2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

CPRCNN: Collaborative Perception Method Based on Two Stages

Jinshan Yuan[1], Puyi Yao[2,*], Tiange Fu[2] and KeXin Gong[2]
[1]Unicom (Qinghai) Industrial Internet Co., Ltd, Qinghai, 810000, China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, China
[*]Corresponding author: yaopuyi@bupt.edu.cn

***Abstract*-Collaborative Perception significantly enhances the accuracy, robustness, and completeness of perception by facilitating the sharing of sensory information among multiple agents. Most existing collaborative perception methods are single-stage approaches, whereas we propose a two-stage object detection method called CPRCNN. In the first stage, information is shared among multiple agents to generate common candidate regions. This collaborative sharing ensures that the initial region proposals are more reliable and comprehensive, leveraging the diverse perspectives and data from multiple agents. In the second stage, the model refines and reselects candidate regions using multi-scale Bird's Eye View information. This fine-tuning process effectively eliminates the influence of background areas on the proposals, optimizing the region proposals and enhancing precision. Experiments conducted on the OPV2V and DAIR-V2X datasets demonstrate that CPRCNN achieves state-of-the-art performance.**

Keywords- Collaborative Perception, Object Detection, Autonomous Driving

I. INTRODUCTION

The rapid advancement of autonomous driving technology has revolutionized the transportation industry, promising safer and more efficient roadways. However, achieving fully autonomous vehicles (AVs) presents numerous challenges, especially in complex urban environments where sensor occlusions and limited fields of view can impair perception systems. Traditional perception methods, which rely primarily on onboard sensors, such as cameras, LiDAR, and radar, have made significant progress. Yet, these methods can be insufficient in scenarios where obstacles are obscured or when precise situational awareness is critical. Cooperative perception emerges as a key solution to address these challenges. By enabling vehicles to share sensor data and insights with each other and with roadside infrastructure, cooperative perception enhances the overall situational awareness of each participating entity [1]. The integration of Vehicle-to-Everything (V2X) communication technologies, such as Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I), plays a crucial role in realizing cooperative perception.

In this paper, we propose an improved two-stage cooperative perception framework by extending existing techniques to collective perception scenarios. Our main contributions are summarized as follows:

- We innovatively propose a two-stage collaborative perception method, CPRCNN. In the first stage, multi-agent information is fused to generate region proposals. In the second stage, these proposals are optimized and refined.
- We conducted comparative experiments on the OPV2V and DAIR-V2X datasets, and the results demonstrate that CPRCNN outperforms state-of-the-art (SOTA) methods.

II. RELATED WORK

A. Collaborative Perception

Collaborative perception in autonomous driving has become a focal point of research, as it significantly enhances the environmental awareness of vehicles through the sharing of data between multiple agents. The introduction of V2VNet [2] marked a breakthrough with a vehicle-to-vehicle communication framework that enhances both joint perception and localization accuracy in dynamic environments, paving the way for further advancements in the field. Building on foundational works, OpenCOOD [3] offers an open-source framework for cooperative perception, facilitating the development and testing of V2X perception algorithms, essential for advancing collaborative perception research. Who2com [4] introduces an innovative approach for selecting communication partners in a collaborative vehicle perception framework, reducing communication overhead and maintaining high perception accuracy through intelligent information sharing, suitable for large-scale deployment. Building on this concept, Where2comm [5] advances the field by proposing a data-driven communication strategy that optimizes the content of exchanged messages.

B. Two-stage 3D Object Detection

Two-stage 3D object detection methods have been pivotal in advancing the accuracy and efficiency of detecting objects in 3D space. Shi et al. enhanced the traditional two-stage pipeline with part-aware modules in Part-A^2 Net [6], notably improving detection accuracy for small and occluded objects by capturing fine-grained part information. Shi et al. proposed PV-RCNN [7], a method that integrates point-voxel features in a two-stage framework, generating high-quality 3D proposals and refining them with point-based and voxel-based features, achieving state-of-the-art performance on multiple benchmarks. In 2023, PV-RCNN++ [8] was introduced as an extension of PV-RCNN, further optimizing the integration of point and voxel features and refining the two-stage architecture to boost detection accuracy and efficiency. These modules dynamically adjust the receptive field during the refinement process, leading to more accurate object detection in complex scenes.

