Privacy-Preserving V2X Collaborative Perception Integrating Unknown Collaborators

Bin Lu[1], Xinyu Xiao[2], Changzhou Zhang[3], Yang Zhou[4], Zhiyu Xiang[1], Hangguan Shan[1], Eryun Liu[1*]

[1]Zhejiang University
[2]Ant Group
[3]Hangzhou Geely Automobile Digital Technology Co., LTD
[4]Hangzhou Fenghua Technology Co., LTD
lubin2022@zju.edu.cn, xinyuxiao@nlpr.ia.ac, zhangchangzhou@geely.com, zy@geely.com, Xiangzy@zju.edu.cn, hshan@zju.edu.cn, eryunliu@zju.edu.cn

Abstract

Vehicle-to-everything (V2X) collaborative perception has recently gained increasing attention in autonomous driving due to its ability to enhance scene understanding by integrating information from other collaborators, e.g. vehicles or infrastructure. Existing algorithms usually share deep features to achieve a trade-off between accuracy and bandwidth. However, most of these methods require joint training of all agents, which results in privacy leakage and is impractical and unacceptable in the real world. Sharing prediction results seems to be a direct solution, but its performance is suboptimal and sensitive to localization noise and communication delay. In this paper, we propose a privacy-preserving collaborative perception framework, where each agent is separately trained with its own dataset and the ego vehicle needs to integrate with completely unknown collaborators. Specifically, we propose MSD, a multi-scale feature fusion method combined with deformable attention, to better fuse features of different agents. We also propose PLDA, a plug-in domain adapter to align the features from unknown collaborators to ego-domain. Extensive experiments on the challenging DAIR-V2X and V2V4Real demonstrate that: 1) MSD achieves remarkable performance, outperforming others by at least 2.8% and 6.7% on DAIR-V2X and V2V4Real, respectively; 2)After domain adaptation, it significantly outperforms the No Fusion, Late Fusion scenarios and can approach or even surpass the performance of joint training. We truly achieves privacy-preserving collaboration, providing a new paradigm for the study of collaborative perception, which is crucial for practical applications.

Introduction

Perceiving the complex driving environment is crucial for the safety of autonomous driving. Recent advancements in deep learning have improved the performance of modern perception systems on many tasks, such as object detection (Lang et al. 2019; Zhou and Tuzel 2018), semantic segmentation (Pan et al. 2020), tracking (Fan et al. 2022b), etc. Despite the remarkable progress, challenges remain. Single-agent perception systems are still subject to many limitations due to single-view constraints. For instance, autonomous vehicles often encounter occlusions, which are difficult to handle due to lack of sensory observations of the occluded area, which can also potentially lead to catastrophic consequences.

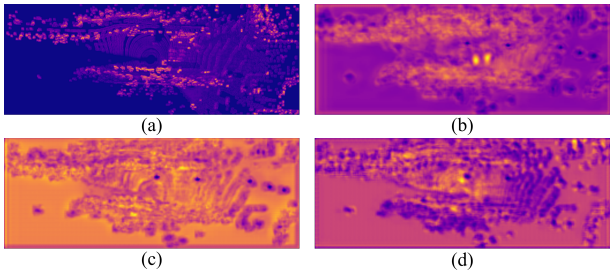

Figure 1: Domain gap across datasets and backbones. (a) feature map by PointPillars trained on DAIR-V2X, (b) feature map by VoxelNet trained on DAIR-V2X, (c) feature map by PointPillars trained on DAIR-V2X with different weights, (d) feaature map by PointPillars trained on V2V4Real.

One solution to this challenge is by sharing information between different agents, which is known as multi-agent perception or collaborative perception (Liu et al. 2023). Depending on the fusion strategy, it can be divided into data-level early collaboration (Chen et al. 2019b), feature-level intermediate collaboration (Hu et al. 2022) and output-level late collaboration (Rawashdeh and Wang 2018). Early collaboration aggregates the raw measurements from all agents, promoting a global perspective. It can achieve better performance but requires a lot of communication bandwidth. On the contrary, late collaboration aggregates each agent's perception outputs, although it is bandwidth-efficient, the sharing of possibly poor perception outputs of a certain vehicle may results in unsatisfying fusion results, and its performance degrades rapidly with the presence of localization noise. Intermediate collaboration aggregates intermediate features across agents, achieving better trade-off between performance and bandwidth. Due to its robustness to noises and the compressibility of features, it is gradually becoming the mainstream choice for collaboration.

Although existing intermediate collaboration algorithms (Wang et al. 2020; Hu et al. 2022; Xu et al. 2022b,a) have achieved significant performance improvements, they are all based on an assumption that all agents are equipped with identical models with the same parameters. However, this is impossible in the real world, especially in autonomous driving. Vehicles and infrastructure products produced by different companies are usually equipped with different detection models, and even vehicles from the same company may have different versions, depending on the vehicle type and updating frequency. In reality, there is a large domain gap between the shared features extracted from heterogeneous agents, which might lead to a dramatic performance drop in collaborative perception.

