DI-V2X: Learning Domain-Invariant Representation for Vehicle-Infrastructure Collaborative 3D Object Detection

Xiang Li[1*], Junbo Yin[1*], Wei Li[2], Chengzhong Xu[3], Ruigang Yang[2], Jianbing Shen[3†]

[1]School of Computer Science, Beijing Institute of Technology
[2]Inceptio
[3]SKL-IOTSC, CIS, University of Macau
{lixianggoing, yinjunbocn}@gmail.com

Abstract

Vehicle-to-Everything (V2X) collaborative perception has recently gained significant attention due to its capability to enhance scene understanding by integrating information from various agents, *e.g.*, vehicles, and infrastructure. However, current works often treat the information from each agent equally, ignoring the inherent domain gap caused by the utilization of different LiDAR sensors of each agent, thus leading to suboptimal performance. In this paper, we propose *DI-V2X*, that aims to learn *D*omain-*I*nvariant representations through a new distillation framework to mitigate the domain discrepancy in the context of **V2X** 3D object detection. *DI-V2X* comprises three essential components: a domain-mixing instance augmentation (DMA) module, a progressive domain-invariant distillation (PDD) module, and a domain-adaptive fusion (DAF) module. Specifically, DMA builds a domain-mixing 3D instance bank for the teacher and student models during training, resulting in aligned data representation. Next, PDD encourages the student models from different domains to gradually learn a domain-invariant feature representation towards the teacher, where the overlapping regions between agents are employed as guidance to facilitate the distillation process. Furthermore, DAF closes the domain gap between the students by incorporating calibration-aware domain-adaptive attention. Extensive experiments on the challenging DAIR-V2X and V2XSet benchmark datasets demonstrate *DI-V2X* achieves remarkable performance, outperforming all the previous V2X models. Code is available at https://github.com/Serenos/DI-V2X.

1 Introduction

Vehicle-infrastructure collaborative perception (Hu et al. 2022; Xu et al. 2023b; Lu et al. 2023), also known as Vehicle-to-Everything (V2X) communication, has recently emerged as a promising approach to achieve full automation. It adequately leverages the sensor data collected from diverse agents, *i.e.*, the vehicles and road-side infrastructures, to precisely perceive the complex driving scene. For instance, in situations where a vehicle's line of sight might be obstructed, information from infrastructures can serve as
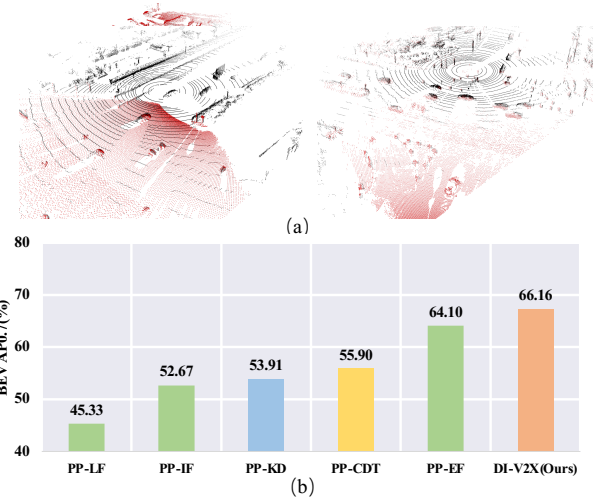
Figure 1: (a) Examples to show the data-level domain gap caused by the different types of LiDAR sensors from the vehicle side (40-beam mechanical LiDAR colored in black) and the infrastructure side (300-beam solid-state LiDAR colored in red). (b) Performance comparison of different solutions to mitigate the domain discrepancy, where our model attains superior performance.

crucial redundancy due to their distinct viewpoints. This collaboration essentially broadens the perception range and reduces occurrences of blind spots, enhancing the overall perception capability in contrast to previous single-vehicle autonomous driving systems.

To effectively fuse the information from various agents, leading V2X approaches tend to employ feature-based intermediate collaboration, which is referred to as intermediate fusion (Wang et al. 2020). This approach retains essential information from each agent at the feature level, which can then be compressed for efficiency. Therefore, intermediate fusion ensures a performance-bandwidth trade-off that is superior to the early fusion or late fusion approaches, where the former needs to transmit raw point cloud data between agents while the latter is vulnerable to incomplete results produced by each model.

Nevertheless, current intermediate fusion models primarily concentrate on enhancing interaction between features from different agents. To date of this work, limited effort has been made to handle the data-level domain gap in the context of V2X collaborative perception. As depicted in Fig. 1(a), the vehicles and infrastructures might have different types of LiDAR sensors, thereby directly fusing the point cloud data or intermediate features from various sources could inevitably corrupt the final performance. Therefore, how to explicitly learn domain-invariant representation from multi-source data in this context remains to be explored.

In this work, we propose a *D*omain-*I*nvariant *V2X* model, *DI-V2X*, for collaborative 3D object detection in autonomous driving scenarios. To achieve this, a new teacher-student distillation model is introduced in *DI-V2X*. During training, we enforce the student models, *i.e.*, the vehicles and infrastructures, to learn a domain-invariant representation aligned with that of the early-fused teacher model, *i.e.*, the point clouds from multiple viewpoints are integrated into a holistic view to train the teacher. During inference, only the student models are retained. To be specific, *DI-V2X* consists of three necessary components: a domain-mixing instance augmentation (DMA) module, a progressive domain-invariant distillation (PDD) module, and a domain-adaptive fusion (DAF) module. The purpose of DMA is to build a mixed ground-truth instance bank during training to align the point cloud inputs of the teacher and students, where instances are drawn from vehicle-side, road-side, or fused-side. After that, PDD aims to progressively transfer the information from the teacher to students across different stages, *i.e.*, both before and after domain-adaptive fusion. For example, before fusion, students are guided to learn the domain-invariant representation separately in non-overlapping areas. While after fusion, we focus on the distillation within the overlapping area since the information has been well aggregated. In the DAF module, features from different domains are adaptively fused, based on their spatial significance. Additionally, DAF enhances the model's resilience to pose errors by integrating calibration offsets, ensuring robust V2X detection performance.

