CoPEFT: Fast Adaptation Framework for Multi-Agent Collaborative Perception with Parameter-Efficient Fine-Tuning

Quanmin Wei[1, 2], **Penglin Dai**[1, 2*], **Wei Li**[1, 2], **Bingyi Liu**[3], **Xiao Wu**[1, 2]

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University
[2] Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education
[3] School of Computer Science and Artificial Intelligence, Wuhan University of Technology
wqm@my.swjtu.edu.cn, penglindai@swjtu.edu.cn, liwei@swjtu.edu.cn, byliu@whut.edu.cn, wuxiaohk@gmail.com

Abstract

Multi-agent collaborative perception is expected to significantly improve perception performance by overcoming the limitations of single-agent perception through exchanging complementary information. However, training a robust collaborative perception model requires collecting sufficient training data that covers all possible collaboration scenarios, which is impractical due to intolerable deployment costs. Hence, the trained model is not robust against new traffic scenarios with inconsistent data distribution and fundamentally restricts its real-world applicability. Further, existing methods, such as domain adaptation, have mitigated this issue by exposing the deployment data during the training stage but incur a high training cost, which is infeasible for resource-constrained agents. In this paper, we propose a Parameter-Efficient Fine-Tuning-based lightweight framework, CoPEFT, for fast adapting a trained collaborative perception model to new deployment environments under low-cost conditions. CoPEFT develops a Collaboration Adapter and Agent Prompt to perform macro-level and micro-level adaptations separately. Specifically, the Collaboration Adapter utilizes the inherent knowledge from training data and limited deployment data to adapt the feature map to new data distribution. The Agent Prompt further enhances the Collaboration Adapter by inserting fine-grained contextual information about the environment. Extensive experiments demonstrate that our CoPEFT surpasses existing methods with less than 1% trainable parameters, proving the effectiveness and efficiency of our proposed method.

Code — https://github.com/fengxueguiren/CoPEFT

Introduction

Collaborative perception allows agents to share complementary information through communication, thereby enhancing a more comprehensive perception (Han et al. 2023). This fundamentally becomes a new paradigm to overcome the long-standing limitations of single-agent perception, such as difficulties in distant and occluded perception (Ren, Chen, and Zhang 2022). Recent studies have highlighted the potential of collaborative perception in various realistic applications, including autonomous driving (Wang et al. 2020), robot automation (Li et al. 2022), and UAV collaborative rescue (Hu et al. 2023). The field of collaborative perception is experiencing rapid growth, driven by the availability of high-quality datasets (Xu et al. 2021b, 2022; Yu et al. 2022), the evolution of powerful fusion methods (Chen et al. 2019; Xu et al. 2021b), and the development of robust collaborative systems (Wei et al. 2024; Li et al. 2024a).

Most existing collaborative perceptions (Xu et al. 2021b; Hu et al. 2024) rely on the assumption that the training and deployment data follow the same distribution, but ensuring this assumption is inherently challenging. In real-world deployment, the distribution inconsistency between the training and deployment data is a common occurrence (Yuan, Xie, and Li 2023), such as unseen road topology during training or different sensing patterns due to updated sensors, which often result in performance degradation of the trained collaborative perception model. The ineffective utilization of popular adaptation methods from other fields, such as transfer learning (Zhu et al. 2023), to address this dilemma is compounded by the collaborative nature and sparsity of

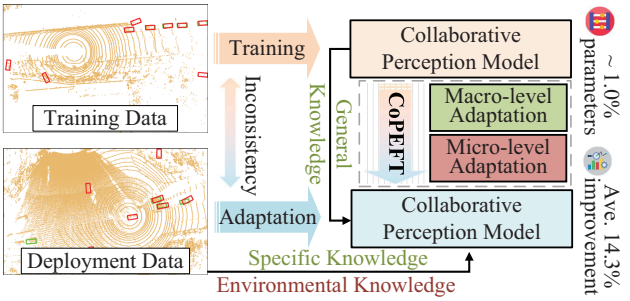

Figure 1: Illustration of CoPEFT. We mitigate the impact of inconsistent data distribution on collaborative perception by dynamically combining general knowledge derived from the training data with specific and environmental knowledge obtained from the deployment data. Here, general knowledge encompasses the general patterns of collaborative perception, specific knowledge represents the disparities between the new deployment and the training data, and environmental knowledge refers to fine-grained contextual information. Consequently, CoPEFT can fast adapt a well-trained model to various deployment environments at a low cost.

*Corresponding author.

the data in collaboration. Although some studies have explored collaborative domain adaptation by integrating deployment data during training (Li et al. 2023; Kong et al. 2023), they still require costly training from scratch for new data, which is unsuitable for resource-constrained agents. To address these issues, we aim to develop a unified and lightweight design that permits fast adaptation of the collaborative perception to inconsistent deployment environments while keeping the cost acceptable. Before that, we still need to answer the following questions. Firstly, given the inherent inconsistency in data distributions, *how can we simultaneously preserve the shared patterns and unique characteristics in training and deployment data?* Secondly, it is still essential to identify the environmental context specific to agents for aiding the adaptation process. So *how can we effectively guide collaborative perception in utilizing fine-grained environmental information?*

To address these questions, we propose a lightweight framework for collaborative perception, namely CoPEFT, which employs Parameter-Efficient Fine-Tuning for fast adaptation to new data distributions. The illustration of CoPEFT is presented in Figure 1. The adaptation process within CoPEFT is structured into two levels to tackle the questions mentioned above separately. From a macro perspective, a learnable Collaboration Adapter, with the assistance of sparse collaborative information, facilitates the dynamic combination of general knowledge from training data with specific knowledge related to deployment data, thereby adapting the feature map to the new distribution of deployment data. From a micro perspective, we develop an Agent Prompt that injects fine-grained environmental knowledge through a virtual agent, further enhancing the adaptation process. These two components jointly guide the trained collaborative perception model toward alignment with the deployment data, achieving significant performance gains while keeping costs at an acceptable level. In summary, CoPEFT seeks to equip collaborative perception with fast adaptation capabilities under limited supervision.

