Dense Distillation (SDD), ii) Point Cloud Painting (PCP), and iii) Point Cloud Reconstruction (REC). The consistent improvement in detection results across both datasets demonstrates the effectiveness of each introduced design. Notably, integrating all three components boosts detection performance by 5.79% and 6.13% on the OPV2V dataset for AP@0.5 and AP@0.7, respectively.



(a) Point cloud (b) Baseline (c) Teacher (d) Student

Figure 6: Visualisation of raw point cloud with features extracted by different models. Red 3D bounding boxes represent the ground truth.

Qualitative Evaluation

The qualitative results are shown in Figure 6. We visualize the original point cloud alongside the corresponding feature maps of three types of models: baseline, teacher, and student. Here, we use a student model supervised solely by detection loss as the Baseline. The observations are as follows: (i) the sparse-to-dense distillation framework successfully compensates for the sparse features of occluded or distant objects (circled in yellow), enabling the student model to produce denser and more robust features; (ii) compared to the Baseline, the student model exhibits a more distinct separation between the background and object regions. This improvement is attributed to the semantic-guided approach that allows the student network to extract meaningful semantic information and focus on perceptually critical features.

Conclusion

This paper proposes DSRC, an innovative collaborative perception framework to enhance robustness against common corruptions. It considers a semantic-guided sparse-to-dense distillation framework to learn density-insensitive and semantic-aware collaborative representation effectively. Meanwhile, a feature-to-point cloud reconstruction approach is introduced to fuse critical perceptual information across agents better. Our method conducts supervised learning from multiple dimensions, including the original point cloud, latent features, and predictions, to stimulate more effective collaboration. Extensive experiments demonstrate that DSRC outperforms state-of-the-art collaborative perception methods in clean and corrupted scenarios.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China, Project No. 2024YFE0200700, Subject No. 2024YFE0200703, the Specific Research Fund of the innovation Platform for Academicians of Hainan Province under Grant YSPTZX202314, the Shanghai Key Research Laboratory of NSAI, the Joint Laboratory on Networked Al Edge Computing, Fudan University-Changan, and the National Natural Science Foundation of China (Grant No. 62250410368).

References

Bang, G.; Choi, K.; Kim, J.; Kum, D.; and Choi, J. W. 2024. RadarDistill: Boosting Radar-based Object Detection Performance via Knowledge Distillation from LiDAR Features. *arXiv preprint arXiv:2403.05061*.

Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; and Fu, S. 2019. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 88–100.

Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; and Zhu, J. 2023. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1032.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Gu, J.; Zhang, J.; Zhang, M.; Meng, W.; Xu, S.; Zhang, J.; and Zhang, X. 2023. FeaCo: Reaching Robust Feature-Level Consensus in Noisy Pose Conditions. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 3628–3636. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Hong, S.; Liu, Y.; Li, Z.; Li, S.; and He, Y. 2024. Multi-agent Collaborative Perception via Motion-aware Robust Communication Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15301–15310.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-Efficient Collaborative Perception via Information Filling with Codebook. arXiv:2405.04966.

Hu, Y.; snd Zixing Lei, S. F.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*.

Huang, T.; Zhang, Y.; Zheng, M.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36.

Ju, B.; Zou, Z.; Ye, X.; Jiang, M.; Tan, X.; Ding, E.; and Wang, J. 2022. Paint and distill: Boosting 3d object detection with semantic passing network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5639–5648.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19994–20006.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lu, Y.; Li, Q.; Liu, B.; Dianati, M.; Feng, C.; Chen, S.; and Wang, Y. 2023. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4812–4818. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ren, J.; Pan, L.; and Liu, Z. 2022. Benchmarking and Analyzing Point Cloud Classification under Corruptions. arXiv:2202.03377.

Wang, B.; Zhang, L.; Wang, Z.; Zhao, Y.; and Zhou, T. 2023a. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8710–8720.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023b. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11186–11196.

Wang, Q.; Wu, Y.; Yang, L.; Zuo, W.; and Hu, Q. 2024. Layer-Specific Knowledge Distillation for Class Incremental Semantic Segmentation. *IEEE Transactions on Image Processing*.

Wang, T.; Chen, G.; Chen, K.; Liu, Z.; Zhang, B.; Knoll, A.; and Jiang, C. 2023c. UMC: A Unified Bandwidth-efficient and Multi-resolution based Collaborative Perception Framework. *arXiv preprint arXiv:2303.12400*.

Wang, T.; Hu, X.; Liu, Z.; and Fu, C.-W. 2022. Sparse2Dense: Learning to densify 3d features for 3d object detection. *Advances in Neural Information Processing Systems*, 35: 38533–38545.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*, 605–621. Springer.

Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; and Ma, J. 2021. OpenCDA: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1155–1162. IEEE.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2022a. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022b. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision*, 107–124. Springer.

Yang, K.; Yang, D.; Zhang, J.; Li, M.; Liu, Y.; Liu, J.; Wang, H.; Sun, P.; and Song, L. 2023. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. arXiv:2307.13929.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. 2022. 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*, 21361–21370.

Yu, H.; Yang, W.; Zhong, J.; Yang, Z.; Fan, S.; Luo, P.; and Nie, Z. 2024. End-to-End Autonomous Driving through V2X Cooperation. arXiv:2404.00717.

Zhang, J.; Yang, K.; Wang, H.; Sun, P.; and Song, L. 2024a. Efficient Vehicular Collaborative Perception Based on Saptial-Temporal Feature Compression. *IEEE Transactions on Vehicular Technology*, 73(11): 16125–16133.

Zhang, J.; Yang, K.; Wang, Y.; Wang, H.; Sun, P.; and Song, L. 2024b. 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*, 12575–12584.

Zhu, C.; Li, L.; Wu, Y.; and Sun, Z. 2024. Saswot: Real-time semantic segmentation architecture search without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7722–7730.