As illustrated in Fig. 1(b), we test different solutions that explicitly tackle the domain gap. 'PP-LF', 'PP-IF', and 'PP-EF' indicate that we apply late fusion, intermediate and early fusion based on PointPillar (PP) (Lang et al. 2019). Knowledge distillation (KD) (Li et al. 2021) is an intuitive solution that adopts an early-fused teacher to guide the intermediate fusion process, while CDT refers to the cross-domain transformer in MPDA (Xu et al. 2023b) to mitigate domain gap. All these solutions yield sub-optimal results compared to the proposed *DI-V2X* model.

To the best of our knowledge, *DI-V2X* makes the early effort to bridge the data-level domain gap caused by sensor differences in the context of V2X collaborative perception, which is achieved by incorporating a domain-invariant distillation framework. The proposed modules PDD, DAF, and DMA prove highly effective in aligning, merging, and augmenting feature representations, ensuring the model acquires domain-invariant capabilities. Furthermore, we demonstrate the leading performance of

DI-V2X through extensive experiments on the challenging DAIR-V2X (Yu et al. 2022) and V2XSet (Xu et al. 2022a) benchmark datasets. It significantly outperforms previous state-of-the-art (SOTA) models like CoAlign (Lu et al. 2023) and V2X-ViT (Xu et al. 2022a) by 5.76% AP@0.7 and 11.5% AP@0.7, respectively.

2   Related Works

V2X Collaborative Perception

V2X collaborative perception approaches can be roughly divided into three categories: early fusion, intermediate fusion, and late fusion. Early fusion models (Chen et al. 2019b; Yu et al. 2022) integrate the information from various agents at the raw data level, *i.e.*, aggregating the 3D point cloud from multiple viewpoints to yield a holistic view of the scene. However, transmitting the raw 3D data between agents inevitably consumes more communication bandwidth, which hinders its application in real-world scenarios. late fusion methods (Xu et al. 2023a; Yu et al. 2023), on the other hand, performs fusion at the output level, *i.e.*, the 3D bounding boxes. While these approaches can decrease the communication cost by only transmitting the perception results, the performance is susceptible due to the presence of noisy and incomplete outputs from each agent, as well as pose errors between agents. To achieve a performance-bandwidth trade-off, intermediate-fusion methods (Liu et al. 2020b,a; Chen et al. 2019a) are becoming popular as they conduct fusion at the feature level. They only transmit compressed feature maps that contain crucial information across agents, thereby significantly saving the communication cost while also remaining robust to pose errors in contrast to early fusion and late fusion models.

In this paper, we mainly investigate the intermediate-fusion approaches. V2VNet (Wang et al. 2020) adopts a graph neural network to model the agent relation and aggregate messages from different vehicles. OPV2V (Xu et al. 2022b) utilizes self-attention to learn interactions between features. DiscoNet (Li et al. 2021) leverages an early-fused teacher to guide the cooperation across different student agents. Where2comm (Hu et al. 2022) focuses on the compression of feature maps, which only keep spatially sparse yet perceptually critical information. CoAlgin (Lu et al. 2023) proposes an agent-object graph optimization method to handle the pose error. V2X-ViT (Xu et al. 2022a) uses a transformer-based structure and a heterogeneous multi-agent self-attention mechanism to facilitate feature fusion.

All these approaches focus on enhancing interactions between the features from different agents or reducing communication costs while paying less attention to the inherent domain gap. To this end, we introduce *DI-V2X* to explicitly learn domain-invariant representation to overcome the data-level domain disparities, which is achieved through a progressive distillation framework.