CoPEFT has two significant advantages. Firstly, it is resource-efficient, as it requires only a modest amount of labeled data and updates less than 1% of the parameters. This feature enables the reuse of a trained model in different deployment environments without incurring expensive adaptations. Secondly, CoPEFT seamlessly integrates with existing collaborative perception systems, functioning as a plug-and-play universal plugin that effectively operates within intermediate and aggregated feature spaces. To validate the effectiveness of our CoPEFT, we conducted extensive experiments on three benchmark datasets for collaborative 3D object detection. The results consistently indicate that our method yields substantial improvements in performance. For instance, by adapting CoAlign (Lu et al. 2023) trained on OPV2V (Xu et al. 2021b) to the DAIR-V2X (Yu et al. 2022) with a 10% data availability rate using CoPEFT, we have doubled the performance at AP@70 compared to counterparts without adaptation or those trained from scratch. In comparison to the DUSA (Kong et al. 2023) that updates all parameters for domain adaptation, CoPEFT improves the perception performance by 7.8% at AP@70 while reducing the number of trainable parameters by 99%.

In summary, our contributions are three-fold. (1) To the best of our knowledge, this is the first comprehensive exploration of fast adaptation for collaborative perception, focusing specifically on alleviating the adverse effects of data inconsistency. (2) We propose a novel fast adaptation solution called CoPEFT, which can be seamlessly integrated with existing collaborative perception systems. It comprises two complementary components: a Collaboration Adapter for macro-level adaptation and an Agent Prompt for micro-level adaptation. (3) Extensive experiments on both simulated and real-world datasets demonstrate the superior performance of CoPEFT in comparison to SOTA methods.

Related Work

Collaborative Perception

Collaborative perception overcomes inherent limitations in single-agent perception by sharing complementary information among agents (Han et al. 2023). Some early works can be categorized into early collaboration that shares raw observations (Zhang et al. 2021; Luo et al. 2023) and late collaboration that transmits perception results (Miller et al. 2020). However, these approaches often fail to strike a balance between communication efficiency and performance (Li et al. 2021), hindering their practical applications. Recently, the intermediate collaboration paradigm, which operates in a compact feature space, has gained popularity as it offers a better performance-bandwidth trade-off (Han et al. 2023).

V2VNet (Wang et al. 2020) represents a milestone in this field, employing a graph neural network to model the dynamic interactions among agents. After that, AttFuse (Xu et al. 2021b) introduces the self-attention to aggregate intermediate features from different agents and release a high-quality OPV2V dataset. To alleviate the adverse impacts of pose errors, CoAlign (Lu et al. 2023) proposes an agent pose correction method and a multi-scale fusion method. To retain the advantages of early collaboration while reducing bandwidth, DiscoNet (Li et al. 2021) and MKD-Cooper (Li et al. 2024b) introduce knowledge distillation to guide the learning of the intermediate collaboration model. Most existing efforts assume that the training data for the collaborative perception model is comparable to the data encountered during deployment. However, this assumption is often deemed impractical in real-world deployment situations.

So far, only a few works, S2R-ViT (Li et al. 2023) and DUSA (Kong et al. 2023), recognize the potential implications of this assumption not holding. These studies employ a technique called unsupervised domain adaptation, where labeled training data is combined with unlabeled deployment data during the training stage to uncover the distribution of the deployment data. They have the following limitations: (1) It is difficult to determine the deployment data during the training stage; (2) When the deployment data changes significantly, it requires costly re-training of the model from scratch; (3) This discriminative-based method may distort the learned features (Tang, Chen, and Jia 2020), thereby affecting the final perceptual performance. In contrast to previous studies that focus solely on simulation-to-reality do-

main adaptation settings, our research not only addresses the aforementioned limitations but also broadens the applicability to a wider range of scenarios.

Parameter-Efficient Fine-Tuning

In natural language processing and computer vision, Parameter-Efficient Fine-Tuning (PEFT in short) offers an efficient alternative to full-parameter fine-tuning for specific tasks (Xin et al. 2024b). The core idea behind PEFT is to achieve comparable performance to full-parameter fine-tuning by updating only a portion of the existing model's or newly added parameters. Inspired by the manually defined prompt (Petroni et al. 2019), the learnable prompt adjusts the model by adding a few parameterized input blocks into the input layer of the trained Transformer model (Jia et al. 2022; Dong et al. 2023; Nie et al. 2023). Some subsequent works have explored adjusting other elements of the Transformer architecture, such as attention block (Li and Liang 2021). Another mainstream research is adapter, which inserts subnetworks containing bottlenecks within the backbone network to fine-tune the output of each layer (Houlsby et al. 2019; Chen et al. 2022; Xin et al. 2024a). However, these works are all targeted at image or language models, which are not compatible with collaborative perception. In contrast, we introduce PEFT as a lightweight plugin that enables fast adaptation by encoding the inconsistency between the deployment and training data in collaborative perception.