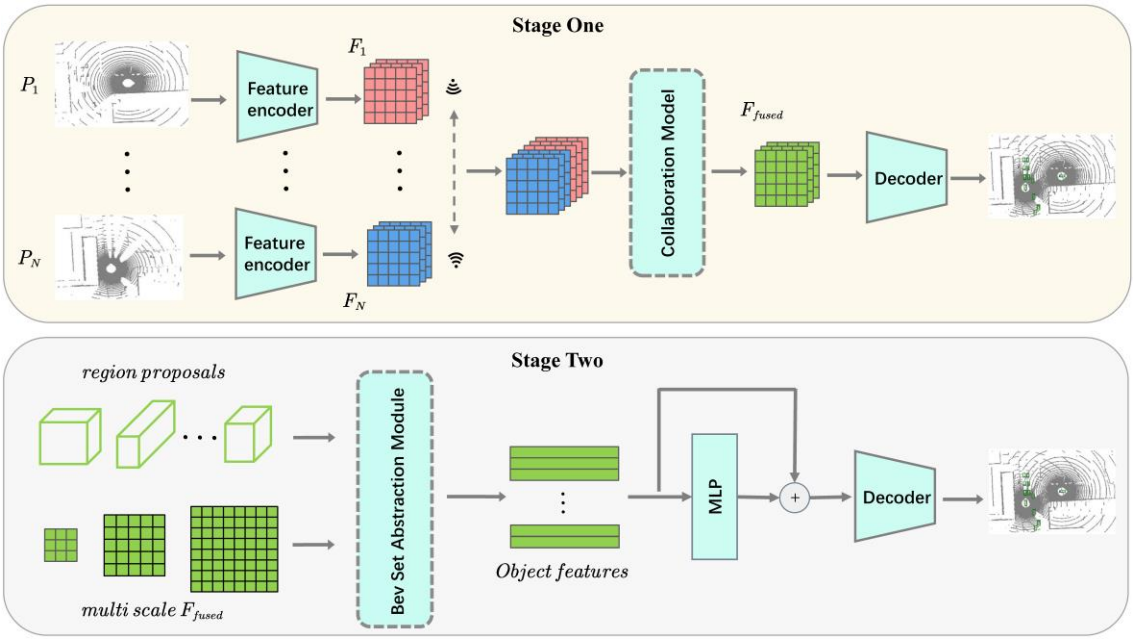

Figure 1. The framework of CPRCNN. The upper image illustrates the structure of the first stage of the model, while the lower image depicts the schematic of the second stage.

III. METHOD

A. *Fisrt-stage collaborative perception*

As shown in Figure 1, our approach consists of two main components. In a scenario with *N* agents, each equipped with sensing, communication, and detection capabilities, the goal of the first phase of collaborative perception is to detect as many potential targets in the scene as possible through distributed collaboration. The first phase of our collaborative perception method, similar to existing intermediate-feature-based collaborative perception, comprises three parts: single-agent feature extraction, cross-agent feature fusion, and proposal generation.

When extracting features from a single agent, we allow all agents to share encoder $f_{encoder}$ to extract BEV features. The process is as follows:

$$F_i = f_{enconder}(P_i)_{i=1,2,3,\cdots,N}, \tag{1}$$

where P_i denotes the point cloud data observed by the i agent, and F_i represents the Bird's Eye View (BEV) features obtained after encoding.

During cross-agent collaboration, agent i receives features F_j and the pose information pos_j from agent j. Agent i then performs projection transformation on the received features using its own pose information pos_i, followed by fusion with its own features F_i. The process is as follows:

$$F_{fused} = f_{fuse}\left(F_i, f_{transform}(pos_i, pos_j, F_j)\right). \tag{2}$$

After obtaining the fused feature F_{fused}, it is fed into the decoder $f_{decoder}$ to generate region proposals. The process is as follows:

$$(p_0, p_1, \ldots, p_n) = f_{decoder}(F_{fused}), \tag{3}$$

where $(p_0, p_1, ..., p_n)$ represents the region proposals detected in the scene, indicating areas where potential objects of interest may be present.

B. *Second-stage refinement processing*

1) Selection of proposals. The number of region proposals often exceeds the number of objects in the scene. Therefore, we have designed a distance-sampling strategy (DSS). Initially, proposals are filtered using score and IoU thresholds, significantly reducing their number. Following this initial screening, proposals are categorized based on distance. This is necessary because point clouds become sparser with increased distance, making it difficult to detect objects at the periphery. Thus, proposals are sampled according to distance to ensure a balanced division into hard, medium, and easy samples.