Domain-Invariant Representation in Point Cloud

In the context of point cloud-based perception, domain gaps may arise from variations in sensor types, weather conditions, and geographical factors. Domain-invariant learn-
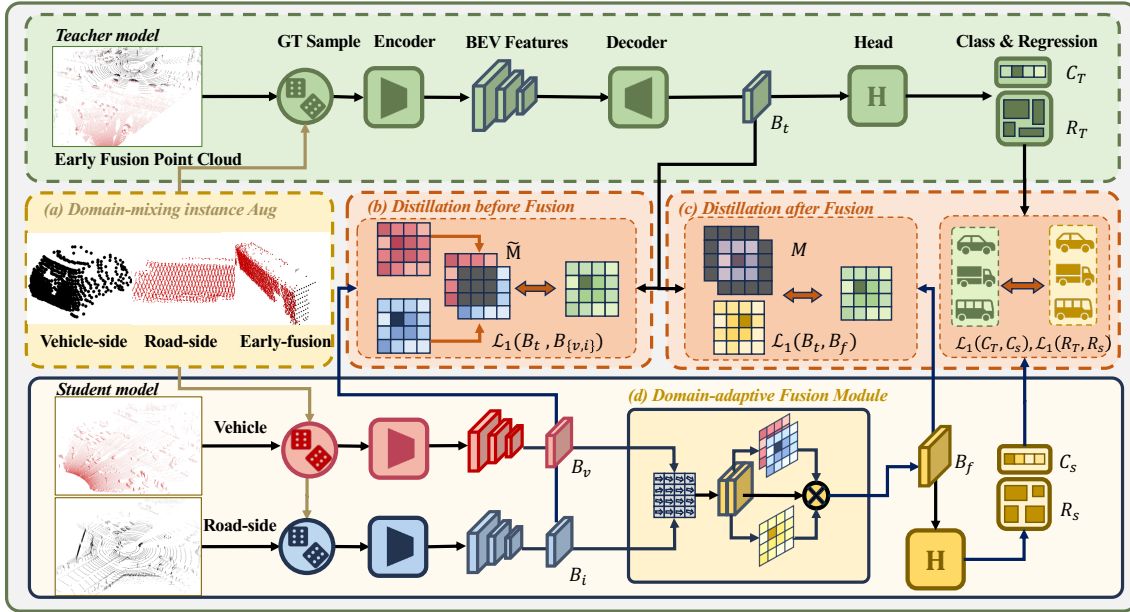
Figure 2: Overall architecture of *DI-V2X*. It consists of a teacher model that takes the holistic-view point cloud as input, along with two student models utilizing point clouds from different domains, *e.g.*, the vehicles and infrastructures. (a). A Domain-Mixing instance Augmentation (DMA) module is used to generate aligned data representation. (b-c). A Progressive Domain-invariant Distillation (PDD) module then aligns the students with the teacher to learn domain-invariant features, both before and after fusion with the guidance of overlap information. (d). A Domain-Adaptive Fusion (DAF) module further enhances the feature with domain- and spatially-adaptive attention. The components within the dashed boxes are only used during training.

ing (Triess et al. 2021) is a promising approach to ensure feature generalization in V2X perception tasks. There are typically two approaches to achieve this: constructing domain-invariant data representation and conducting domain-invariant feature learning. To attain domain-invariant data, some works tend to apply data augmentation strategies (Chen et al. 2020; Lee et al. 2021; Nekrasov et al. 2021) to transfer the multi-source data into a unified domain. LiDomAug (Ryu, Hwang, and Park 2023) proposes a sophisticated framework to generate a dense point cloud scene with sequential frames to adapt to random configurations and reduce sensor bias. Other works apply scene-level augmentation (Alonso et al. 2021) or re-sample the LiDAR scans to match the data distribution of target domain (Langer et al. 2020). On the other hand, some works focus on domain-invariant representation learning to create representations that capture underlying patterns and characteristics common across different domains.

Most of the aforementioned approaches target a domain adaptation task that seeks to transfer the knowledge from a source domain to a target domain. In contrast, we focus on acquiring a generalized representation capable of unifying point clouds from diverse domains (*e.g.,* vehicle-side and road-side) for improved performance in V2X perception.

3 The Proposed *DI-V2X* Framework

In typical driving scenarios with V2X collaborative perception, vehicles and infrastructures are commonly equipped with different types of LiDAR sensors, which will result in

data-level domain gap. In this work, we aim to tackle the V2X collaborative 3D object detection task from the perspective of domain-invariant feature learning. In this section, we will introduce the overall framework. Next, we will delve into the details of each module.

Overall Architecture

As illustrated in Fig. 2, *DI-V2X* adopts a teacher-student distillation framework that aims to learn domain-invariant representations for the student models. Formally, let $\mathbf{P}_v \in \mathbb{R}^{N_v \times 4}$ and $\mathbf{P}_i \in \mathbb{R}^{N_i \times 4}$ denote the vehicle-side and road-side point clouds which have been temporally synchronized. We first fuse the two point clouds from distinct viewpoints to attain a holistic viewpoint cloud $\mathbf{P}_e$ by leveraging their pose information. Then, $\mathbf{P}_e$ is utilized to train the teacher.

To be specific, we employ the Domain-Mixing instance Augmentation (DMA) module to get augmented inputs $\{\mathbf{P}'_v, \mathbf{P}'_i, \mathbf{P}'_e\} = Aug(\{\mathbf{P}_v, \mathbf{P}_i, \mathbf{P}_e\})$. The perception model is instantiated by a PointPillars (Lang et al. 2019) 3D detector here. We discretize the $\mathbf{P}'_e$ into voxels and employ a 3D encoder (*e.g.*, VoxelNet (Yan, Mao, and Li 2018)) to extract bird-eye-view (BEV) feature maps denoted as $\mathbf{B}_t \in \mathbb{R}^{H \times W \times C}$. Finally, a detection head predicts the class and regression results $\mathbf{C}_T = \{\mathbf{c}_k\}_{k=1}^K, \mathbf{R}_T = \{\mathbf{r}_k\}_{k=1}^K$. $\mathbf{r}_k = (x, y, z, h, w, l, \theta)$, where (x, y, z) and (h, w, k) represent the center coordinate and size of an object and θ indicates the orientation along the X-Y plane.