In the context of collaborative perception, there is also a relevant method known as MACP (Ma et al. 2024), which introduces the concept of PEFT to transfer a single-agent perception model to multi-agent perception. Different from MACP, our goal is to achieve fast adaptation to new deployed scenario of collaborative perception with low cost by leveraging the complementary interaction of the proposed Collaboration Adapter and Agent Prompt

Methodology

Overall Architecture

Consider a set of N agents, denoted as $\mathsf{A} = \{\mathsf{A}_1, \ldots, \mathsf{A}_N\}$, that are present in the current perceptual environment. Each agent is equipped with perceptual and computational capabilities. The goal is to encourage better 3D object detection through the cooperative sharing of complementary information among agents. Specifically, this paper focuses on intermediate collaboration that achieves a performance-bandwidth trade-off. For an ego agent A_i with local observation $\mathbf{O}_i$ and perception output $\mathbf{Y}_i$, the pipeline of our CoPEFT for collaborative 3D object detection is as follows

$$\mathbf{F}_i = f_{\text{enc}}\left(\mathbf{O}_i\right) \tag{1a}$$

$$\widehat{\mathbf{F}}_i = f_{\text{c\_ada1}}\left(\mathbf{F}_i\right), \mathbf{P}_i = f_{\text{a\_pro}}\left(\widehat{\mathbf{F}}_i\right), \tag{1b}$$

$$\mathbf{H}_i = f_{\text{fus}}\left(\widehat{\mathbf{F}}_i, \left\{\widehat{\mathbf{F}}_j\right\}_{j\in\mathsf{A}, j\neq i}, \mathbf{P}_i\right), \tag{1c}$$

$$\widehat{\mathbf{H}}_i = f_{\text{c\_ada2}}\left(\mathbf{H}_i\right), \tag{1d}$$

$$\mathbf{Y}_i = f_{\text{det}}\left(\widehat{\mathbf{H}}_i\right), \tag{1e}$$

where step 1a extracts intermediate BEV feature $\mathbf{F}_i$ from $\mathbf{O}_i$ using an Encoder Network f_{enc}, step 1b generates adapted feature $\widehat{\mathbf{F}}_i$ via a Collaboration Adapter $f_{\text{c\_ada1}}$ and fine-grained prompt $\mathbf{P}_i$ via an Agent Prompt module $f_{\text{a\_pro}}$, step 1c merges intermediate features with a Fusion Network f_{fus} to generates aggregated feature $\mathbf{H}_i$, step 1d adapts $\mathbf{H}_i$ using another Collaboration Adapter $f_{\text{c\_ada2}}$, and step 1e outputs the final detection results $\mathbf{Y}_i$ by a Decoder Network f_{det}. There are two aspects that require special attention. Firstly, in intermediate collaboration, each agent needs to standardize the coordinate system and send $\mathbf{F}_i$ after performing feature extraction (i.e., step 1a); and the remaining steps will be executed after receiving all data sent by other agents. Secondly, when steps 1b and 1d are removed, Equation 1 degenerates into the standard intermediate collaboration. This plug-and-play manner endows the CoPEFT with the flexibility to adapt to various collaborative perception systems.

Deploying a trained collaborative perception model directly in a new environment significantly increases the fatal risk due to potential inconsistency in data distribution. As shown in Figure 2, after collecting a small amount of data with acceptable cost, we freeze most parameters and update only about 1% of them (including the parameters of the Collaboration Adapter, the Agent Prompt, and the Decoder Network) to adapt to new environments.

Macro-level Adaptation via Collaboration Adapter

Fast adapting collaborative perception model with new data poses a non-trivial problem under acceptable costs. On the one hand, updating only a small subset of parameters (e.g., Decoder Network) has a finite model's adaptability. On the other hand, insufficient data increases the risk of overfitting caused by noise. We note the homogeneity between the training and deployment data, which describe potentially general knowledge associated with collaborative perception, although significant differences may accompany them. Therefore, a potential solution to mitigate this dilemma is to dynamically combine general knowledge from training data with specific knowledge from limited deployment data. To concretely implement this idea, we propose the Collaboration Adapter from a macro perspective.

Specifically, the Collaboration Adapter $f_{\text{c\_ada}} : \mathbb{R}^{N\times D} \rightarrow \mathbb{R}^{N\times D}$ is a lightweight network for fast adaptation under limited supervision signals, where D denotes the feature dimension. As depicted in subgraph at the upper left corner of Figure 2, BEV feature $\mathbf{F}_i \in \mathbb{R}^{N\times D}$ is first transformed via a convolutional adapter. Distinct from conventional adapter methods (Houlsby et al. 2019; Chen et al. 2022), we adopt convolutional layers $\{\text{Conv}_{\text{up}}, \text{Conv}_{\text{down}}\}$ with a default bottleneck rate of 4 instead of linear ones to match sparse BEV inputs. With $\mathbf{F}_i$ as the input, the Collaboration Adapter $f_{\text{c\_ada}}$ can be expressed as follows

$$\begin{aligned}\widehat{\mathbf{F}}_i &= f_{\text{c\_ada}}\left(\mathbf{F}_i\right), \\ &= \underbrace{\mathbf{F}_i}_{\text{general knowledge}} \oplus \underbrace{S \odot \text{Conv}_{\text{up}}\, \sigma\left(\text{Conv}_{\text{down}}\, \mathbf{F}_i\right)}_{\text{specific knowledge}},\end{aligned} \tag{2}$$

