TransIFF: An Instance-Level Feature Fusion Framework for Vehicle-Infrastructure Cooperative 3D Detection with Transformers

Ziming Chen
Beihang University
chenzm@buaa.edu.cn

Yifeng Shi*
Baidu Inc.
shiyifeng@baidu.com

Jinrang Jia
Baidu Inc.
jjr5401@163.com

Abstract

Cooperation between vehicles and infrastructure is vital to enhancing the safety of autonomous driving. Two significant and contradictory challenges now stand in the collaborative perception: fusion accuracy and communication bandwidth. Previous intermediate fusion methods that transmit features balance the accuracy and bandwidth compared with early fusion and late fusion, but usually have problems with feature alignment and domain gaps, and the bandwidth usage still falls short of the industrial application standard to our best knowledge.

In this paper, we propose TransIFF, an instance-level feature fusion framework with transformers that can effectively reduce bandwidth usage. Furthermore, it can align the domain gaps between vehicle and infrastructure features, and improve the robustness of feature fusion, leading to a high cooperative perception accuracy. TransIFF is composed of three components: a vehicle-side network, an infrastructure-side network, and a vehicle-infrastructure fusion network. Initially, the vehicle-side and infrastructure-side networks independently generate instance-level features. Subsequently, the infrastructure-side instance-level features are transmitted to the vehicles, significantly reducing the communication bandwidth usage. Finally, in the vehicle-infrastructure fusion network, Cross-Domain Adaptation (CDA) module is designed to align the feature domains, followed by Feature Magnet (FM) module which can adaptively fuse the instance features and achieve a robust feature fusion. TransIFF yields state-of-the-art performance on the widely used real-world vehicle-infrastructure cooperative benchmark DAIR-V2X, achieving 59.62% AP with only 2^{12} bytes bandwidth consumption.

1. Introduction

Accurate environment perception [14, 15, 40] is an important topic in the field of autonomous driving. Currently, autonomous vehicles mainly rely on the on-board LiDAR sensors [15, 31, 32] to obtain dense point clouds of the surrounding environment and perform target detection. However, limited by the height of the sensor installation, the vehicle perception system faces challenges such as occlusion-induced blind spots and long-distance perception instability, which hinder the development of autonomous driving. In recent years, there has been an increasing amount of research focused on utilizing both vehicle-side and infrastructure-side information to achieve Vehicle-to-Everything (V2X) cooperative perception [22, 36, 37, 8] to address these issues. Thanks to the high installation height of infrastructure-side sensors, autonomous vehicles can achieve a global perspective and long-distance perception by receiving information from infrastructure-side sensors, significantly improving the perception ability.

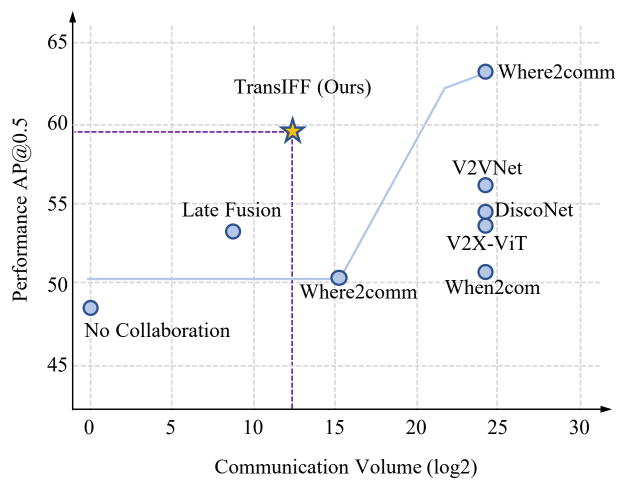

Figure 1. **TransIFF achieves superior performance-bandwidth trade-off on DAIR-V2X benchmark.** Compared with other intermediate fusion methods, TransIFF can achieve over 2^{12} times less communication volume, and still have an AP of nearly 60%. The curve of Where2comm means its performance at different bandwidth usage.

*Corresponding author

One of the challenges of vehicle-infrastructure collaborative perception is that information needs to be sent from infrastructure-side equipment to autonomous vehicles, which requires communication bandwidth resources. However, industrial communication systems can hardly afford huge communication consumption in real-time [42, 13]. Therefore, how to reduce communication bandwidth consumption while ensuring the accuracy of cooperative perception is crucial. Intermediate fusion [6, 19, 30] transmits feature information, offering a trade-off between bandwidth occupation and accuracy. However, despite the reduction in data volume compared to early fusion, the bandwidth occupation of intermediate fusion still falls short of the industrial standard. Moreover, most intermediate fusion methods face the challenge of spatial alignment [11, 27, 30], which places high requirements on the real-time pose between the vehicle and infrastructure-side equipment, resulting in insufficient robustness of feature fusion. Additionally, the features from the vehicle-side and the infrastructure-side sensors belong to two domains, and domain gaps [35] in the features can also affect the accuracy of collaborative perception.

In this work, we propose TransIFF, a robust and effective instance-level feature fusion framework based on transformers. Our key idea is to reduce the bandwidth consumption by transmitting instance-level features instead of the entire features, while improving the precision of collaborative perception through aligning the domain gaps and achieving a robust and adaptive feature fusion with a transformer that can get rid of the dependence on high-precision pose.