2) BEV feature extraction based on proposals. As shown in Figure 2, we utilize the BEV set abstraction module (BSA) to extract features from the proposals. Each region proposal has seven attributes $(x, y, z, l, w, h, \theta)$, where (x, y, z) represents the center point coordinates, (l, w, h) denotes the length, width, and height, and θ indicates the rotation angle of the bounding box. Proposals can be projected onto the BEV features, and RoiPooling is applied to extract information across multi-scale BEV features. These feature maps are flattened and concatenated, and then processed through a neural network to obtain the aggregated information.

3) Object detection based on proposals. The filtered proposals are matched with ground truths (GTs) based on the principle that if the IoU between one proposal and one GT exceeds a certain threshold, they are considered a match. When a proposal matches multiple GTs, the one with the highest IoU

is selected as its label. Proposals that cannot be matched are designated as negative samples. The information extracted by the BSA module is first fed into an MLP (Multi-Layer Perceptron), and then passed to the detection head to train the model.

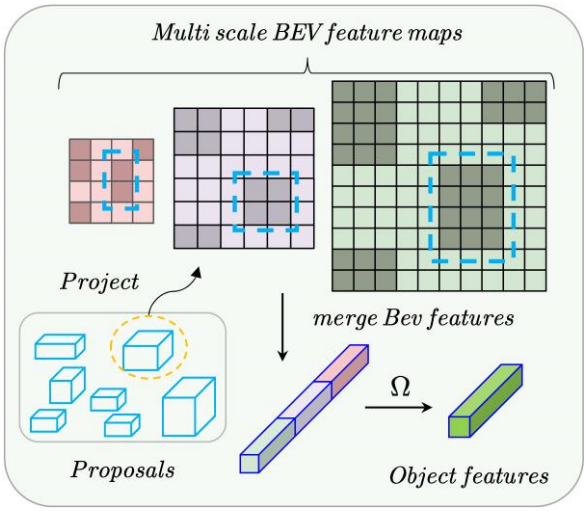

Figure 2. The framework of BEV set abstraction module.

C. Loss function design

As shown in Figure 1, our method consists of two parts, and accordingly, the loss function also has two components.

$$L_{stage1} = l_{reg}(f_{d1}(F_{fused}), GT_s) + l_{cls}(f_{d1}(F_{fused}), GT_s), (4)$$

$$L_{stage2} = l_{reg}(f_{d2}(F_{object}), GT_o) + l_{cls}(f_{d2}(F_{object}), GT_o), (5)$$

where L_{stage1} represents the loss for the first stage, where f_{d1} denotes the decoder for the first stage, and GT_s represents the ground truth labels for the scene. L_{stage2} represents the loss for the second stage, where f_{d2} denotes the decoder for the second stage, F_{object} represents the object-level features after passing through the MLP, and GT_o represents the corresponding ground truth labels.

IV. EXPERIMENTS

Table 1. Performance comparison on OPV2V and DAIR-V2X dataset.

Mothod	Publication	OPV2V. (*AP@0.5/0.7*)	DAIR-V2X. (*AP@0.5/0.7*)
V2VNet [2]	ECCV 2020	0.935/0.740	0.664/0.402
DiscoNet [10]	NeurIPS 2021	0.916/0.791	0.736/0.583
Where2comm [5]	NeurIPS 2022	0.944/0.855	0.752/0.588
V2X-ViT [11]	ECCV 2022	0.946/0.856	0.704/0.531
CoBEVT [12]	RAL 2022	0.914/0.862	0.639/0.517
CoAlign [13]	ICRA 2023	0.966/0.912	0.746/0.604
Ours		**0.967/0.916**	**0.751/0.613**

A. Datasets and Evaluation criteria

OPV2V [3] is a large-scale simulated dataset designed for evaluating CP models, featuring 3D LiDAR point cloud data from a variety of environments and scenarios. The training set comprises 6374 scenes, while the test set includes 2170 scenes. Object detection within this dataset typically covers a range of $x \in [-140m, 140m]$ and $y \in [-40m, 40m]$. DAIR-V2X [9] is a real-world CP dataset used for evaluating CP models. It features scenes involving both vehicle and roadside unit agents. The dataset consists of 4811 scenes for training and 1789 scenes for testing. The typical range for object detection within this dataset is $x \in [-100m, 100m]$ and $y \in [-40m, 40m]$.

When evaluating object detection algorithms, we typically use the Intersection over Union (IoU) metric, which measures the overlap ratio between detected objects and ground truths. The most commonly adopted standards are the accuracy at an IoU of 0.5 (denoted as AP@0.5) and the accuracy at an IoU of 0.7 (denoted as AP@0.7).