where $\oplus$ denotes the element-wise addition, $\odot$ denotes the

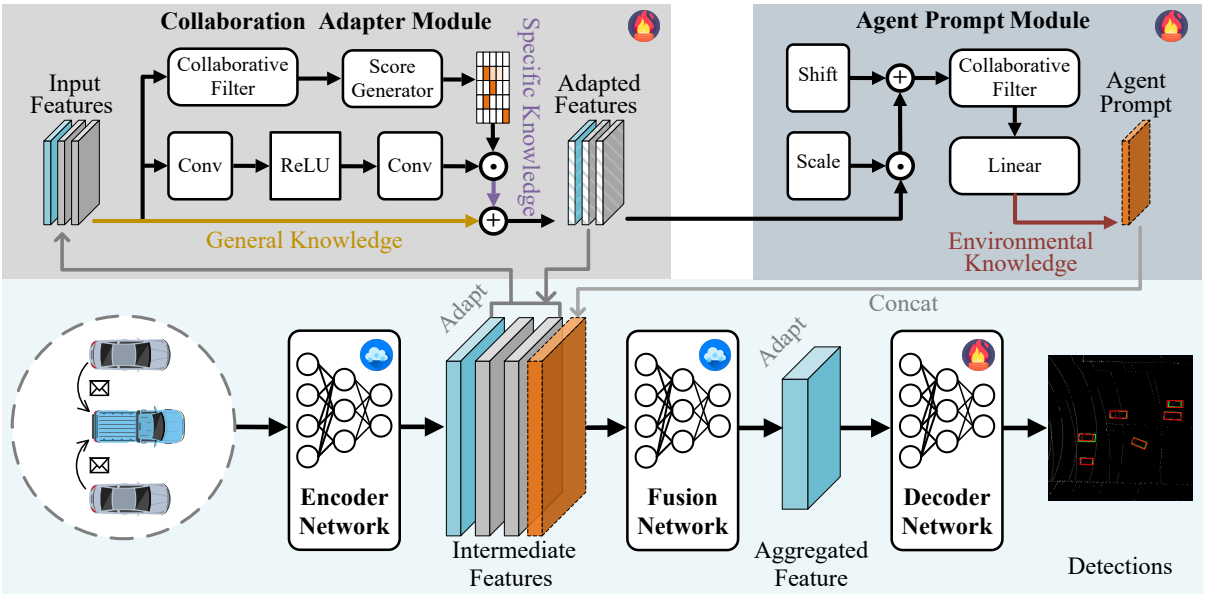

Figure 2: The overall architecture of CoPEFT. It involves standard components in intermediate collaboration augmented with two lightweight elements: a Collaboration Adapter and an Agent Prompt. (a) The Collaboration Adapter, guided by several collaborative perception priors, adapts the feature maps from a macro-level perspective for new data. (b) The Agent Prompt offers fine-grained environmental information from a micro-level perspective, which can be conceptualized as the insertion of a virtual agent to further assist in adapting feature maps. By updating only the parameters of the Collaboration Adapter, Agent Prompt, and Decoder Network, CoPEFT effectively realizes the dynamic combination of general, specific, and environmental knowledge for fast adaptation.

element-wise multiplication, and σ is the ReLU (Nair and Hinton 2010). S is a modulation score containing the priors from collaboration, which will be described in detail below.

Furthermore, the significance of inter-agent interaction and foreground confidence information is undeniable in collaborative perception. To leverage these valuable priors, we add a parallel branch into standard adapter architecture, consisting of a Collaborative Filter $\mathrm{ColF} : \mathbb{R}^{N\times D} \to \mathbb{R}^{1\times D}$ and a Score Generator $\mathrm{ScoG} : \mathbb{R}^{1\times D} \to \mathbb{R}^{1\times D}$ in series to obtain a modulation score $S \in \mathbb{R}^{1\times D}$. Specifically, these two components are implemented using parameter-free max pooling Max and convolutional layer $\mathrm{Conv}_{1\times 1}$ with kernal size 1×1, formulated as

$$\begin{aligned} S &= \mathrm{ScoG}(\mathrm{ColF}(\mathbf{F}_i)), \\ &= \mathrm{Conv}_{1\times 1}(\mathrm{Max}(\mathbf{F}_i)). \end{aligned} \tag{3}$$

In the standard CoPEFT, the aggregated feature $\mathbf{H}_i$ is adapted using another non-sharing Collaboration Adapter. The macro-level adaptation can refer to Equations 2 and 3.

Micro-level Adaptation via Agent Prompt

The Collaboration Adapter inherently provides global adaptation, referred to as macro-level adaptation, that is shared among arbitrary inputs. However, it falls short in capturing the fine-grained information in collaborative perception, where different agents occupy distinct environments. To overcome this limitation, we propose the concept of Agent Prompt, which enhances the adaptation capability from a micro-level perspective. Agent Prompt is derived from a learnable prompt but offers several notable features. Unlike the existing prompts (Houlsby et al. 2019; Xin et al. 2024a), which are typically randomly initialized and concatenated with the embeddings of other input blocks to collectively serve as the input of the Transformer layer, this improved Agent Prompt aligns with our design goal. Specifically, it is initialized with the output of the Collaboration Adapter, enabling awareness of the input instance. Furthermore, it extends into the general intermediate feature space to accommodate diverse collaborative perception systems.