The student models share the most designs with the

teacher model, *e.g.*, voxel configuration, 3D encoder, and detection head. The augmented point clouds $\{\mathbf{P}'_v, \mathbf{P}'_i\}$ are sent to a share-weight 3D encoder to extract BEV features $\mathbf{B}_v, \mathbf{B}_i \in \mathbb{R}^{H \times W \times C}$, respectively. Then, a Domain-adaptive Fusion (DAF) module is presented to fuse the feature maps from different domains to get $\mathbf{B}_f \in \mathbb{R}^{H \times W \times C}$. During training, the proposed Progressive Domain-invariant Distillation (PDD) module aligns various student features ($\mathbf{B}_v, \mathbf{B}_i$ and $\mathbf{B}_f$) with the teacher BEV feature ($\mathbf{B}_t$) both before and after fusion, which is guided by the overlapping and non-overlapping mask $\mathbf{M}, \tilde{\mathbf{M}} \in \{0,1\}^{H \times W}$. In this way, the students can learn a domain-invariant representation that is adaptable for multi-source data in the V2X context. After training, only the student is employed for inference

Domain-Mixing Instance Augmentation

For 3D perception in autonomous driving scenarios, instances are more informative than the noisy background. Therefore, we propose to augment the input point cloud with domain-mixing instances, aligning the data distribution between the teacher and students. This also facilitates the subsequent distillation process. Specifically, let $\mathbf{P}_v \in \mathbb{R}^{N_v \times 4}, \mathbf{P}_i \in \mathbb{R}^{N_i \times 4}$ be the point clouds from vehicle and infrastructure. We first project $\mathbf{P}_i$ to the coordinate system of ego-vehicle, such that $\mathbf{P}_i^T = \mathbf{T}_{(i \rightarrow v)} \mathbf{P}_i^T$, where $\mathbf{T}_{(i \rightarrow v)} \in \mathbb{R}^{4 \times 4}$ is the transformation matrix from infrastructure to vehicle system. Then we leverage the ground-truth bounding box $\mathbf{B}_{gt} = \{\mathbf{b}_k\}$ to get instances from $\mathbf{P}_v$ and $\mathbf{P}_i$. Instance from different domains corresponding to the same ground-truth object will be merged to get an early fusion instance $\mathbf{p}_k = \text{Concat}(\mathbf{p}_k^v, \mathbf{p}_k^i) \in \mathbb{R}^{N_k \times 4}$, where $\mathbf{p}_k^v \in \mathbf{P}_v$ and $\mathbf{p}_k^i \in \mathbf{P}_i$ are instance points indexed by $\mathbf{b}_k$ from both domains. As relative positions between agents change dynamically with the motion of the ego-vehicle, some instances may only come from individual domains, while some others may come right from the overlapping area that has been early fused. To determine the domain source of each instance, we divide these instances into three categories by computing the point proportion from each domain:

$$
\begin{aligned}
D_i &= \{\mathbf{p}_k | N_k^v / (N_k^v + N_k^i) < \tau_l\} \\
D_f &= \{\mathbf{p}_k | \tau_l < N_k^v / (N_k^v + N_k^i) < \tau_h\} \\
D_v &= \{\mathbf{p}_k | N_k^v / (N_k^v + N_k^i) > \tau_h\}
\end{aligned}
\tag{1}
$$

where N_k^v and N_k^i represent the point number from the vehicle side and infrastructure side, respectively, and τ_l, τ_h indicate the thresholds. Then, we can get an instance bank $D_{mixed} = D_i \cup D_f \cup D_v$, which contains mixed instances from all the domains, *i.e.*, including the fused domain. During training, we randomly sample instances from D_{mixed} according to certain probability and add these instances to both the teacher and students.

By involving instances from different domains, we enhance the diversity of the training data. Furthermore, from each student's perspective, information from other domains is incorporated through instance-level mixup (Zhang et al. 2018). This approach essentially aligns the data distributions between the teacher and student models, thereby resulting in

more generalized features after the subsequent knowledge distillation process.

Progressive Domain-Invariant Distillation

In this section, we focus on acquiring domain-invariant features through progressive domain-invariant distillation. To achieve this, We adopt a two-stage distillation strategy, *i.e.*, distillation before and after the domain-adaptive fusion (DAF) module. The first distillation stage is to align the students' distributions with the teacher model as inputs for DAF, which is essential for accurate information fusion. However, we empirically found that directly conducting distillation over the whole feature map between students and the teacher yields suboptimal performance. To address this, we opt for distillation in the non-overlapping area during the first stage. In the second stage, since the student features have been well fused by DAF, we can concentrate on distillation over the overlapping area. This two-stage distillation process enforces the student models to match the feature representations of the teacher model from different regions, reducing the gap between students.

Distillation Before Fusion. To perform the first-stage distillation, we need to first compute the overlapping mask to determine the overlapping area. We present the vehicle perception region as a rectangle $\mathbf{A}_v = (x_v, y_v, 2R_x, 2R_y, \theta_v)$, where R_x, R_y are the max range along X-Y-plane and θ denotes the rotation. All the computations are unified under the vehicle coordinate system, thus $(x_v, y_v) = (0, 0)$ and $\theta_v = 0$. The perception region of the infrastructure side is then transformed to the vehicle side, and get a new rectangle $\mathbf{A}_i = (x_i, y_i, 2R_x, 2R_y, \theta_i)$. Then we can compute the overlapping area between A_v and A_i, which is $\mathbf{P}_{overlap} = Intersection(\mathbf{A}_v, \mathbf{A}_i)$. The obtained $\mathbf{P}_{overlap}$ (*i.e.*, a polygon) is then downsampled to match the size of feature map $\mathbf{B}_v \in \mathbb{R}^{H \times W \times C}$. The overlapping mask $\mathbf{M} \in \{0,1\}^{H \times W}$ is then defined by:

$$
\mathbf{M}(i, j) = \left\{ \begin{array}{lc} 1 & , \quad \text{if } (i,j) \in \mathbf{P}_{overlap} \\ 0 & , \quad \quad \text{otherwise} \end{array} \right.
\tag{2}
$$

where $\mathbf{M}(i, j) \in 0, 1$ denotes the binary value at the coordinate of (i, j). By focusing distillation only on the non-overlapping area $\tilde{\mathbf{M}}$, we allow each student to concentrate on learning representations that align their respective domains. This avoids strict constraint that enforces incomplete student features to learn towards complete features from the teacher. The distillation loss is then expressed by:

$$
\begin{aligned}
\mathcal{L}_{da} &= \mathcal{L}_1(\mathbf{B}_t, \mathbf{B}_v \odot \tilde{\mathbf{M}}_v) + \mathcal{L}_1(\mathbf{B}_t, \mathbf{B}_i \odot \tilde{\mathbf{M}}_i) \\
&= \frac{1}{HW} \sum_m^H \sum_m^W |\mathbf{B}_t(m, n) - \mathbf{B}_v(m, n)| \times \tilde{\mathbf{M}}_v(m, n) \\
&+ \frac{1}{HW} \sum_m^H \sum_n^W |\mathbf{B}_t(m, n) - \mathbf{B}_i(m, n)| \times \tilde{\mathbf{M}}_i(m, n)
\end{aligned}
\tag{3}
$$

where $\mathcal{L}_1$ is the L1 loss and $\odot$ denotes the element-wise multiplication with channel-wise broadcasting. $\tilde{M}_v, \tilde{M}_i$ are the non-overlapping mask for vehicle and infrastructure.

Distillation After Fusion. In this phase, the student features from various domains have been effectively merged using the DAF module. As a result, we obtain a highly capable fused representation denoted as $\mathbf{B}_f$, which can be aligned with the teacher's feature representation $\mathbf{B}_t$. Intuitively, $\mathbf{B}_t$ is obtained through the early collaboration of mixed point cloud data, which inherently involves minimal information loss. By enforcing the intermediate fused features $\mathbf{B}_f$ to progressively align with $\mathbf{B}_t$, we effectively ensure that the essential knowledge gained through the early fusion stage is consistently integrated throughout the learning process, resulting in the domain-invariant feature representations. Formally, the second-stage distillation loss can be denoted as:

$$
\begin{aligned}
\mathcal{L}_f &= \mathcal{L}_1(\mathbf{B}_t, \mathbf{B}_f \odot \mathbf{M}_v) \\
&= \frac{1}{HW} \sum_m^H \sum_m^W |\mathbf{B}_t(m,n) - \mathbf{B}_f(m,n)| \times \tilde{\mathbf{M}}_v(m,n)
\end{aligned}
\tag{4}
$$

Moreover, we could go beyond the feature level alignment and extend to the prediction level. Since our final target is to decode the final 3D bounding boxes from both $\mathbf{B}_f$, ensuring alignment at the prediction level further contributes to the consistency and accuracy of the results. The prediction-level distillation can be formulated as:

$$
\begin{aligned}
\mathcal{L}_p &= \mathcal{L}_{class} + \mathcal{L}_{regression} \\
&= \frac{1}{K} \sum_k^K (|\mathbf{c}_k - \mathbf{c}_k^s| + |\mathbf{r}_k - \mathbf{r}_k^s|)
\end{aligned}
\tag{5}
$$

where K is the number of predictions, $\mathbf{C}_T = \{\mathbf{c}_k\}_{k=1}^K, \mathbf{R}_T = \{\mathbf{r}_k\}_{k=1}^K$ indicate the class and regression prediction of teacher model and $\mathbf{C}_S = \{\mathbf{c}_k^s\}_{k=1}^K, \mathbf{R}_S = \{\mathbf{r}_k^s\}_{k=1}^K$ refer to the class and regression prediction of student model after feature fusion. During training, the overall loss function is

$$
\mathcal{L} = \mathcal{L}_{detect} + \lambda_{kd}(\mathcal{L}_{da} + \mathcal{L}_f + \mathcal{L}_p)
\tag{6}
$$

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

Domain-Adaptive Fusion

The target of the DAF module is to aggregate the features from both vehicles and infrastructure, creating an enhanced representation that contains valuable information from each domain. However, this fusion process comes with two main challenges: potential misalignment of the pose between two sides and the design of a suitable feature interaction strategy. The relative pose of vehicles and infrastructure in the real world is vulnerable due to facts such as sensor noise, dynamic motion, and inconsistency in varying timestamps, which will impact the accuracy of V2X perception. To address this, we proposed to leverage the calibration offset to dynamically correct the potential pose errors. Regarding feature collaboration, our approach emphasizes essential domain-specific information across spatial locations.

The overview of DAF is illustrated in Fig 3. We first predict the calibration offset with a convolution layer to better align $\mathbf{B}_i$ with $\mathbf{B}_v$, which is denoted as:

$$
\mathbf{\Delta}_{(i \to v)} = \text{Conv}(\text{Concat}(\mathbf{B}_v, \mathbf{B}_i)) \in \mathbb{R}^{H \times W \times 2}.
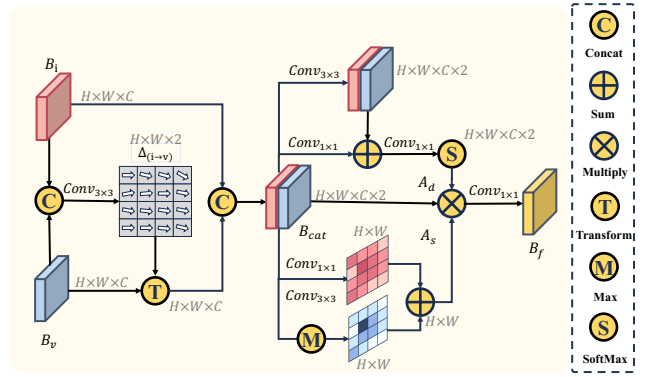\tag{7}
$$

Figure 3: Illustration of the Domain-Adaptive Fusion module. Features from different domains are first aligned by a collaboration offset to reduce pose errors. Then, the aligned features are further refined by domain-adaptive and spatially-adaptive attention mechanisms.