Specifically, TransIFF is composed of three components: a vehicle-side network, an infrastructure-side network, and a vehicle-infrastructure fusion network. The vehicle-side and infrastructure-side networks are equipped with isomorphic convolutional backbones to extract 3D features, which are designed to reduce the domain gaps between the extracted features, as compared to heterogeneous backbones. Transformer decoders and a FFN module are utilized as the detection head to predict the initial bounding boxes of each side by a set of object queries. To further narrow the domain gap, the network weights of the two sides are shared and pre-trained through hybrid training. Subsequently, the output features from the respective transformer decoders will pass through a feature filter module to extract high-confidence features. Only the high-confidence features of the infrastructure-side are sent to the vehicles, which greatly reduce communication bandwidth consumption.

In the vehicle-infrastructure fusion network, the features from the vehicle-side and infrastructure-side go through Cross-Domain Adaptation (CDA) module to further align their domains. The CDA module transforms the infrastructure-side position encoding into the vehicle coordinate system by pose transformation matrix and then uses two parallel cross-domain attention to align the domain gaps. Moreover, the CDA module enhances the features of a single side by utilizing information from the other side. Finally, the features of both sides are fused by a Feature Magnet (FM) module to output the collaborative perception results.

Overall, incorporating all the components, our proposed TransIFF can achieve a remarkable low bandwidth consumption, while maintaining a high collaborative perception precision (Fig. 1). To summarize, the main contributions are as follows:

- We propose TransIFF, a transformer-based vehicle-infrastructure collaborative perception fusion framework, which can realize best performance-bandwidth trade-off by transmitting instance-level features.
- To address the domain gap issue in the intermediate fusion, we design Cross-Domain Adaptation (CDA) module. Additionally, vehicle-infrastructure isomorphic backbones and hybrid pre-training are also used to help achieving domain alignment.
- We introduce several simple yet effective designs to improve the robustness of feature fusion, such as unified positional encoding and Feature Magnet (FM) module. Our TransIFF achieves state-of-the-art performance on the widely used real-world benchmark DAIR-V2X, achieving 59.62% AP while only taking up 2^{12} bytes bandwidth.

2. Related Work

Cooperative Perception Datasets. In order to promote the development of the cooperative perception, many high-quality datasets have been released these years. Rope3D [41] and WIBAM [25] are two datasets for infrastructure-side 3D detection. OPV2V [38], V2X-Sim [18] and V2XSet [37] are simulated datasets that focus on multi-agent cooperative perception. V2XP-ASG [33] is an open adversarial scene generation framework which can create challenging scenes used as simulated V2X dataset. DAIR-V2X [42] is the only real world vehicle-infrastructure cooperative benchmark which supports both camera and LiDAR modalities. However, originally DAIR-V2X only annotates 3D boxes within the range of camera's view in vehicle-side. CoAlign [24] supplements the missing 3D box annotations to enable the 360 degree detection.

Cooperative 3D Object Detection. Utilizing cooperative perception to assist single-vehicle autonomous driving has attracted more research attentions [12, 21]. CSVNet [28] first uses multiple sets of images between two vehicles to achieve better accuracy in controlling steer angle. Early and late fusion schemes are first introduced in a vehicle-infrastructure cooperative perception system in

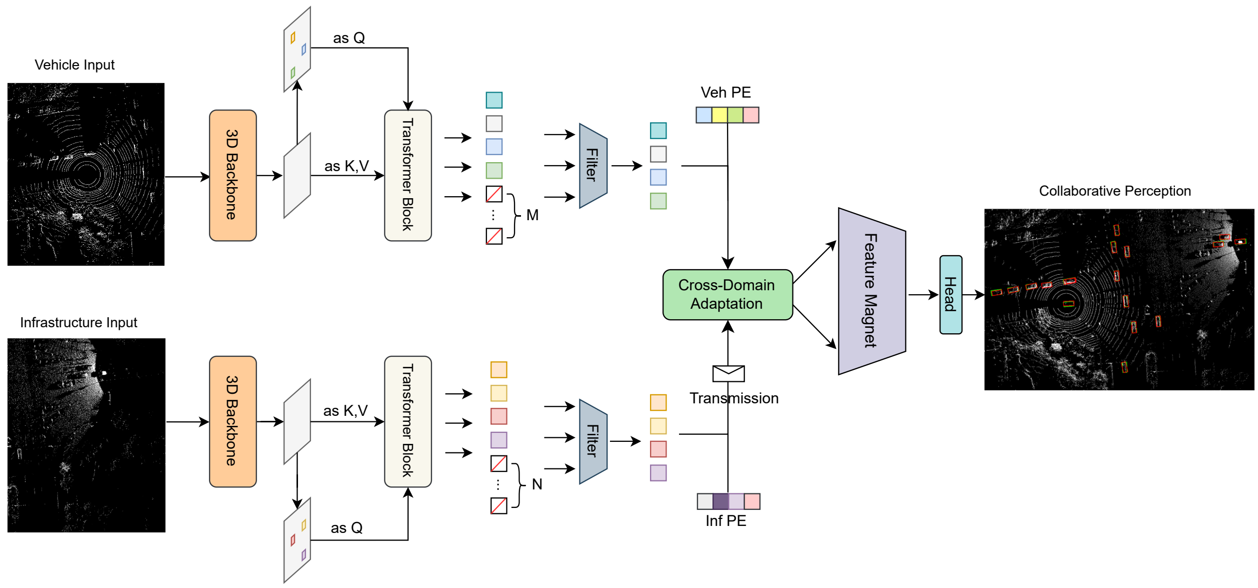

Figure 2. **Overall pipeline of the TransIFF.** Given LiDAR input from the vehicle side and infrastructure side, TransIFF first extracts instance-level features through the vehicle-side and infrastructure-side network. Then, transmits infrastructure-side instance-level features to the vehicle side and performs Cross-Domain Adaptation (CDA) to align feature domains. Finally, the Feature Magnet (FM) module is built to achieve adaptive feature fusion. Green boxes are predicted bounding boxes, while red boxes are ground truth bounding boxes.