As shown in the subgraph at the upper right corner of Figure 2, the Agent Prompt module $f_{\text{a\_pro}} : \mathbb{R}^{N\times D} \to \mathbb{R}^{1\times D}$ initially employs the parameter-efficient SST (Perez et al. 2018; Liu, Nguyen, and Fang 2023) to modulate the output $\widehat{\mathbf{F}}_i$ of the Collaboration Adapter $f_{\text{c\_ada1}}$, thereby generating environmental context information $\mathbf{E}_i \in \mathbb{R}^{N\times D}$:

$$\mathbf{E}_i = \mathrm{Scale} \odot \widehat{\mathbf{F}}_i \oplus \mathrm{Shift}, \tag{4}$$

where $\mathrm{Scale} \in \mathbb{R}^{C}$ and $\mathrm{Shift} \in \mathbb{R}^{C}$ are scaling and shifting operator for transformation, and C is channel. Next, it passes a ColF and enters a linear layer Linear to improve the expressive capabilities. This process yields an Agent Prompt $\mathbf{P}_i \in \mathbb{R}^{1\times D}$ that matches size of a single intermediate feature $\mathbf{F}_i$ to provide fine-grained environmental knowledge:

$$\mathbf{P}_i = \underbrace{\mathrm{Linear}(\mathrm{ColF}(\mathbf{E}_i))}_{\text{environmental knowledge}}. \tag{5}$$

Method	Publication	Parameter	AP@50	AP@70
None	-	0/12,896,384	0.429	0.217
Training from Scratch	-	12,896,384/12,896,384	0.423	0.210
Decoder Network only	-	5,140/12,901,524	0.515	0.276
SSF (Lian et al. 2022)	NeurIPS 2022	5,780/12,902,164	0.518	0.280
Adapter (Houlsby et al. 2019; Chen et al. 2022)	ICML 2019, NeurIPS 2022	42,420/12,938,804	0.579	0.380
MACP (Ma et al. 2024)	WACV 2024	43,060/12,939,444	<u>0.597</u>	<u>0.389</u>
DUSA (Kong et al. 2023)	ACM MM 2023	14,213,266/14,213,266	0.514	0.340
CoPEFT (Ours)	AAAI 2025	111,270/13,007,654	**0.610**	**0.418**

Table 1: Collaborative 3D object detection on the DAIR-V2X dataset, where the shared base model, CoAlign (Lu et al. 2023), is trained on the OPV2V dataset. None denotes the direct deployment of the collaborative perception model to new scenarios without any adaptation. The optimal and sub-optimal performances are highlighted in **bold** and <u>underline</u>, respectively.

	1%		**2%**		**5%**		**20%**	
Method	AP@50	AP@70	AP@50	AP@70	AP@50	AP@70	AP@50	AP@70
None	0.429	0.217	0.429	0.217	0.429	0.217	0.429	0.217
Training from Scratch	0.139	0.053	0.199	0.069	0.332	0.137	0.599	0.395
Decoder Network only	0.432	0.183	0.469	0.228	0.497	0.240	0.532	0.292
SSF	**0.507**	<u>0.221</u>	<u>0.513</u>	0.264	0.532	0.285	0.567	0.333
Adapter	0.315	0.089	0.466	0.199	<u>0.575</u>	0.352	0.619	0.409
MACP	<u>0.455</u>	0.192	<u>0.513</u>	0.239	<u>0.575</u>	<u>0.362</u>	<u>0.623</u>	<u>0.414</u>
CoPEFT (Ours)	**0.507**	**0.268**	**0.517**	**0.302**	**0.596**	**0.384**	**0.627**	**0.434**

Table 2: Collaborative 3D object detection under different data availability rates.

Finally, the Agent Prompt $\mathbf{P}_i$ is concatenated with the existing adapted features $\{\widehat{\mathbf{F}}_i, \left\{\widehat{\mathbf{F}}_j\right\}_{j\in\mathsf{A},j\neq i}\}$ as the input $\mathbf{I}_i \in \mathbb{R}^{N+1\times D}$ to the Fusion Network f_{fus}. Intuitively, the insertion of the Agent Prompt can be seen as a virtual agent A_{N+1} participating in the collaborative perception process. Its fusion interaction with the Fusion Network enhances the complementary adaptation for the Collaboration Adapter.

Experiments

Dataset

We conduct extensive experiments on three public benchmark datasets for multi-agent collaborative perception: OPV2V (Xu et al. 2021b), DAIR-V2X (Yu et al. 2022), and V2XSet (Xu et al. 2022) datasets. Specifically, the OPV2V dataset, a large-scale simulation dataset designed to simulate vehicle-to-vehicle interactions, comprises 11K frames of data and 232K 3D bounding boxes. Each frame includes an average of 3 agents, ranging from a minimum of 2 to a maximum of 7. The V2XSet is a vehicle-to-everything simulation dataset. Similar to the OPV2V dataset, it is jointly collected from the high-fidelity simulators CARLA (Dosovitskiy et al. 2017) and OpenCDA (Xu et al. 2021a). It contains 11K frames of data from both the intelligent vehicle and intelligent infrastructure perspectives. Finally, the DAIR-V2X dataset is the first vehicle-to-everything real dataset. DAIR-V2X is more challenging compared to other simulation datasets due to inevitable noise. We follow existing works (Lu et al. 2023; Li et al. 2024a) and employ supplementary 3D annotations for DAIR-V2X.

Experimental Setup

Evaluation Metrics. We select collaborative 3D object detection accuracy as the experimental evaluation metric. We fix the evaluation area as $x \in [-100m, 100m]$ and $y \in [-40m, 40m]$ for all datasets, thereby excluding objects outside this spatial range. The experimental results are quantified using Average Precisions (AP) at Intersection-over-Union (IoU) thresholds of 50 and 70, denoted as AP@50 and AP@70, respectively.