Collaboration between heterogeneous agents has always been a difficulty. HM-ViT (Xiang, Xu, and Ma 2023) designs a heterogeneous 3D graph transformer for inter-agent and intra-agent interactions to efficiently fuse features from multi-view images and Lidar point clouds. DI-V2X (Li et al. 2024) proposes a progressive domain-invariant distillation module to encourage student models from different domains to gradually learn a domain-invariant feature representation towards the teacher and a domain-adaptive fusion module to fill domain gaps. However, all these methods need to be implemented through joint training of all agents on the same dataset, which may cause privacy leakage and bring security risks, so it is almost impossible to meet in the real world. HEAL (Lu et al. 2024) proposes to sustain a unified features space, and new agents align their own features to it through backward alignment to join the collaboration. But it is not easy to establish a unified standard that could be generally accepted by everyone, and backward alignment is time-consuming, it will change the weight of its own model, which may affect single-agent perception. MPDA (Xu et al. 2023a) proposes a learnable feature resizer to align features in multiple dimensions and a sparse cross-domain transformer for domain adaptation. Although it can realize independent training of backbones, the weight changes of the fusion module and the detection head could still weaken the performance of self-perception, which may lead to terrible consequences when there is no collaborator in reality.

In this paper, we consider the actual deployment of collaborative perception, that is, each agent trains its own model separately with its own dataset, the ego vehicle needs to integrate information from completely unknown collaborators. At the same time, when there is no collaborator in the actual scene, the ego vehicle should maintain its ability to complete self-perception. Individual training can avoid any form of information leakage, regardless of the data or model used. However, since the collaborator is trained independently thus completely unknown, there is inevitably a domain gap between the collaborator and the ego vehicle. As can be seen from Figure 1, there is a clear gap between the features extracted by different backbones, and the same is true for backbones with the same structure but different weights. It can also be found that the inconsistency of the training dataset will also increase the gap between features. In this regard, as shown in Figure 2, we propose a privacy-preserving collaborative perception framework with two training stages: single-agent training and adaptation training. In the first stage, each agent uses its own data to train its own detector, and in order to achieve collaboration, the ego vehicle's detector also includes a fusion module. In the second stage, each agent first runs its detector on the same public dataset to generate an adaptation dataset, which contains shared features. Then each agent uses this dataset to train its own feature domain adapter. This individual training design not only ensures the independence of each agent in achieving single-agent perception, but also eliminates automotive companies' concerns about information leakage. Furthermore, the design of the adapter as a plug-and-play module enables it to better adapt to the high update speed of each automotive company's detectors today. By retaining a new version of the adapter, efficient collaboration can be continuously achieved.

To evaluate the proposed method, we conduct extensive experiments on two real-world datasets, DAIR-V2X and V2V4Real. Results show MSD's remarkable performance, outperforming all compared SOTA baselines. Through our framework, after feature domain adaptation, the performance can be close to or even exceed that of joint training. Our contributions can be summarized as follows:

1) We propose a privacy-preserving collaborative perception framework that enables individual training of each agent. This is a framework that is truly suitable for practical deployment and can ensure the information security of all agents, providing a new paradigm for the study of collaborative perception.

2) We propose a multi-scale feature fusion method combined with deformable attention (MSD), which is highly robust to various noises and achieves remarkable performance.

3) We propose a lightweight plug-and-play adapter (PLDA), which exhibits strong cross-model and cross-dataset domain adaptation capabilities, and can be easily combined with many popular fusion algorithms.

Related Work

3D Lidar Detection

3D detection is a prerequisite for the success of autonomous driving. Compared to other sensors, Lidar based 3D detection can offers superior performance. 3D Lidar detection methods can be divided into three categories: point-based (Yang et al. 2020), voxel-based (Yan, Mao, and Li 2018), and point-voxel-based (Yang et al. 2019). Point-based methods can extract more detailed features but have a larger computational overhead. PointRCNN (Shi, Wang, and Li 2019) adopts a two-stage strategy. The first stage generates some proposals and the second stage finetunes and optimizes them through feature aggregation. Voxel-based methods have high computational efficiency, but information loss occurs due to discretization operations. VoxelNet (Zhou and Tuzel 2018) uses 3D convolution to extract voxel features and compress them into BEV features for detection. PointPillars (Lang et al. 2019) directly uses 2D convolution to process pillar features, further improving computational efficiency. Point-voxel-based methods aim to combine the advantages of both, which can take into account both speed and accuracy. SA-

SSD (He et al. 2020) deeply mines the structural information of the target through auxiliary learning. PV-RCNN (Shi et al. 2020) combines the efficiency of 3D sparse convolution with the flexible receptive field of PointNet (Qi et al. 2017) to extract more discriminative point cloud features.