method [2]. WIBAM [25] model develops a weak supervision method for fine-tuning models of traffic observation cameras. Cooper [7] combines raw point clouds from different vehicles and designs SPOD network. V2VNet [30] first proposes the intermediate fusion method, which utilizes a spatially aware graph neural network to fuse information from multiple vehicles. DiscoNet [19] promotes better performance-bandwidth trade-off for multi-agent perception by training a DiscoGraph via knowledge distillation. V2X-ViT [37] achieves feature fusion across vehicles and infrastructures based on a vision transformer framework. SyncNet [17] is proposed for time-domain synchronization. Where2comm [13] introduces spatial confidence maps to reduce the communication bandwidth consumption. AdaFusion [26] proposed three adaptive bird's eye view (BEV) feature fusion models to further increase the perception accuracy. MPDA [35] first presents domain gap problem among different agents and utilize the entire feature maps to unify the pattern, suffering from a huge communication consumption.

Transformer in Vision. Transformer [29] has revolutionized the field of natural language processing. It is composed of multi-head self-attention mechanisms, which enable the network to capture complex relationships between elements in a sequence. Since attention has no ability to learn location information, additional positional encoding always be encoded into the features. Transformer has been proven to be highly effective in computer vision tasks, such as image classification [5, 9], object detection [3, 4], and segmentation [34]. Additionally, the parallel processing capabilities of the transformer have made it feasible for real-time video field [1, 20].

3. Methodology

3.1. Overall

The architecture of the TransIFF is depicted in Fig. 2. To obtain the initialization weight of the model, we alternately use the input data from the vehicle-side and infrastructure-side for training. After that, the LiDAR clouds from both sides are fed into the isomorphic 3D Backbone for preliminary feature extraction. These features from the vehicle side and the infrastructure side are then fed into a transformer block in parallel to extract instance-level features. The filter module is then employed to eliminate features with low confidence levels, This can preliminarily reduce the bandwidth of transmission and decrease the generation of false positive results. To align the domain of features from both sides, we propose Cross-Domain Adaptation (CDA) module. In this module, the position encoding of the infrastructure-side is transformed into the coordinate system of the vehicle-side, which implicitly encodes the spatial information into the instance-level feature of the infrastructure side. The instance-level features from both sides then perform cross-domain attention in parallel to align domains. Furthermore, we introduce the Feature Magnet (FM) module, which fuses features by deduplicating the same instance-level features, thereby generating

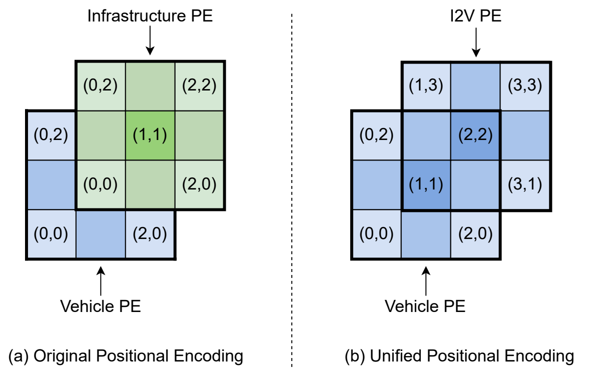

Figure 3. **Positional encoding comparison of (a) Original Positional Encoding and (b) Unified positional encoding.** The original positional encodings of the vehicle side and infrastructure side are generated independently. In contrast, we transform infrastructure-side positional encoding to a vehicle coordinate system to generate a unified positional encoding, where both sides share the same coordinate space.

unique instance-level features. Finally, the instance-level features are input through the transformer decoder and detection head for collaborative perception results.

3.2. Unified Positional Encoding

Due to the diverse coordinate systems used by the infrastructure and the vehicle, positional encoding is essential for instance-level fusion, which demands consistency in the feature representation from several sources. Positional encoding has already been extensively employed in transformers to include location information in features that can address self-attention mechanism flaws.

The ultimate goal of vehicle-infrastructure cooperative tasks is to enhance and supplement the perception abilities of vehicles. Therefore, it is appropriate to use the vehicle coordinate system as a unified coordinate system. Firstly, we calculate the relative position relationship between the vehicle and the infrastructure from the BEV perspective. Due to the absence of height information, the transformation relationship of the position is reduced from six degrees of freedom $(t_x, t_y, t_z, \theta_p, \theta_r, \theta_y)$ to three degrees of freedom (t_x, t_y, θ_y). Then, we transform the independent position encoding of the infrastructure-side into the vehicle-side BEV space through a transformation matrix, the formulas are listed as follows:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta_y & -\sin\theta_y & 0 \\ \sin\theta_y & \cos\theta_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ 0 \end{bmatrix} \tag{1}$$

Finally, due to the numerical precision issues caused by the transformation, we use the nearest neighbor method to align the transformed position encoding to the vehicle side, the fomulas are listed as follows:

$$\begin{bmatrix} x_{out} \\ y_{out} \\ 1 \end{bmatrix} = \begin{bmatrix} round(x') \\ round(y') \\ 1 \end{bmatrix} \tag{2}$$

where $round()$ denotes rounding operator. Following above steps, unified positional encoding is generated, as shown in Fig.3.