Table 2. Effectiveness of modules in the model on the OPV2V Dataset.

DSS	BSA	AP@0.5	AP@0.7
		0.928	0.894
√		0.953	0.906
	√	0.943	0.913
√	√	**0.967**	**0.916**

B. Comparative result analysis

To validate the effectiveness of our model, we compared it with previous state-of-the-art models, V2VNet [2], DiscoNet [10], Where2comm [5], V2X-ViT [11], CoBEVT [12] and CoAlign [13]. As shown in Table 1, our model outperforms these past SOTA models, achieving the best performance. Compared to the previous best model CoAlign [13], our precision on dataset OPV2V improved by 0.4% at AP@0.7, and on dataset DAIR-V2X, it improved by 0.9% at AP@0.7.

C. Ablation Studies

To validate the effectiveness of the various modules we designed, we conducted a series of ablation studies on the OPV2V dataset. As shown in Table 2, our DSS strategy and BSA module both contribute significantly to enhancing the model's performance. When these two components are combined, they synergistically lead to the best overall performance, demonstrating the complementary nature of their functionalities.

V. CONCLUSION

Our method, CPRCNN, adopts a two-stage approach for collaborative perception to achieve better detection performance. In the first stage, we perform cross-agent interaction to extract high-quality proposals that include as many potential objects as possible. In the second stage, we first filter out the high-quality proposals and then use the BSA module to aggregate BEV features. Finally, we employ the detection head to generate the detection results. This two-stage design enhances the model's ability to capture both global and local information, thereby improving its generalization capability. In future research, we plan to explore more proposal selection strategies to further enhance model performance and address lightweight issues, ensuring that the model maintains high accuracy with minimal computational overhead.

REFERENCES

[1] Caillot A, Ouerghi S, Vasseur P, Boutteau R, and Dupuis Y 2022 Survey on cooperative perception in an automotive context *IEEE Trans. Intell. Transp. Syst.* **23** 14204-14223

[2] Wang T, Manivasagam S, Liang M, Yang B, Zeng W, Tu J, and Urtasun R 2020 V2VNet: Vehicle-to-vehicle communication for joint perception and prediction *Eur. Conf. Comput. Vis.*

[3] Xu R, Xiang H, Xia X, Han X, Liu J, and Ma J 2021 OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication *IEEE Int. Conf. Robot. Autom.* 2583-2589

[4] Liu Y, Tian J, Ma C, Glaser N, Kuo C, and Kira Z 2020 Who2com: Collaborative perception via learnable handshake communication *IEEE Int. Conf. Robot. Autom.* 6876-6883

[5] Hu Y, Fang S, Lei Z, Zhong Y, and Chen S 2022 Where2comm: Communication-efficient collaborative perception via spatial confidence maps *ArXiv, abs/2209.12836*

[6] Shi S, Wang Z, Wang X, and Li H 2019 Part-A2 net: 3d part-aware and aggregation neural network for object detection from point cloud *ArXiv, abs/1907.03670*

[7] Shi S, Guo C, Jiang L, Wang Z, Shi J, Wang X, and Li H 2019 PV-RCNN: Point-voxel feature set abstraction for 3d object detection *IEEE Conf. Comput. Vis. Pattern Recognit.* 10526-10535

[8] Shi S, Jiang L, Deng J, Wang Z, Guo C, Shi J, Wang X, and Li H 2021 PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3d object detection *Int. J. Comput. Vis* **131** 531-551

[9] Yu H, Luo Y, Shu M, Huo Y, Yang Z, Shi Y, Guo Z, Li H, Hu X, Yuan J, and Nie Z 2022 DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection *IEEE Conf. Comput. Vis. Pattern Recognit.* 21329-21338

[10] Li Y, Ren S, Wu P, Chen S, Feng C, and Zhang W 2021 Learning distilled collaboration graph for multi-agent perception *ArXiv, abs/2111.00643*

[11] Xu R, Xiang H, Tu Z, Xia X, Yang M, and Ma J 2022 V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer *ArXiv, abs/2203.10638*

[12] Xu R, Tu Z, Xiang H, Shao W, Zhou B, and Ma J 2022 CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers *ArXiv, abs/2207.02202*

[13] Lu Y, Li Q, Liu B, Dianat M, Feng C, Chen S, and Wang Y 2022 Robust collaborative 3d object detection in presence of pose errors *IEEE Int. Conf. Robot. Autom.* 4812-4818