Then we can use this offset to transform the infrastructure feature $\mathbf{B}_i$ to align with the vehicle feature $\mathbf{B}_v$.

$$
\mathbf{B}_i^{'}(p_k) = \mathbf{B}_i(p_k + \mathbf{\Delta}_{(i \to v)}(p_k)), 0 \le k < HW
\tag{8}
$$

where $p_k \in \mathbb{R}^2$ indicates the position of k-th grid and $\mathbf{B}(p)$ means the feature value at position p. This operation enables the dynamic integration of information from neighboring points, thereby mitigating potential pose errors.

Furthermore, we aim to extract both domain-adaptive and spatially-adaptive features from $\mathbf{B}_i$ and $\mathbf{B}_v$. This allows the model to leverage both the unique information present in each domain and the shared spatial context, enhancing the overall representation. Mathematically, we first aggregate feature maps by concatenation operation to get $\mathbf{B}_{cat} = \text{Concat}(\mathbf{B}_v, \mathbf{B}i^{'}) \in \mathbb{R}^{H \times W \times C \times 2}$. Domain-adaptive attention can be formulated as:

$$
\mathbf{A}_d = \text{Softmax}(\text{Conv}(\mathbf{B}_{cat})) \in \mathbb{R}^{H \times W \times C \times 2}
\tag{9}
$$

where $\text{Conv}(\cdot)$ is implemented by a 3×3 convolution layer and two 1×1 convolution layers with residual structure. $\text{Softmax}(\cdot)$ is used to generate the weight for each domain. The domain-adaptive attention can dynamically adjust the weight of each domain in each region, resulting in a more flexible collaboration. Spatial attention is also crucial because it forces the network to focus on the areas with high objectness scores or areas that can provide essential context information. Here, we use convolution of various kernel sizes to aggregate features with different receptive fields and a $Maximum$ function to select important areas. The spatially-adaptive attention can be formulated as:

$$
\mathbf{A}_s = \text{Conv}(\mathbf{B}_{cat}) + \max(\mathbf{B}_{cat})
\tag{10}
$$

where $\mathbf{A}_s \in \mathbb{R}^{H \times W}$, $\max(\cdot)$ is the maximum value along the channel and domain dimension and $\text{Conv}(\cdot)$ is two convolution layers with different kernel size. Finally, we get the fused feature considering both domain and spatial attention:

$$
\mathbf{B}_f = \text{Conv}(\mathbf{A}_d \odot \mathbf{A}_s \cdot \mathbf{B}_{cat}) \in \mathbb{R}^{H \times W \times C}
\tag{11}
$$

Fusion	Model	Publication	BEV AP@0.5	BEV AP@0.7
No Fusion	PP (Lang et al. 2019)	CVPR 2019	65.21	54.26
Late Fusion	PP-LF	CVPR2019	63.68	45.33
Early Fusion	PP-EF	CVPR2019	76.55	64.10
	DI-V2X (teacher)	-	**79.81**	**67.08**
	PP-IF	CVPR 2019	72.85	52.67
	V2VNet (Wang et al. 2020)	ECCV 2020	70.97	47.63
	DiscoNet (Li et al. 2021)	NeurIPS 2021	73.87	59.61
Intermediate Fusion	OPV2V (Xu et al. 2022b)	ICRA 2022	73.30	55.33
	V2X-ViT (Xu et al. 2022a)	ECCV 2022	71.81	54.94
	Where2comm (Hu et al. 2022)	NeurIPS 2022	74.09	59.66
	CoAlign (Lu et al. 2023)	ICRA 2023	<u>74.60</u>	<u>60.40</u>
	DI-V2X (student)	-	**78.82**(+4.22)	**66.16**(+5.76)

Table 1: Comparison with state-of-the-art methods on DAIR-V2X val dataset. We apply PointPillars (denoted as PP) as the detection model to benchmark different fusion approaches (*i.e.*, no/early/intermediate/late fusion). *DI-V2X* (teacher) refers to the early fused teacher trained with DMA strategy, while *DI-V2X* (student) is the intermediate fusion model. The best and the second-best approaches in each column are indicated by bold and underlined formatting.

Fusion	Model	AP@0.5	AP@0.7
No Fusion		60.6	40.2
Late Fusion	PP	72.7	62.0
Early Fusion		81.9	71.9
	PP-IF	76.9	49.2
	DiscoNet	84.4	69.5
Intermediate	Where2comm	85.5	65.4
	V2X-ViT	<u>88.2</u>	<u>71.2</u>
	DI-V2X	**92.7**	**82.7**

Table 2: Performance comparasion on V2XSet dataset.

where $\odot$ means element-wise multiplication with channel wise broadcast and $\cdot$ is the element-wise multiplication. Thus, spatially-adaptive attention can provide a robust attention map by aggregating multi-granularity features.