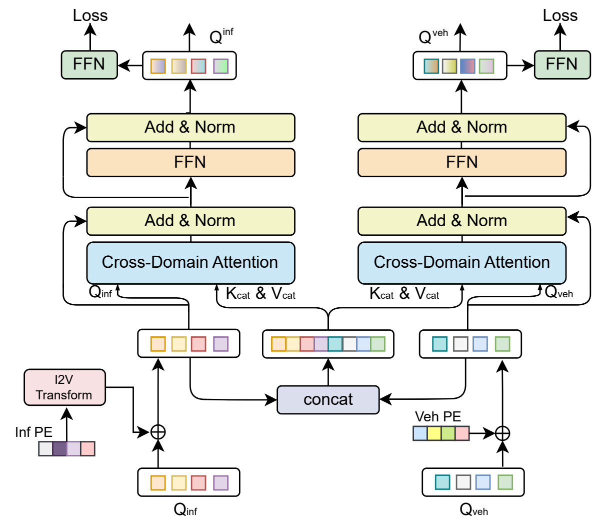

Figure 4. **Cross-Domain Adaptation Module.** It is a two-stream Cross-Domain Adaptation (CDA) module. Infrastructure positional encoding is transformed to vehicle coordinate system for domain alignment. Cross-domain attention between vehicle and infrastructure side are performed to adaptively learn from each other.

3.3. Cross-Domain Adaption

To overcome the domain gap problem between the instance-level features extracted from the vehicle and infrastructure sides, we propose a Cross-Domain Adaptation (CDA) module, as shown in Fig. 4.

In the CDA module, infrastructure-side positional encoding is transformed to vehicle-side coordinate system to align features in space. Then, using the top-ranking grid features from each side as Q_{veh}, Q_{inf}, and the concatenated features from both sides as K_{cat} and V_{cat}, the module performs cross-domain attention in parallel. The formulas for cross-domain attention operation are listed as follows:

$$\begin{aligned} Q^{veh} &= \text{Cross-Domain Attention}(Q_{veh}, K_{cat}, V_{cat}) \\ &= \text{softmax}\left(\frac{Q_{veh} K_{cat}^T}{\sqrt{d_k}}\right) V_{cat} \end{aligned} \tag{3}$$

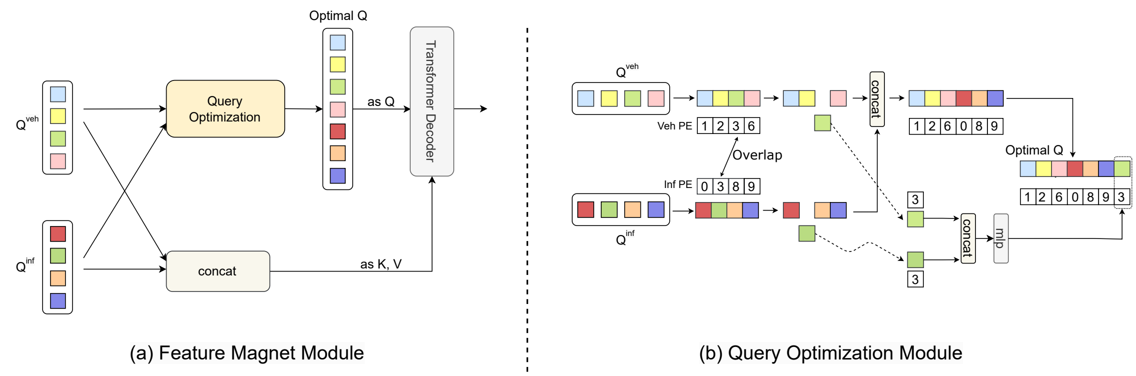

Figure 5. **(a) Feature Magnet Module.** In the Feature Magnet module, the queries from vehicle-side and infrastructure-side are sent to Query Optimization module to obtain optimal queries and then sent to a transformer decoder with concatenated features to fuse features. **(b) Query Optimization Module.** In the Query Optimization module, similar queries from different sides indicate that they correspond to the same object in space. To avoid redundancy, the Query Optimization module splits out similar query pairs, fuses them, and concatenates them back to generate optimal queries, where each query corresponds to a unique object.

$$
\begin{aligned}
Q^{inf} &= \text{Cross-Domain Attention}(Q_{inf}, K_{cat}, V_{cat}) \\
&= \text{softmax}\left(\frac{Q_{inf}K_{cat}^{T}}{\sqrt{d_k}}\right)V_{cat}
\end{aligned} \tag{4}
$$

Through this operation, the queries on one side can adaptively learn and extract heterogeneous feature information from the other side to reduce the domain gap. Due to the fact that the output of the cross-domain attention in the CDA module is a weighted sum of the V_{cat}, which contains all the information from the two domains, the CDA module essentially aims to bring both domains closer to a common intermediate representation. Specifically, by utilizing the shared V_{cat} in the cross-domain attention mechanisms, the module encourages convergence between the two domains towards a central representation.