Settings and Implementation Details. To evaluate the effectiveness of our CoPEFT in fast adaptation for collaborative perception under low-cost conditions, we consider collaborative 3D object detection using a small amount of labeled data available with a default proportion set at 10%. This setting is reasonable as the annotation process of small data can be completed with acceptable manual effort and time, aided by automatic annotation tools. Since our method is universal for any collaborative perception method, by default, we employ a multi-scale fusion-based CoAlign method (Lu et al. 2023) as the base model to concretely implement CoPEFT. Additionally, to further evaluate the flexibility of CoPEFT, we also use two other collaborative perception models, AttFuse (Xu et al. 2021b) and MKD-Cooper (Li et al. 2024b), and vary the availability rates of the data, specifically 1%, 2%, 5%, and 20%. In contrast, the unsu-

Fusion	Method	AP@50	AP@70
AttFuse	None	0.442	0.203
	Training from Scratch	0.326	0.167
	Adapter	0.552	0.359
	MACP	0.533	0.350
	DUSA	0.457	0.324
	CoPEFT (Ours)	**0.554**	**0.374**
MKD-Cooper	None	0.320	0.157
	Training from Scratch	0.480	0.300
	Adapter	0.548	0.348
	MACP	0.537	0.339
	DUSA	0.460	0.319
	CoPEFT (Ours)	**0.554**	**0.362**

Table 3: Collaborative 3D object detection using AttFuse (Xu et al. 2021b) and MKD-Cooper (Li et al. 2024b) as base model.

pervised domain adaptation method (DUSA) requires 100% unlabeled deployment data to achieve competitive results.

We first train the collaborative perception models using their default settings to simulate the trained models. Then, we update the parameters of the Collaboration Adapter, the Agent Prompt, and the Decoder Network by utilizing partially available deployment data while keeping the parameters of the backbone network frozen. CoPEFT is optimized by the Adam optimizer with a learning rate of 0.002 and a batch size of 2. The maximum epoch for all methods is fixed at 20. Since the proportion of trainable parameters of CoPEFT is less than 1%, the adaptation time can saved by tens of times compared to some traditional domain adaptation methods (e.g., DUSA). All experiments are implemented with PyTorch on an NVIDIA 3090 GPU.

Quantitative Evaluation

As shown in Table 1 and Table 2, our CoPEFT, which has less than 1% trainable parameters, outperforms all baseline methods, including PEFT methods developed for other fields, PEFT, and domain adaptation methods tailored for collaborative perception. For instance, when adapting with only 10% of the deployment data, CoPEFT improves the detection performance by an average of 19.1% compared to the unadapted baseline, is 8.7% higher than the domain adaptation method DUSA for collaborative perception, and demonstrates promising improvements compared to PEFT methods for other fields. Several crucial observations can be revealed from these tables. Firstly, due to the substantial inconsistency between the training and deployment data, the collaborative perception model suffers from severe performance degradation, as demonstrated by the performance in the "None" rows. Secondly, training a model from scratch for a new environment requires a large amount of labeled data and excessive training costs. Thirdly, existing PEFT methods fail to achieve satisfactory adaptation performance relative to ours, potentially because they lack alignment with collaborative perception. Finally, although the DUSA for domain adaptation does not require labeled data, its performance boost is not substantial and relies on a large amount of deployment data. Consequently, our CoPEFT offers a novel strategy for collaborative perception to adapt to the new deployment environment under low-cost conditions, particularly in terms of low training and data costs, which correspond to a small number of trainable parameters and a minimal amount of labeled data, respectively.

Method	AP@50	AP@70
None	0.918	0.839
Training from Scratch	0.871	0.699
Adapter	0.928	0.849
MACP	0.930	0.851
DUSA	0.889	0.842
CoPEFT (Ours)	**0.933**	**0.854**

Table 4: Collaborative 3D object detection is adapted from the OPV2V to the V2XSet dataset.

Collaboration Adapter	Agent Prompt	AP@50	AP@70
-	-	0.429	0.217
✓	-	0.604	0.408
-	✓	0.578	0.372
✓	✓	**0.610**	**0.418**

Table 5: Ablation study on main components.

To validate the flexibility of CoPEFT, we report the results using alternative collaborative perception models as the base model in Table 3. Note that the table only includes the competitive comparison methods, and the subsequent results are presented in a similar format. It is apparent that these models are not robust against changes in the deployment environment, while CoPEFT endows them with the capability to adapt. Furthermore, the experimental results under relatively similar training and deployment distributions are detailed in Table 4, where the two datasets adopt a simulation collection scheme with comparable configurations. Despite minor performance degradation in collaborative perception models due to slight distributional differences, CoPEFT consistently delivers performance enhancements.

Qualitative Evaluation

To intuitively illustrate the superiority of CoPEFT, we present the results of qualitative comparison in Figure 3. We can observe that CoPEFT achieves better results in 3D object detection. Specifically, CoPEFT has made significant contributions in several aspects, including reducing false positives, enhancing true positives, and improving matching accuracy. These results are consistent with the quantitative evaluation mentioned above, proving that CoPEFT can effectively eliminate the impact of inconsistency between training and deployment data on collaborative perception.

Convo-lution	Collabora-tive Filter	Score Generator	AP@50	AP@70
-	-	-	0.579	0.380
✓	-	-	0.588	0.389
-	✓	-	0.588	0.392
-	-	✓	0.594	0.395
-	✓	✓	0.589	0.397
✓	✓	-	0.583	0.386
✓	✓	✓	**0.604**	**0.408**

Table 6: Ablation study on Collaboration Adapter.