Collaborative Perception

Due to the limited field of view, the performance of single-vehicle perception cannot break through the limitations of occlusions and long distance issues. Collaborative perception studies how to use the information of neighboring agents to enhance the perception ability of the ego vehicle itself. V2VNet (Wang et al. 2020) is a pioneering work that generate motion forecasts by sharing intermediate features. AttFuse (Xu et al. 2022c) uses self-attention mechanism to fuse features while F-Cooper (Chen et al. 2019a) adopts the maxout operation. Where2comm (Hu et al. 2022) implement spatial location selection to save communication bandwidth by introducing spatial confidence map and How2comm (Yang et al. 2024) extend the selection to the element level to further achieve for more efficient collaboration. At the same time, there are some transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020; Liu et al. 2021) based methods. V2X-ViT (Xu et al. 2022b) uses window-based attention to fuse features of multiple agents while CoBEVT (Xu et al. 2022a) designs a fused axial attention module which captures sparsely local and global spatial interactions across agents to efficiently fuse features. HM-ViT (Xiang, Xu, and Ma 2023) combines the two to achieve multi-modal collaboration. HEAL (Lu et al. 2024) and UMC (Wang et al. 2023) both propose multi-scale feature fusion methods and achieve good performance.

Domain Adaptation

In deep learning, the inconsistency of data distribution and downstream tasks can cause domain gaps and thus affect the performance of the model. Domain adaptation is a transfer learning technology that can efficiently reduce the domain gap by adaptively adjusting the source domain to the target domain. Domain adaptation has been widely used in various tasks of computer vision (Fan et al. 2022a; Xu et al. 2021; Du et al. 2021; Shao et al. 2021). (Yosinski et al. 2014) comprehensively explores feature transferability of deep convolution neural networks, and then several works of feature domain adaptation emerged. DLID (Chopra, Balakrishnan, and Gopalan 2013) trains a joint source and target CNN architecture with two adaptation layers. DDC (Tzeng et al. 2014) applies a single linear kernel to one layer to minimize Maximum Mean Discepancy (MMD) while DAN (Long et al. 2015) minimizes MMD with multiple kernels applied to multiple layers. ReverseGrad (Ganin and Lempitsky 2015) adds a binary classifier to explicitly confuse the two domains. In V2X collaborative perception, there is a large domain gap between the features of different agents. Therefore, inspired by this, we use domain adaptation to perform feature alignment for better collaboration.

Proposed Method

In this paper, we explore the feasibility of collaborative perception in real-world deployments, where different agents train their own model independently using their own dataset. We mainly focus on the task of Lidar-based 3D detection in autonomous driving, but the methodology can also be applied to other autonomous tasks as long as they broadcast features for collaboration. Our goal is to build a robust collaborative perception framework that fully preserves the privacy of all vehicles. Figure 2 shows the overall architecture of our framework.

Overall Architecture

For simplicity, consider a scenario with two agents, an ego vehicle and a collaborator (vehicle or infrastructure), both of which are equipped with Lidar and have the ability to perceive, communicate and detect. The ultimate goal is to improve the perception performance of the ego vehicle by integrating the information from the collaborator. In reality, for the protection of privacy and intellectual property rights, different agents usually train their own models separately on their own datasets. Therefore, we propose a privacy-preserving collaborative perception framework, as shown in Figure 2. Let $O_{e/c}$ denote the raw observation, where subscripts e and c denote the ego vehicle and collaborator, respectively, our framework works as follows:

$$F_e = f_{backbone,e}(O_e), \tag{1}$$

$$F_c = f_{backbone,c}(O_c), \tag{2}$$

$$M_{c\to e} = \{F_c, P_c\}, \tag{3}$$

$$F_c = f_{align}(F_c, P_c, P_e), \tag{4}$$

$$F_{c\to e} = f_{adapt}(F_c), \tag{5}$$

$$F_e^{'} = f_{fusion}(F_c, F_{c\to e}), \tag{6}$$

$$D_e = f_{head,e}(F_c^{'}), \tag{7}$$

where $F_{e/c}$ is the feature extracted from backbone $f_{backbone,e/c}$, $M_{c\to e}$ is the message transmitted from the collaborator to the ego vehicle, which contains its feature F_c and pose P_c. Then, according to the poses P_c, P_e of the two, F_c is aligned to the ego vehicle's coordinate system through feature alignment f_{align}. After domain adaptation f_{adapt} and feature fusion f_{fusion}, the fused feature $F_e^{'}$ is obtained, and finally it is passed through the detection head f_{head} to get the detection results D_e. Note that single-agent detection only includes step (1) and (7) with $F^{'} = F$.

Unlike existing methods that require agents to share model information and conduct joint training, in our framework, all agents' detectors, including backbone and detection head, and fusion module for ego vehicle, are trained independently on their own datasets. This undoubtedly avoids any form of information leakage. At the same time, the training of the domain adapter only requires all agents to run their detectors on the same public dataset to obtain feature pairs, so our framework protects the privacy of all participants well. Our framework has two training stages, single-agent training and adaptation training, which are introduced in the following sections.