4 Experiments

Dataset

DAIR-V2X. We employ the challenging DAIR-V2X (Yu et al. 2022) for evaluating our model and other SOTA approaches. DAIR-V2X is a real-world V2X collaborative perception dataset, which is collected by well-equipped vehicles and infrastructures. The road-side infrastructures are equipped with 300-beam solid-state LiDAR with $100°$ horizontal FOV, while the vehicle uses 40-beam mechanical LiDAR with $360°$ horizontal FOV. Totally 9K synchronized vehicle and infrastructure LiDAR frame pairs are sampled from 100 representative scenes at 10Hz. DAIR-V2X provides 3D object detection labels within the camera's field of view. We utilize these comprehensive 3D labels that encompass a 360-degree detection range, as mentioned by prior works (Lu et al. 2023).

V2XSet. We also evaluate our method on another V2X calibration dataset V2XSet (Xu et al. 2022a), which contains 6694 training data and 1920 validation data generated by the simulator. Each scene contains point clouds collected from at least 2 and at most 7 agents. These agents are equipped with 36-beam LIDAR with $360°$ horizontal FOV.

Experimental Setup

Evaluation Metrics. V2X collaborative 3D object detection aims to achieve both high detection performance and minimal communication costs. Our focus is on enhancing the detection performance without increasing the communication bandwidth compared with other intermediate fusion approaches. AP (average precision) is employed to evaluate the detection performance in BEV at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7.

Implementation Details. We set the point cloud range to $[-100, 100] \times [-40, 40] \times [-3.5, 1.5]$ meters defined in the vehicle coordinate system with the voxel size as $[0.4, 0.4, 5]$ meters along XYZ axes. For scene-level data augmentation, we adopt random flip, rotation, and scaling. The proposed DMA randomly samples the instances from the database with the probability of 0.8, 0.1, and 0.1 for D_f, D_v, D_i. We adopt the PointPillars (Lang et al. 2019) as the detector to align with other V2X approaches (Lu et al. 2023), while our approach is versatile to any BEV-based 3D object detectors. The feature map after fusion (*i.e.*,B_i, B_v) is of size $252 \times 100 \times 256$. We set λ_{kd} as 1 and the thresholds τ_l, τ_h as 0.2 and 0.8. All student models are trained on 4 NVIDIA Tesla V100 GPUs with a batch size of 4 for 40 epochs.

Main Results

Performance Benchmarking. Table 1 presents a comparison of various V2X approaches, including early, intermediate, and late fusion methods. PP refers to PointPillars cotrained with mixed data of vehicle and infrastructure. As presented in Table 1, we can observe that the late fusion method PP-LF, which directly combines the detection results from two sides using non-maximum suppression (NMS) even results in a performance decrease of 1.53% in terms of AP@0.5, probably due to the pose error. The naive intermediate fusion method PP-IF, which is obtained by directly summing the features from both domains, improves the performance by 7.64% in terms of AP@0.5. Furthermore, the early fusion method PP-EF surpasses other fusion types by a large margin. By utilizing domain-mixing augmentation,

Model	Train	Test			Avg.
		Veh.	Inf.	V. & I.	AP@0.7
PP	Veh.	73.08	31.53	52.62	52.41
	Inf.	32.21	**65.80**	26.39	41.46
PP-LF		63.62	65.66	45.33	58.20
PP-EF		73.19	47.44	64.10	61.57
PP-IF	V. & I.	68.55	7.45	52.67	26.22
PP-KD		69.91	17.82	53.91	47.11
PP-CDT		66.22	41.73	55.90	54.61
Ours-T	V. & I.	**76.74**	53.34	**67.08**	**65.72**
Ours-S		73.04	45.91	66.16	61.70

Table 3: Domain generalization ability of different solutions on DAIR-V2X val set. 'Veh.' and 'Inf.' indicate the vehicle and infrastructure domain, and 'V. & I.' means using data from both sides. 'Avg.' is the average AP.

our *DI-V2X* (teacher) achieves the highest performance in terms of both BEV AP@0.5 and AP@0.7.

Additionally, we conduct a comparison between our *DI-V2X* and some competitive intermediate fusion methods. The results of OPV2V and CoAlign are borrowed from (Lu et al. 2023). Besides, we reimplement V2VNet, V2X-ViT and DiscoNet and achieve slightly better performance than the results reported in CoAlign. We can see that *DI-V2X* (student) surpasses all these intermediate fusion methods. For example, it outperforms the previously best-performing model CoAlign by 4.22% AP@0.5. This demonstrates the importance of domain-invariant learning.

In Table 2, we present performance comparisons on another challenging dataset, V2XSet (Xu et al. 2022a). We access different fusion strategies and compare our model with SOTA models like DiscoNet, Where2comm, and V2X-ViT. The results indicate that *DI-V2X* outperforms these approaches significantly on the V2XSet validation set, demonstrating the generalizability of our model.

Domain Generalization Ability of *DI-V2X*. As our method aims to learn domain-invariant representation, we analyze the effectiveness of *DI-V2X* in improving the model generalization ability in this section. To this end, we evaluate the transfer learning results (*i.e.*, training on Veh. & Inf. and testing on Veh., Inf, or Veh. & Inf.). We compare with some common solutions that mitigate domain gap, *e.g.*, PP-LF, PP-EF, PP-IF, PP-KD, and PP-CDT. PP-KD adopts an early-fusion teacher to guide the collaboration through knowledge distillation while PP-CDT leverages a cross-domain transformer to align the domain distribution (Xu et al. 2023b). In cases of communication failure, vehicles may not receive signals from roadside infrastructure, *i.e.*, testing only with vehicle data. Here, our model achieves competitive results with a model trained with only the vehicle data. Among the intermediate fusion approaches, our model also gives the best performance when testing only with the infrastructure data. When data from both sides are available, our model obtains the best performance. In essence, our model attains the best average accuracy across various cases, showing its superior generalization capability.