To further facilitate the adaptation of features belonging to distinct domains, we employ an isomorphic backbone. Both the vehicle-side network and infrastructure-side network use the same 3D backbone to extract features. During training, we initialize the backbone using the weights generated by the vehicle-infrastructure hybrid training as pre-training weights to further reduce the domain gap.

The output features are then fed into a feed-forward network (FFN) to produce prediction results and calculate the corresponding loss respectively. By incorporating features from the other side through CDA, our approach enhances the detection results on each side, making the output queries more informative and useful.

3.4. Feature Magnet Module

Through the CDA module, we generate instance-level features for both the vehicle and infrastructure sides in a similar domain. To fuse the features from both sides, we design a Feature Magent (FM) module which is composed of a Query Optimization module and a transformer decoder.

An intuitive idea is to concatenate the features as Q, K, and V, and then implemente a self-attention operation. However, the concatenated queries contain repeated obstacles in the vehicle-infrastructure common-view area, which negatively impact the bipartite matching and learning difficulty of the network. Therefore, we propose a deduplication operation named Query Optimization module to generate optimal queries without duplicate obstacles, as shown in Fig.5 (b). The optimal queries are of three types: obstacles from the vehicle and infrastructure sides appear separately, and deduplicated obstacles are shared by both sides. The Query Optimization module establishes a one-to-one matching relationship between a query and a ground truth, significantly lowering the network's learning difficulty. We do not perform deduplication on K and V because they require complete information.

In the Query Optimization module, firstly, we find the features with same positional encoding between different sides. The positional encodings with same values indicate that the corresponding features have similar coordinates and space overlaps, representing the same object from two perspectives. Subsequently, we pick out pairs of instance-level features and split them out. For instance, considering the case where only one pair of queries overlap, as shown in Fig. 5, the position encodings of the instance-level features represented by the two green squares from both sides are equal to 3, which means that they overlap in space and may be the same object. Supposing that there are N queries on the vehicle-side denoted as $Q^{veh} \in \mathbb{R}^{N \times C}$, and M queries on the infrastructure-side denoted as $Q^{inf} \in \mathbb{R}^{M \times C}$. We split the pairing overlapped queries out which can be respectively represented as Q_i^{veh} and Q_j^{inf}. After that, we

Method	Reference	Fusion Type	Average Precision (AP)		**Comm** $\downarrow$ (log2)
			IOU@0.5	IOU@0.7	
PointPillars [16]	CVPR 2019	Veh.-Only	50.03	43.57	0
PointPillars [16]	CVPR 2019	Late	54.32	44.58	8.39
Where2comm [13]	NIPS 2022	Intermediate	50.99	39.11	12.39
Where2comm [13]	NIPS 2022	Intermediate	51.01	39.10	15.58
Where2comm [13]	NIPS 2022	Intermediate	58.46	44.46	18.19
Where2comm [13]	NIPS 2022	Intermediate	63.71	48.89	22.35
When2com [22]	CVPR 2020	Intermediate	51.12	36.17	24.62
DiscoNet [19]	NIPS 2021	Intermediate	54.29	44.88	24.62
V2VNet [30]	ECCV 2020	Intermediate	56.01	42.25	24.62
V2X-ViT [37]	ECCV 2022	Intermediate	54.26	43.35	24.62
TransIFF(Ours)	-	Intermediate	59.62	46.03	12.39

Table 1. **Vehicle-infrastructure collaborative 3D detection performance on the DAIR-V2X *val* set.** "Comm" denotes the communication volume used to measure the calculated with Equation 8. Methods are ranked from low to high according to the communication volume. For the fusion type: 1) **Veh.-Only** denotes no collaboration which only uses perception results of the vehicle. 2) **Late** means late fusion. 3) **Intermediate** denotes intermediate fusion.

concatenate the remaining queries on the instance dimension as $Q_{rest} \in \mathbb{R}^{(M+N-2)\times C}$. The formula is as follows:

$$Q_{rest} = (\{Q_n^{veh}\}_{0\le n\le N} \setminus Q_i^{veh}) \| (\{Q_m^{inf}\}_{0\le m\le M} \setminus Q_j^{inf}) \quad (5)$$

According to the $Q_i^{veh} \in \mathbb{R}^{1\times C}$ and $Q_j^{inf} \in \mathbb{R}^{1\times C}$, which have the same positional encoding, they are concatenated on the feature dimension and reduced to the original feature size after passing through a Multilayer Perception (MLP) as $Q_{i\&j} \in \mathbb{R}^{1\times C}$:

$$Q_{i\&j} = \mathrm{MLP}(Q_i^{veh} \| Q_j^{inf}) \quad (6)$$

Finally, we concatenate Q_{rest} and $Q_{i\&j}$ to generate the preferred optimal queries.

$$Optimal\ Q = Q_{rest} \| Q_{i\&j} \quad (7)$$

After the Query Optimization module, as shown in Fig.5(a), we use the optimal queries as Q and the concatenation of vehicle-side and infrastructure-side features output by the CDA module as K, V. The Q, K, V are sent to a transformer decoder to further fuse instance-level features. Compared with fusion methods based on pixel-level feature alignment, our attention-based method does not rely on high-precision poses, and can achieve more robust perception effects.