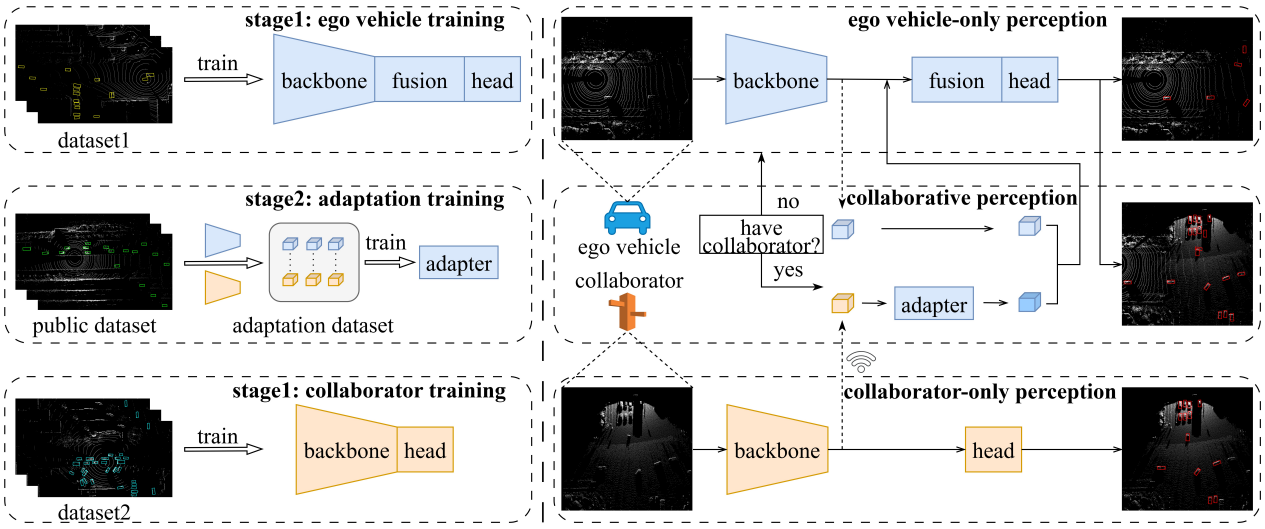

Figure 2: The overall architecture of the privacy-preserving collaborative perception framework. The left part is the training phase, which include two stages, while the right part is the inference phase. In our experiments, fusion and adapter are the proposed MDS and PLDA, respectively.

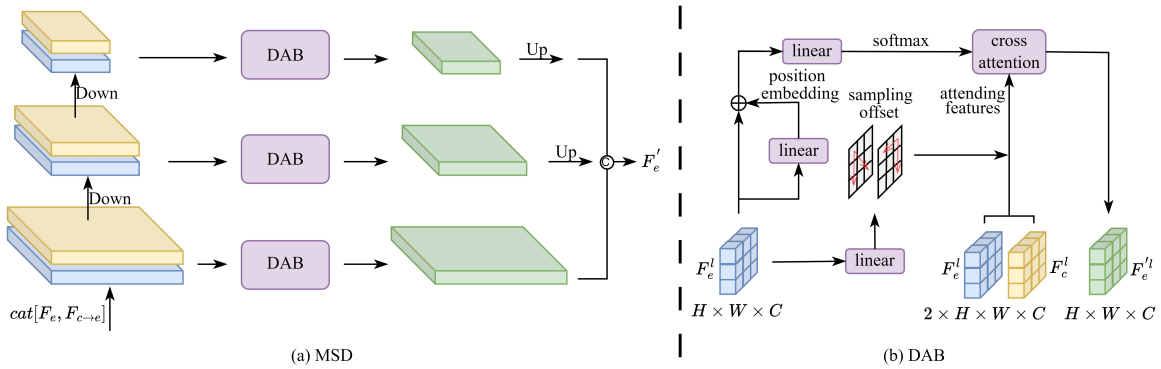

Figure 3: (a) the architecture of MSD. (b) the architecture of deformable attention block (DAB).

Single-agent Training

In this stage, we need to train separate detectors for all agents participating in the collaboration. Since the ego vehicle needs to fuse the features of all agents during the final inference, the ego vehicle's detector should also contain a feature fusion module. To promote better feature integration, we propose MSD, as shown in Figure 3(a).

Due to the presence of positioning noise and transmission delay, the features of the collaborator after spatial transformation will inevitably have discretization errors and misalignments, so we apply a multi-scale pyramidal fusion to enhance its robustness. This is because each BEV grid in the higher-scale features has a larger actual receptive field, allowing for larger positioning errors. At the same time, combining fusion at different scales can also promote better global and local interactions. Specifically, we use convolution for downsampling and transposed convolution for upsampling.