Model	DMA	PDD	DAF	AP@0.5	AP@0.7
PP-IF	-	-	-	72.85	52.67
DG-V2X	✓	-	-	73.90	54.02
	-	✓		74.17	55.89
	-	-	✓	75.76	60.89
	-	✓	✓	77.37	61.66
	✓	✓	✓	**78.82**	**66.16**

Table 4: Ablation studies of *DI-V2X* to verify each proposed module on the DAIR-V2X val set. Our final model surpasses the baseline fusion model, *i.e.*, PP-IF, by 5.97% and 13.49% in terms of AP@0.5 and AP@0.7, respectively.

Ablation Study. We investigate the contribution of each proposed module in this section, where an intermediate fusion model PP-IF is used as the baseline. PP-IF refers to PointPillars trained by summing features from both domains. As illustrated in Table 4, the three proposed modules are all beneficial to the final performance. To be specific, DMA improves the AP@0.7 by 1.35% and AP@0.5 by 1.05%, which indicates that domain-mixing instances sampling for both agents enriches the data diversity and aligns the distribution. DAF improves the AP@0.7 by 8.22% and AP@0.5 by 2.91% by implicitly reducing the sensor noise of the two agents and enriching their feature interaction. To evaluate PDD, we equip PP-IF with an early fused teacher model and distill features both before and after fusion guided by an overlapping mask and improve the AP@0.7 by 3.22% and AP@0.5 1.32% The performance can be further boosted when applying three modules, which proves the effectiveness of *DI-V2X*.

5 Conclusion

In this work, we presented a novel approach, *DI-V2X*, to address a critical challenge in V2X collaborative perception, *i.e.*, data-level domain gap due to sensor differences across agents. The core is to leverage a domain-invariant distillation framework to enhance the generalization of the feature representation. To achieve that, a Domain-Mixing instance Augmentation (DMA) strategy is proposed to ensure the alignment of data distributions between teacher and student models. Then, a Progressive Domain-invariant Distillation (PDD) module leverages overlap information shared between agents to facilitate the distillation both before and after domain fusion. The Domain-Adaptive Fusion (DAF) module further aggregates features from each domain by considering both domain-adaptive and spatially relevant information. It also accounts for potential pose errors by incorporating calibration offsets. Experimental results conducted on the challenging datasets DAIR-V2X and V2XSet demonstrate the remarkable performance of our approach.

6 Acknowledgments

References

Alonso, I.; Riazuelo, L.; Montesano, L.; and Murillo, A. C. 2021. Domain Adaptation in LiDAR Semantic Segmentation by Aligning Class Distributions. arXiv:2010.12239.

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*.

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*.

Chen, Y.; Hu, V. T.; Gavves, E.; Mensink, T.; and Mettes. 2020. PointMixup: Augmentation for Point Clouds. In *ECCV*.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *NeurIPS*.

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 *CVPR*.

Langer, F.; Milioto, A.; Haag, A.; Behley, J.; and Stachniss, C. 2020. Domain Transfer for Semantic Segmentation of LiDAR Data using Deep Neural Networks. In *IROS*.

Lee, D.; Lee, J.; Lee, J.; Lee, H.; Lee, M.; Woo, S.; and Lee, S. 2021. Regularization Strategy for Point Cloud via Rigidly Mixed Sample. In *CVPR*.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning Distilled Collaboration Graph for Multi-Agent Perception. In *NeurIPS*.

Liu, Y.-C.; Tian, J.; Glaser, N.; and Kira, Z. 2020a. When2com: Multi-Agent Perception via Communication Graph Grouping. In *CVPR*.

Liu, Y.-C.; Tian, J.; Ma, C.-Y.; Glaser, N.; Kuo, C.-W.; and Kira, Z. 2020b. Who2com: Collaborative Perception via Learnable Handshake Communication. In *ICRA*.

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 *ICRA*.

Nekrasov, A.; Schult, J.; Litany, O.; Leibe, B.; and Engelmann, F. 2021. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In *2021 International Conference on 3D Vision (3DV)*.

Ryu, K.; Hwang, S.; and Park, J. 2023. Instant Domain Augmentation for LiDAR Semantic Segmentation. In *CVPR*.

Triess, L. T.; Dreissig, M.; Rist, C. B.; and Marius Zöllner, J. 2021. A Survey on Deep Domain Adaptation for LiDAR Perception. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*.

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 *ECCV*.

Xu, R.; Chen, W.; Xiang, H.; Xia, X.; Liu, L.; and Ma, J. 2023a. Model-agnostic multi-agent perception framework. In *ICRA*.

Xu, R.; Li, J.; Dong, X.; Yu, H.; and Ma, J. 2023b. Bridging the domain gap for multi-agent perception. In *ICRA*.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022a. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022b. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *ICRA*.

Yan, Y.; Mao, Y.; and Li, B. 2018. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18(10): 3337.

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 *CVPR*.

Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N.; Song, J.; Yuan, J.; Luo, P.; and Nie, Z. 2023. V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *CVPR*.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.