Overall, our proposed FM module can efficiently and accurately fuse instance-level features from both sides, improving the accuracy of the cooperative perception.

4. Experiments

4.1. Setup

Dataset. We evaluate our proposed method on the widely used real-world vehicle-infrastructure cooperative benchmark DAIR-V2X [42], which contains 9K cooperative frames with a single vehicle and infrastructure-side unit. Originally, the DAIR-V2X does not label objects outside the view of the on-board camera. To enable 360-degree detection range around the vehicle, we use the relabeled dataset following the approach in Where2comm [13] and CoAlign [24]. The vehicle-infrastructure cooperative annotations in the DAIR-V2X only provide labels for motorized vehicles, and does not include non-motorized vehicles or pedestrians. Consistent with the approach taken in Where2comm, we merge the four classes: 'Car', 'Van', 'Bus', and 'Truck', into a single category 'Car' for computing the AP. The percpetion range is set to $x \in [-100.8m, 100.8m]$, $y \in [-40m, 40m]$.

Evaluation metrics. Following official protocol, we use average precision (AP) [10] as main metrics, with Intersection Over Union (IOU) thresholds of 0.5 and 0.7. The communication volume results are presented in logarithmic scale base 2, and the message size is counted in bytes. Mathematically, for a given feature map $F \in \mathbb{R}^{H\times W\times C}$, the communication volume is computed as follows:

$$Comm(F) = log_2(H \times W \times C \times 32/8) \quad (8)$$

where 32 is multiplied since each number is represented using the float32 data type, 8 is divided as the metric byte is used. To ensure a fair and direct comparison of communication results, we do not account for any additional compression of data, features, or models.

Implementation details. Our proposed TransIFF is trained using a batch size of 4 on 4 Tesla V100 GPUs for a total of 20 epochs. We use the AdamW optimizer [23] with an initial learning rate of $1e-4$ and weight decay of $1e-2$. For the LiDAR-based 3D object detection task, we employ SECOND [39] as our 3D backbone. The features are rep-

F	P	C	Q	Average Precision (AP)			Comm
				IOU@0.3	IOU@0.5	IOU@0.7	(log2)
				53.93	48.93	32.91	16.64
✓				53.75	48.97	32.28	12.39
	✓			59.34	54.73	44.92	16.64
✓	✓			59.22	54.88	44.95	12.39
✓	✓		✓	61.81	55.90	45.06	12.39
✓	✓	✓		61.56	57.44	45.85	12.39
✓	✓	✓	✓	**63.11**	**59.62**	**46.03**	**12.39**

Table 2. **Ablation Study on different modules of the TransIFF on the DAIR-V2X *val* set**. 'F' denotes the filter module in the vehicle-side and infrastructure-side, 'P' denotes the positional encoding in the CDA module, 'C' denotes cross-domain attention in the CDA module, and 'Q' denotes the query optimization in FM module.

Strategy	Average Precision (AP)		
	IOU@0.3	IOU@0.5	IOU@0.7
None	54.98	50.52	34.01
Concat	61.92	58.18	44.77
Embedding	**63.11**	**59.62**	**46.03**
Embedding*	62.84	59.43	45.88

Table 3. **Comparison of different positional encoding strategies in the CDA module on the DAIR-V2X *val* set.** '**None**' denotes no positional encoding method used. '**Concat**' means that the positions are directly concat to the query features. '**Embedding**' denotes the query positions are embedded into d-dimensional positional encoding using a MLP, and then are element-wise summed with the query features. '**Embedding***' means that the MLP is shared between the vehicle-side and infrastructure-side in the CDA module.

resented as a bird's-eye-view (BEV) map with dimensions of (504, 200, 128) and a resolution of 0.4m/pixel in length and width. The max instance-level feature size is 200 for both vehicle-side and infrastructure-side with a threshold of filter as 0.1.

4.2. Main Results

Table 1 presents a comparison between the proposed TransIFF and previous methods in terms of the trade-off between detection performance and communication volume. TransIFF achieves an impressive AP performance of 59.62% and 46.03% at IOU thresholds of 0.5 and 0.7, respectively, while utilizing only $2^{12.39}$ bytes of bandwidth. Notably, when compared to the state-of-the-art Where2Comm [13] method under the same communication volume, the TransIFF exhibits a significant improvement of 16.92% and 17.69% in AP performance at IOU thresholds of 0.5 and 0.7, respectively. When considering a similar AP performance (59.62% for TransIFF and 58.46% for Where2comm), TransIFF occupies $2^{5.8}$ times

Method/Metric	3D AP@0.5		
Noise Level σ_t/σ_r $(m/°)$	0.2/0.2	0.4/0.4	0.6/0.6
MASH [11]	50.9	50.9	50.9
FPV-RCNN [43]	56.9	49.5	44.4
V2VNet [30]	58.1	57.1	56.3
V2X-ViT [37]	56.0	54.5	53.3
CoAlign [24]	59.2	57.9	57.0
TransIFF	**59.3**	**58.7**	**58.2**

Table 4. **Detection performance on DAIR-V2X *val* set with pose noises.** The pose noises follow a Gaussian distribution. TransIFF demonstrates robust performance under increasing noise levels.

less bandwidth consumption than the Where2comm, which sufficiently proves the superiority of our model in balancing bandwidth occupation and detection performance. Compared with other intermediate fusion methods (DiscoNet [19], V2VNet [30], and V2X-ViT[37]), our TransIFF can save $2^{12.23}$ times bandwidth consumption while maintaining the highest perception performance. Our results are lower than the Where2comm with the highest bandwidth usage ($AP = 63.71\%$ when $Comm = 2^{24.62}$ bytes), because transmitting high confidence level features may do lose some information.