In driving scenarios, the scope of the vehicle's response to various road conditions and events is often limited. It is crucial for the vehicle to understand the surrounding road conditions within a certain range, and in most cases this is enough for the vehicle to make a sufficiently wise decision. Therefore, we adopt the deformable attention block (DAB) based on deformable cross attention (Zhu et al. 2020) to fuse features of each scale, as shown in Figure 3(b).

We take each grid in F_e^l, where the superscript l indicates the lth scale, as a reference point and extract the corresponding feature as the initial query, and encode the locations of reference points into a position embedding through a linear layer. The attention scores are learned from the initial queries via a linear layer and the softmax function. Subsequently, a linear layer is used to learn the offset maps, which provides the 2D spatial offset $\{\triangle q_s | 1 \leq s \leq N_s\}$ for each query q. We sample the keypoints based on the learned offset maps and extract keypoints' features to form the attending feature. After the cross attention layer, DAB outputs the

enhanced feature for each query q as:

$$DAB(q) = \sum_{a=1}^{A} W_a [\sum_{s=1}^{N_s} \phi(W_b F_e^l(q)) F_e^l(q + \triangle q_s) + \sum_{s=1}^{N_s} \phi(W_c F_e^l(q)) F_c^l(q + \triangle q_s)], \tag{8}$$

where a index the attention head, $W_{a/b/c}$ are the learnable parameters and ϕ is the softmax function. Finally, we concatenate the features of different scales to generate the fused feature $F_e^{'}$.

We use the smooth $l1$ loss for bounding box regression and a focal loss for classification. Thus for the basic object detection task, we minimize the following detection loss:

$$\mathcal{L}_{det} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{cls}\mathcal{L}_{cls}, \tag{9}$$

where λ_{reg} and λ_{cls} are the hyperparameters to control the weights pf regression loss and classification loss.

Adaptation Training

After the first stage, all participants have acquired and possessed the ability to perceive the environment independently. In this stage, we consider how to achieve efficient collaboration between the ego vehicle and the collaborator. Since the first stage is carried out independently, the collaborator is now completely unknown to the ego vehicle, and there is definitely a large domain gap between the features $F_e \in R^{H_e \times W_e \times C_e}$ and $F_c \in R^{H_c \times W_c \times C_c}$ of the two. Therefore, we design a plug-in domain adapter (PLDA) to align the feature of the collaborator to the ego domain, as shown in Figure 4.

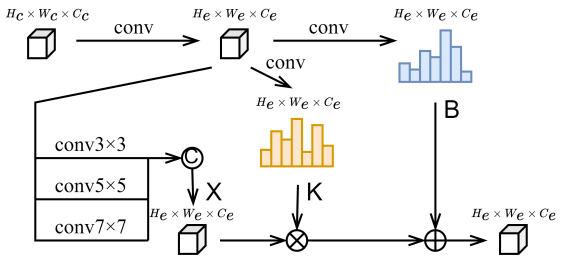

Figure 4: The architecture of PLDA.

The process of domain adaptation is as follows:

$$F_c = f_{reshape}(F_c), \tag{10}$$

$$K, X, B = f_K(F_c), f_X(F_c), f_B(F_c), \tag{11}$$

$$F_{c \to e} = KX + B, \tag{12}$$

where $f_{reshape}$, f_K, f_X and f_B are all convolution blocks. First, we use convolution to align the shape of F_c to F_e, that is, $F_c \in R^{H_c \times W_c \times C_c} \rightarrow F_c \in R^{H_e \times W_e \times C_e}$. Then, we use convolutions with three kernel size of 7×7, 5×5, and 3×3 for feature enhancement, further refine the resized feature, and finally generate a more robust feature representation $X \in R^{H_e \times W_e \times C_e}$. Next, we try to model the adaptation process between the two feature domains. Specifically, by learning a scaling factor $K \in R^{H_e \times W_e \times C_e}$ and a translation factor $B \in R^{H_e \times W_e \times C_e}$ for each element, we finally combine translation and scaling, that is, $KX + B$, to obtain the adapted feature $F_{c \to e}$. It is worth noting that compared with traditional linear transformations, neural network based methods have stronger modeling capabilities, thereby learning more complex and accurate transformations. Compared with MPDA (Xu et al. 2023a) which adopts many cross-attention blocks, we only use a small number of convolutions, which greatly reduces the computational overhead. At the same time, PLDA can be used as a plug-and-play module without changing any part of the original detector, thus not sacrificing the performance of self-perception.

To train the adapter, the ego vehicle and the collaborator need to run their respective detectors on the same public dataset to generate a training set containing shared features, as shown in Figure 2. Specifically, we adopt the mean squared error to supervise the training, as follow:

$$\mathcal{L} = \mathcal{L}_{mse}(F_{c \to e}, F_e), \tag{13}$$

where F_e and $F_{c \to e}$ are the features of the ego vehicle and the features of the collaborator after adaptation, respectively.