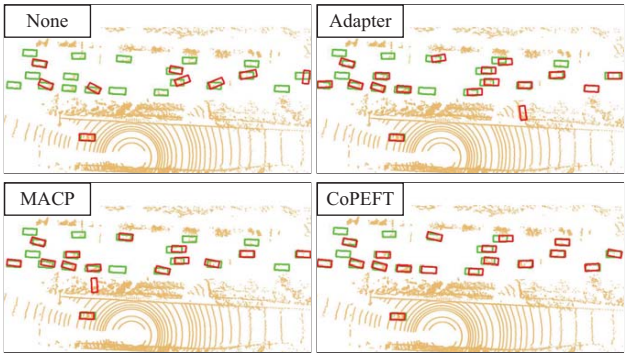

Figure 3: Qualitative comparison. The green and red 3D bounding boxes represent ground truth and prediction, respectively. Best viewed in color.

Ablation Study

Effectiveness of Main Components. We check the effectiveness of each component and report the ablation results in Table 5. The first row represents the performance without adaptation, while the second and third rows indicate the performance using only a specific component. The final row represents the complete CoPEFT. Introducing the Collaboration Adapter or Agent Prompt achieves substantial performance improvement compared to the base model without adaptation. Notably, Collaboration Adapters contribute an additional 3.6% improvement at AP@70 over the individual Agent Prompt. We further combine the two components, resulting in an average performance boost of 19.1% over the base model. These results highlight the effectiveness of each design within CoPEFT.

Internal Components of Collaboration Adapter and Agent Prompt. As shown in Tables 6 and 7, additional ablation experiments are conducted to gradually incorporate internal designs into the Collaboration Adapter and Agent Prompt. These tables follow a similar organization, where the first row corresponds to the naive PEFT method developed for other domains, and the last row represents the complete module. Both complete modules within CoPEFT surpass the baseline and their respective ablation counterparts in performance. Specifically, the Collaboration Adapter and Agent Prompt achieve average improvements of 2.6% and 5.8% over the baseline, respectively. Furthermore, they

Instance-aware	Collaborative Filter	AP@50	AP@70
-	-	0.526	0.308
✓	-	0.572	0.360
✓	✓	**0.578**	**0.372**

Table 7: Ablation study on Agent Prompt.

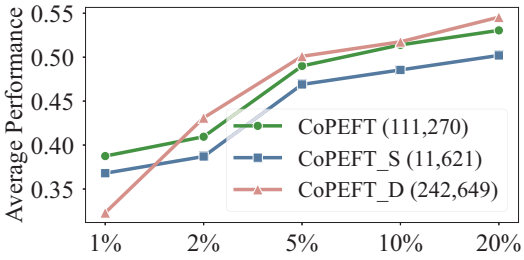

Figure 4: Comparisons of CoPEFT variations.

demonstrate varying degrees of enhancement for the incomplete ablation variants. Thus, the experimental results unquestionably validate the effectiveness of integrating collaborative perception priors into the two elements of CoPEFT.

Analysis on CoPEFT variants. We analyze the effect of incorporating our CoPEFT into various positions. The specifics of the variants, including parameters and performances, are outlined in Figure 4. Besides the standard CoPEFT that applies to the intermediate and aggregation feature space, we also introduce a more lightweight version, CoPEFT_S, which only adapts intermediate features, and a powerful CoPEFT_D, which inserts extra Collaboration Adapters into each layer of the Fusion Network. Note that none of these three variants have additional bandwidth requirements as they operate on the ego agent. The results demonstrate that CoPEFT_S significantly reduces the number of tunable parameters, yet does not offer any performance advantage. In addition, CoPEFT_D yields satisfactory results when more deployment data is available, but with roughly double parameters compared to CoPEFT. Therefore, considering the trade-off between training cost and performance, we prefer the standard CoPEFT.

Conclusion

In this paper, we investigate the performance degradation of collaborative perception in inconsistent deployment data with training data. We propose a general framework, called CoPEFT, to fast adapt collaborative perception models to the new deployment environment under acceptable costs. This framework consists of the Collaboration Adapter and Agent Prompt. The Collaboration Adapter focuses on macro-level adaptation by aligning feature maps with the distribution of the deployment data. Conversely, the Agent Prompt pursues micro-level adaptation by incorporating fine-grained environmental information. CoPEFT only updates less than 1% of parameters with the cooperation of the two components, rendering it an efficient and effective solution.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant Numbers 62172342, 62372387 and 62001400, the Natural Science Foundation of Hebei Province, China under Grant Number 2022105003, Key R&D Program of Guangxi Zhuang Autonomous Region, China (Grant No. AB22080038, AB22080039), Sichuan Science and Technology Program (Grant No. 2024NSFSC0494), China Postdoctoral Science Foundation (Grant No. 2020T130547, No.2021M702713) and Fundamental Research Funds for the Central Universities (2682024ZTPY044).

References

Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2019. Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds. In *International Conference on Distributed Computing Systems (ICDCS)*, 514–524.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 16664–16678.

Dong, B.; Zhou, P.; Yan, S.; and Zuo, W. 2023. LPT: Long-tailed Prompt Tuning for Image Classification. In *International Conference on Learning Representations (ICLR)*.

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

Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; and Li, Y. 2023. Collaborative Perception in Autonomous Driving: Methods, Datasets, and Challenges. *IEEE Intelligent Transportation Systems Magazine*, 15: 131–151.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning (ICML)*, 2790–2799.