Furthermore, we evaluate the execution time of TransIFF for inference. Experiments show that TransIFF operates at 0.16 seconds per execution on an NVIDIA V100 GPU using 32-bit floating point, demonstrating its efficiency for real-time perception.

4.3. Ablation Study

In this subsection, we investigate the impact of each component in our method. All ablation results are reported on the DAIR-V2X validation set.

Effectiveness of Different Modules in TransIFF. As reported in the first 2 rows of Table 2, the filter module can effectively reduce the communication bandwidth from $2^{16.64}$ bytes to $2^{12.39}$ bytes. In addition, comparing the results of using positional encoding in the first four lines, it can be seen that using positional encoding will greatly improve the perception results. Furthermore, using cross-domain attention and Query Optimization in the vehicle-infrastructure fusion network can provide additional improvements of 3.89%/4.74%/1.08% in AP under three IOU thresholds.

Different Positional Encoding Strategies. Since position encoding is crucial to the transformer-based network, in this subsection, we compare several different commonly used position encoding methods. As can be seen from Table 3, the method of using embedding performs better than concat. Moreover, there is a slight advantage to utilizing an independent MLP weight for the vehicle-side and infrastructure-side, as opposed to using a same weight for both.

Analysis on Robustness against Localization Error. We

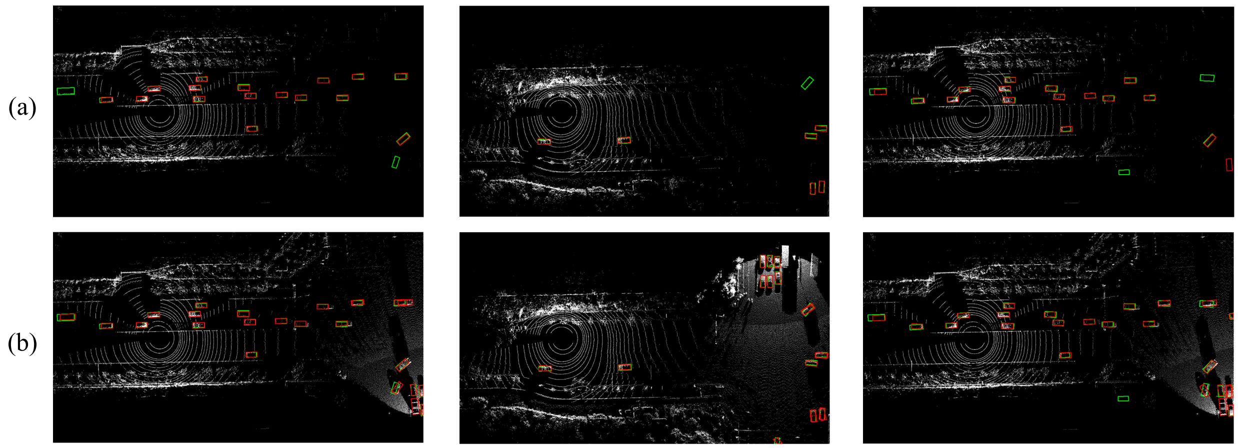

Figure 6. **Visualization of predictions from (a) No Collaboration and (b) TransIFF.** In the first row and second row, we respectively visualize the perception results of the model from the vehicle-side without collaboration fusion and TransIFF. Green boxes are the predicted bounding boxes, while red boxes are the ground truth bounding boxes. It is shown that the TransIFF can effectively compensate for the blind spots on the vehicle-side and improves detection performance.

conduct an experiment to illustrate the model's performance under varying conditions of localization error, represented by different levels of pose noise. The results are presented in Table 4. For each level of noise $\sigma_t/\sigma_r\ (m/^{\circ})$, we compare the AP performance of TransIFF at IOU thresholds of 0.5 with that of several other methods. As can be seen from the table, even as the level of localization error increases, TransIFF continues to exhibit strong resistance to the noise, maintaining a high detection performance.

Influence of the Query Optimization. In this paper, Query Optimization is utilized to remove duplicate elements from the query. We hypothesize that repeated elements could introduce ambiguity in the calculation of the subsequent bipartite matching loss, resulting in increased difficulty in learning. Our experimental results in Table 5 demonstrate the effectiveness of our approach, and we observe that the learning-based method outperforms average pooling.

4.4. Qualitative Results

We conduct a qualitative analysis of the model's performance using typical samples from the DAIR-V2X dataset, as shown in Fig.6. The first row (a) shows the perception results without collaboration, while the second row (b) shows the results obtained using the TransIFF framework for vehicle-infrastructure collaborative perception. Overall, TransIFF is able to generate accurate results over a large field of view. Compared to the model without collaboration, TransIFF supplements the infrastructure-side field of view and incorporates relevant features, resulting in the detection of previously undetectable objects and expanding the range of perception. Even in the field of view area within the range of only one side, the detection accuracy is still improved (e.g., the left-most object in the first row). The qualitative analysis confirms the effectiveness of the TransIFF framework for improving cooperative perception.