Experiments

Dataset

DAIR-V2X We employ the challenging DAIR-V2X (Yu et al. 2022) for evaluating our method and other SOTA approaches. DAIR-V2X is the first large-scale, multi-modality, multi-view dataset from real scenarios for vehicle-infrastructure cooperative autonomous driving. It has 9000 frames featuring one vehicle and one road side unit, both equipped with a Lidar and a camera. The original dataset does not provide annotations outside the camera's view, so we use the full-view annotations provided by CoAlign (Lu et al. 2023). In our experiments, we merge the four categories of car, van, truck and bus into one category.

V2V4Real V2V4Real (Xu et al. 2023b) is the first large-scale real-world dataset for vehicle-to-vehicle cooperative perception in autonomous driving. It is collected by two vehicles simultaneously in the same scene, providing multi-view multi-sensor datastream, covering various road sections including city roads, highways and freeways. It supports multiple tasks such as 3D object detection, object tracking and sim2real domain adaptation. Our experiments only focus on Lidar-based 3D object detection.

Implementation Details

On the dataset DAIR-V2X, the maximum communication distance is set to 100m, and on V2V4Real it is 70m. The evaluation range in x and y directions are [-102.4m, 102.4m] and [-38.4m, 38.4m], respectively. During training, we randomly select an agent as the ego, while during inference, on the contrary, a fixed one is selected. All models are trained end-to-end for for 60 epochs on RTX 3090. Adam (Kingma and Ba 2014) is used for optimization. An initial learning rate of 0.001 is selected, and is multiplied by 0.1 every 20 epochs during training. Early stop is used to find the best

	SDS		SDD		DS		DD		
	AP0.5	AP0.7	AP0.5	AP0.7	AP0.5	AP0.7	AP0.5	AP0.7	Inference time(ms)
No Fusion	71.7	60.4	71.7	60.4	71.7	60.4	71.7	60.4	/
Late Fusion	69.2	50.1	72.5	59.3	69.1	50.0	72.9	57.6	/
Intersame	81.1	68.0	81.1	68.0	81.1	68.0	81.1	68.0	/
Interbefore	68.3	57.6	68.2	58.0	66.6	56.1	68.0	57.5	/
MPDA(Xu et al. 2023a)	81.0	67.9	81.0	67.7	81.2	68.0	81.0	67.9	34.02
PLDA(Ours)	81.2	68.1	81.1	67.9	81.5	68.5	80.9	67.8	10.18

Table 1: 3D detection performance. SDS, SDD, DS, DD are the four scenarios introduced before. 'Interbefore' represents the direct fusion of the features of the ego vehicle and the collaborator without domain adaptation. 'Intersame' means that the ego vehicle and the collaborator have exactly the same backbone, which is equivalent to joint training. Inference time only includes the adapter.

	DAIR-V2X		V2V4Real	
	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	69.1	58.7	53.1	37.8
Late Fusion	67.1	48.5	67.0	40.2
Fcooper(Chen et al. 2019a)	77.0	57.6	<u>70.7</u>	38.4
AttFuse(Xu et al. 2022c)	75.0	58.1	67.2	40.8
Where2comm(Hu et al. 2022)	77.2	61.7	68.6	<u>43.2</u>
V2X-ViT(Xu et al. 2022b)	<u>80.8</u>	<u>65.2</u>	68.9	42.0
CoBEVT(Xu et al. 2022a)	77.5	62.3	69.3	36.7
V2VNet(Wang et al. 2020)	79.4	62.8	68.4	41.8
MSD(Ours)	**81.1**	**68.0**	**72.2**	**49.9**

Table 2: Performance Comparison on DAIR-V2X and V2V4Real datasets. The bold ones are the highest, and the underlined ones are the second.

epoch. In all experiments, λ_{reg} is set to 2, λ_{cls} is set to 1 and N_s is set to 9.

Experiments Setup

When evaluating the adapter, we consider the following four scenarios according to the backbone structure, weights and training set:

- SDS: Same structure, different weights, same training set. That is, the ego vehicle and the collaborator are equipped with backbones with the same structure but different weights trained from the same dataset;
- SDD: Same structure, different weights, different training set. That is, the ego vehicle and the collaborator are equipped with backbones with the same structure but different weights trained from different dataset;
- DS: Different structure, same training set. That is, the ego vehicle and the collaborator are equipped with backbones with different structures, but the training set is the same;
- DD: Different structure, different training set. That is, the ego vehicle and the collaborator are equipped with backbones with different structures, and the training set is also different.

In all experiments, the backbone of the ego vehicle is all PointPillars, while the collaborator is PointPillars or VoxelNet in different scenarios.