Hu, Y.; Fang, S.; Xie, W.; and Chen, S. 2023. Aerial Monocular 3d Object Detection. *IEEE Robotics and Automation Letters*, 8(4): 1959–1966.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-Efficient Collaborative Perception via Information Filling with Codebook. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15481–15490.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 709–727.

Kong, X.; Jiang, W.; Jia, J.; Shi, Y.; Xu, R.; and Liu, S. 2023. DUSA: Decoupled Unsupervised Sim2Real Adaptation for Vehicle-to-Everything Collaborative Perception. In *ACM International Conference on Multimedia (ACM MM)*, 1943–1954.

Li, J.; Xu, R.; Liu, X.; Li, B.; Zou, Q.; Ma, J.; and Yu, H. 2023. S2R-ViT for Multi-Agent Cooperative Perception: Bridging the Gap from Simulation to Reality. *arXiv preprint arXiv:2307.07935*.

Li, X.; Yin, J.; Li, W.; Xu, C.; Yang, R.; and Shen, J. 2024a. DI-V2X: Learning Domain-Invariant Representation for Vehicle-Infrastructure Collaborative 3D Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 3208–3215.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 4582–4597.

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 (NeurIPS)*, 34: 29541–29552.

Li, Y.; Zhang, J.; Ma, D.; Wang, Y.; and Feng, C. 2022. Multi-Robot Scene Completion: Towards Task-Agnostic Collaborative Perception. In *Conference on Robot Learning (CoRL)*, 2062–2072.

Li, Z.; Liang, H.; Wang, H.; Zhao, M.; Wang, J.; and Zheng, X. 2024b. MKD-Cooper: Cooperative 3D Object Detection for Autonomous Driving via Multi-teacher Knowledge Distillation. *IEEE Transactions on Intelligent Vehicles*.

Lian, D.; Zhou, D.; Feng, J.; and Wang, X. 2022. Scaling & Shifting Your Features: A New Baseline for Efficient Model Tuning. *Advances in Neural Information Processing Systems*, 35: 109–123.

Liu, Z.; Nguyen, T.-K.; and Fang, Y. 2023. On Generalized Degree Fairness in Graph Neural Networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 4525–4533.

Lu, Y.; Li, Q.; Liu, B.; Dianat, M.; Feng, C.; Chen, S.; and Wang, Y. 2023. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4812–4818.

Luo, G.; Shao, C.; Cheng, N.; Zhou, H.; Zhang, H.; Yuan, Q.; and Li, J. 2023. EdgeCooper: Network-Aware Cooperative LiDAR Perception for Enhanced Vehicular Awareness. *IEEE Journal on Selected Areas in Communications*, 42: 207–222.

Ma, Y.; Lu, J.; Cui, C.; Zhao, S.; Cao, X.; Ye, W.; and Wang, Z. 2024. MACP: Efficient Model Adaptation for Cooperative Perception. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3373–3382.

Miller, A.; Rim, K.; Chopra, P.; Kelkar, P.; and Likhachev, M. 2020. Cooperative Perception and Localization for Cooperative Driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1256–1262.

Nair, V.; and Hinton, G. E. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on Machine Learning (ICML)*, 807–814.

Nie, X.; Ni, B.; Chang, J.; Meng, G.; Huo, C.; Xiang, S.; and Tian, Q. 2023. Pro-tuning: Unified prompt tuning for vision tasks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI conference on artificial intelligence (AAAI)*, volume 32.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2463–2473.

Ren, S.; Chen, S.; and Zhang, W. 2022. Collaborative Perception for Autonomous Driving: Current Status and Future Trend. *ArXiv*, abs/2208.10371.

Tang, H.; Chen, K.; and Jia, K. 2020. Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8725–8735.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; Tu, J.; and Urtasun, R. 2020. V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction. In *European Conference on Computer Vision (ECCV)*, 605–621.

Wei, S.; Wei, Y.; Hu, Y.; Lu, Y.; Zhong, Y.; Chen, S.; and Zhang, Y. 2024. Asynchrony-Robust Collaborative Perception via Bird's Eye View Flow. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 28462–28477.

Xin, Y.; Du, J.; Wang, Q.; Lin, Z.; and Yan, K. 2024a. VMT-Adapter: Parameter-Efficient Transfer Learning for Multi-Task Dense Understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 14, 16085–16093.

Xin, Y.; Luo, S.; Zhou, H.; Du, J.; Liu, X.; Fan, Y.; Li, Q.; and Du, Y. 2024b. Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey. *arXiv preprint arXiv:2402.02242*.

Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; and Ma, J. 2021a. OpenCDA: An Open Cooperative Driving Automation Framework Integrated with Co-Simulation. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1155–1162.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *European Conference on Computer Vision (ECCV)*, 107–124.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Liu, J.; and Ma, J. 2021b. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *International Conference on Robotics and Automation (ICRA)*, 2583–2589.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21329–21338.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust Test-Time Adaptation in Dynamic Scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15922–15932.

Zhang, X.; Zhang, A.; Sun, J.; Zhu, X.; Guo, Y. E.; Qian, F.; and Mao, Z. M. 2021. EMP: Edge-assisted Multi-vehicle Perception. In *International Conference on Mobile Computing and Networking (MobiCom)*, 545–558.

Zhu, Z.; Lin, K.; Jain, A. K.; and Zhou, J. 2023. Transfer Learning in Deep Reinforcement Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.