Strategy	Average Precision (AP)		
	IOU@0.3	IOU@0.5	IOU@0.7
None	61.56	57.44	45.85
Avg Pooling	62.65	57.88	45.77
Concat + MLP	**63.11**	**59.62**	**46.03**

Table 5. **Comparison of different strategies in the Query Optimization on the DAIR-V2X *val* set.** '**None**' denotes no Query Optimization is used. '**Avg Pooling**' means that element-wise average pooling are used in the two query features of both sides. '**Concat + MLP**' means the strategy we used in the Query Optimization module.

5. Conclusion

TransIFF is a feature fusion framework based on transformers that aims to enhance the cooperative perception of autonomous driving by improving fusion accuracy and reducing communication bandwidth. It is composed of three components: a vehicle-side network, an infrastructure-side network, and a vehicle-infrastructure fusion network. The vehicle-side and infrastructure-side networks extract 3D instance-level features, which are then aligned using a Cross-Domain Adaptation (CDA) module in the fusion network. An adaptive feature fusion is performed using a Feature Magnet (FM) module, which output the cooperative results. TransIFF achieves a state-of-the-art performance on the DAIR-V2X benchmark with significantly reduced communication bandwidth usage, making it a robust and effective framework for vehicle-infrastructure cooperative perception.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021. 3

[2] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1852–1864, 2020. 3

[3] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099, June 2022. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[5] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, October 2021. 3

[6] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. 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*, SEC '19, page 88–100, New York, NY, USA, 2019. Association for Computing Machinery. 2

[7] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019. 3

[8] Weizhe Chen, Runsheng Xu, Hao Xiang, Lantao Liu, and Jiaqi Ma. Model-agnostic multi-agent perception framework. *arXiv preprint arXiv:2203.13168*, 2022. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 6

[11] Nathaniel Glaser, Yen-Cheng Liu, Junjiao Tian, and Zsolt Kira. Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2406–2413. IEEE, 2021. 2, 7

[12] Laurens Hobert, Andreas Festag, Ignacio Llatser, Luciano Altomare, Filippo Visintainer, and Andras Kovacs. Enhancements of v2x communication in support of cooperative autonomous driving. *IEEE communications magazine*, 53(12):64–70, 2015. 2

[13] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Thirty-sixth Conference on Neural Information Processing Systems (Neurips)*, November 2022. 2, 3, 6, 7

[14] Bo Ju, Wei Yang, Jinrang Jia, Xiaoqing Ye, Qu Chen, Xiao Tan, Hao Sun, Yifeng Shi, and Errui Ding. Danet: Dimension apart network for radar object detection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 533–539, New York, NY, USA, 2021. Association for Computing Machinery. 1

[15] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[16] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 6

[17] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 316–332. Springer, 2022. 3

[18] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: A virtual collaborative perception dataset and benchmark for autonomous driving. 2022. 2

[19] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 2, 3, 6, 7

[20] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14040–14049, October 2021. 3

[21] Shaoshan Liu, Bo Yu, Jie Tang, and Qi Zhu. Towards fully intelligent transportation through infrastructure-vehicle cooperative autonomous driving: Challenges and opportunities. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 1323–1326. IEEE, 2021. 2

[22] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 6

[23] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam, 2018. 6

[24] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. *arXiv preprint arXiv:2211.07214*, 2022. 2, 6, 7

[25] Jamie Mackenzie Matthew Howe, Ian Reid. Weakly supervised training of monocular 3d object detectors using wide baseline multi-view traffic camera data. *32nd British Machine Vision Conference, BMVC 2021*, 2021. 2, 3

[26] Donghao Qiao and Farhana Zulkernine. Adaptive feature fusion for cooperative perception using lidar point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1195, January 2023. 3

[27] Nicholas Vadivelu, Mengye Ren, James Tu, Jingkang Wang, and Raquel Urtasun. Learning to communicate and correct pose errors. In Jens Kober, Fabio Ramos, and Claire Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 1195–1210. PMLR, 16–18 Nov 2021. 2

[28] Rodolfo Valiente, Mahdi Zaman, Sedat Ozer, and Yaser P. Fallah. Controlling steering angle for cooperative self-driving vehicles utilizing cnn and lstm-based deep networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2423–2428, 2019. 2

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[30] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*, pages 605–621. Springer, 2020. 2, 3, 6, 7

[31] Hai Wu, Chenglu Wen, Wei Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. In *AAAI*, 2023. 1

[32] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5418–5427, June 2022. 1

[33] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception. *arXiv preprint arXiv:2209.13679*, 2022. 2

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 3

[35] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 3

[36] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning (CoRL)*, 2022. 1

[37] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6, 7

[38] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2

[39] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 6

[40] Xiaoqing Ye, Liang Du, Yifeng Shi, Yingying Li, Xiao Tan, Jianfeng Feng, Errui Ding, and Shilei Wen. Monocular 3d object detection via feature domain adaptation. In *European Conference on Computer Vision*, pages 17–34. Springer, 2020. 1

[41] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21341–21350, June 2022. 2

[42] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. 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 (CVPR)*, pages 21361–21370, June 2022. 2, 6

[43] Yunshuang Yuan, Hao Cheng, and Monika Sester. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):3054–3061, 2022. 7