Quantitative Evaluation

Performance Benchmarking Table 2 compares the detection performance of the proposed MSD with various models on three datasets. We consider two typical baselines. No Fusion means single-agent detection without any collaboration. Late Fusion is to transmit the proposals and get the final detection results through non-maximum suppression (NMS). As for intermediate approaches, we consider six state-of-art methods Fcooper, AttFuse, Where2comm, V2X-ViT, CoBEVT and V2VNet. It can be seen that all collaborative methods exceed the No Fusion beseline, except for Late Fusion on DAIR-V2X, probably due to the pose error. This fully demonstrates the advantage of collaboration. MSD outperforms all previous methods on both real dataset and virtual dataset, which proves its superiority and robustness to real-world noises. Specifically, on DAIR-V2X, the AP0.7 of MSD is at least 2.8% higher than that of others, while it is 6.7% on V2V4Real.

Adaptation Performance

Table 1 depicts the performance comparison of various methods in the four scenarios of SDS, SDD, DS and DD mentioned above. It is worth noting that in all experiments, the detector of the ego vehicle is trained on DAIR-V2X. In adition, to verify the cross dataset adaptation ability of PLDA, the detector of the collaborator is trained using V2V4Real in both SDD and DD scenarios. Similarly, due to the pose noises, the performance of Late Fusion has decreased compared to No Fusion. It can be seen that when domain adaptation is not performed, the AP0.5 and AP0.7 of the intermediate fusion, denoted by Interbefore, are lower than the No Fusion baseline. This performance degradation also reveals the negative impact of the domain gap. It can be seen that compared with MPDA, both can achieve performance close to joint training, while PLDA can even surpass it, and the inference time of PLDA is only about 1/3 of that of MPDA, which proves that PLDA is is lightweight and more suitable for actual deployment. After domain adaptation, the performance can match or even exceed the results of joint training. Specifically, in the DS scenario, after domain adaptation, PLDA's AP0.5 and AP0.7 exceed Intersame by 0.4% and 0.5%, No Fusion by 9.8% and 8.1%, Late Fusion by 12.4% and 18.5%, respectively.

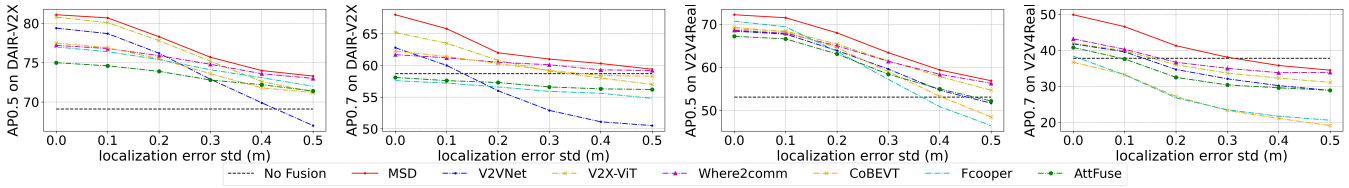

Figure 5: Robustness to localization error on DAIR-V2X and V2V4Real.

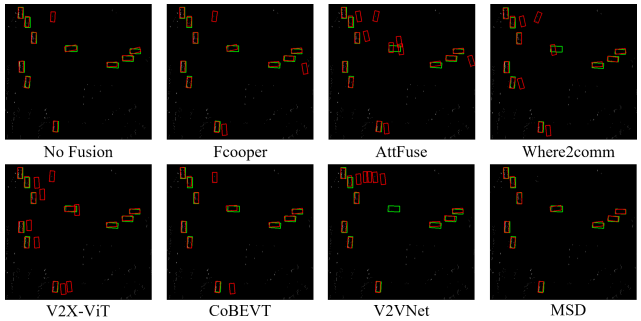

Figure 6: Qualitative comparison results from DAIR-V2X. Green and red boxes denote ground truths and detection results.

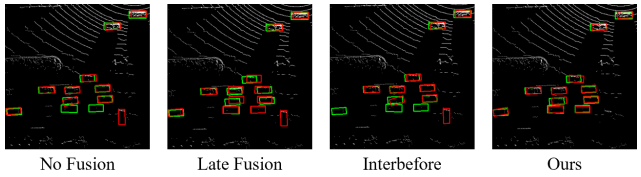

Figure 7: 3D detection visualization. Green and red boxes denote ground truths and detection results.

The experimental results verify the feasibility of the framework we proposed. When the collaborative participants are trained independently and thus unknown to each other, offline domain adaptation can achieve the same performance as joint training. At the same time, the adapter as a plug-in will not affect the original self-perception. This privacy-preserving collaborative perception framework is more in line with actual deployment and provides a new paradigm for the study of collaborative perception.

Qualitative Results

We conduct a qualitative analysis of MSD's performance by visualizing typical samples from the DAIR-V2X dataset. Figure 6 shows the detection results of No Fusion and all compared intermediate approaches. MSD predicts more accurate bounding boxes while other approaches exhibit larger displacements. We sum the values of all channels to visualize the features. As shown in Figure 8, before domain adaptation, there is a clear visible gap between the features of the ego vehicle and the collaborator. After domain adaptation, the two are visually closer, which proves the effectiveness of PLDA. Moreover, we visualize the detection results in the DS scenario in Figure 7. Obviously, without domain adaptation, there are many missed and false detections, while after domain adaptation, the bounding boxes are more accurate and significantly better than those of No Fusion and Late Fusion.

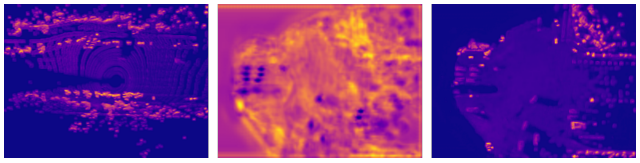

Figure 8: Visualization of features before and after domain adaptation. The left is the ego vehicle's feature, the middle is collaborator's feature before domain adaptation, the right is collaborator's feature after domain adaptation.

Ablation Study

Robustness to Localization errors We verify the performance of MSD at different levels of localization error on both DAIR-V2X and V2V4Real datasets. Specifically, we add Gaussian noise with a variance of 0 and a mean ranging from 0 to 0.5 meter to the collaborator's pose. As can be seen from Figure 5, as the pose error increases, the performance of all models decreases due to the misalignment of the features. On DAIR-V2X, when the error reaches a large value, the AP0.7 of Fcooper, AttFuse, CoBEVT and V2VNet are even lower than the No Fusion baseline. Noticeably, MSD outperforms all other models and No Fusion at every error level. This comparison demonstrates the robustness of MSD against collaboration pose noises. A reasonable explanation is that the deformable fusion way can adaptively select and focus on the key areas, and it also benefits from the multi-scale setting.

Conclusion

We propose a multi-scale deformable attention fusion mechanism that achieves good performance with less computational overhead. We also propose a privacy-preserving collaborative perception framework, which is a realistic framework and provides a new paradigm for the study of collaborative perception. Experimental results show that offline domain adaptation can achieve or even exceed the performance of joint training. We hope that our work can accelerate the actual implementation of collaborative perception. In the future, we will expand our work to domain adaptation between different modalities like RGB cameras.

Acknowledgments

This work is supported in part by Zhejiang Province Key R&D programs, China (Grant No. 2025C01039, 2024C01017, 2024C01010), and in part by the National Natural Science Foundation of China under Grants 62027805, U21B2029, and U21A20456, and in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010006.

References

Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; and Fu, S. 2019a. 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.

Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2019b. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 514–524. IEEE.

Chopra, S.; Balakrishnan, S.; and Gopalan, R. 2013. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2. Citeseer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Z.; Li, J.; Su, H.; Zhu, L.; and Lu, K. 2021. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3937–3946.

Fan, Z.; He, Y.; Wang, Z.; Wu, K.; Liu, H.; and He, J. 2022a. Reconstruction-aware prior distillation for semi-supervised point cloud completion. *arXiv preprint arXiv:2204.09186*.

Fan, Z.; Zhu, Y.; He, Y.; Sun, Q.; Liu, H.; and He, J. 2022b. Deep learning on monocular object pose detection and tracking: A comprehensive overview. *ACM Computing Surveys*, 55(4): 1–40.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

He, C.; Zeng, H.; Huang, J.; Hua, X.-S.; and Zhang, L. 2020. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11873–11882.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35: 4874–4886.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Liu, S.; Gao, C.; Chen, Y.; Peng, X.; Kong, X.; Wang, K.; Xu, R.; Jiang, W.; Xiang, H.; Ma, J.; et al. 2023. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Lu, Y.; Hu, Y.; Zhong, Y.; Wang, D.; Chen, S.; and Wang, Y. 2024. An Extensible Framework for Open Heterogeneous Collaborative Perception. *arXiv preprint arXiv:2401.13964*.

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.

Pan, B.; Sun, J.; Leung, H. Y. T.; Andonian, A.; and Zhou, B. 2020. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3): 4867–4873.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Rawashdeh, Z. Y.; and Wang, Z. 2018. Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3961–3966. IEEE.

Shao, W.; Zhao, S.; Zhang, Z.; Wang, S.; Rahaman, M. S.; Song, A.; and Salim, F. D. 2021. FADACS: A few-shot adversarial domain adaptation architecture for context-aware parking availability sensing. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 1–10. IEEE.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10529–10538.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Chen, G.; Chen, K.; Liu, Z.; Zhang, B.; Knoll, A.; and Jiang, C. 2023. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8187–8196.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 605–621. Springer.

Xiang, H.; Xu, R.; and Ma, J. 2023. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 284–295.

Xu, R.; Li, J.; Dong, X.; Yu, H.; and Ma, J. 2023a. Bridging the domain gap for multi-agent perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6035–6042. 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. *arXiv preprint arXiv:2207.02202*.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. 2023b. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13712–13722.

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.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022c. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589. IEEE.

Xu, T.; Chen, W.; Wang, P.; Wang, F.; Li, H.; and Jin, R. 2021. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2024. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36.

Yang, Z.; Sun, Y.; Liu, S.; and Jia, J. 2020. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11040–11048.

Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; and Jia, J. 2019. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1951–1960.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

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.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.