INSTINCT: Instance-Level Interaction Architecture for Query-Based Collaborative Perception

Yunjiang Xu[1], Lingzhi Li[1]*, Jin Wang[2]*, Yupeng Ouyang[1], Benyuan Yang[2]
[1]School of Computer Science and Technology, Soochow University
[2]School of Future Science and Engineering, Soochow University
{yjxu95, ouyangchina1}@gmail.com, {lilingzhi, wjin1985, byyang}@suda.edu.cn

Abstract

*Collaborative perception systems overcome single-vehicle limitations in long-range detection and occlusion scenarios by integrating multi-agent sensory data, improving accuracy and safety. However, frequent cooperative interactions and real-time requirements impose stringent bandwidth constraints. Previous works proves that query-based instance-level interaction reduces bandwidth demands and manual priors, however, LiDAR-focused implementations in collaborative perception remain underdeveloped, with performance still trailing state-of-the-art approaches. To bridge this gap, we propose **INSTINCT** (**INST**ance-level **IN**tera**C**tion Archi**T**ecture), a novel collaborative perception framework featuring three core components: 1) a quality-aware filtering mechanism for high-quality instance feature selection; 2) a dual-branch detection routing scheme to decouple collaboration-irrelevant and collaboration-relevant instances; and 3) a Cross Agent Local Instance Fusion module to aggregate local hybrid instance features. Additionally, we enhance the ground truth (GT) sampling technique to facilitate training with diverse hybrid instance features. Extensive experiments across multiple datasets demonstrate that INSTINCT achieves superior performance. Specifically, our method achieves an improvement in accuracy 13.23%/33.08% in DAIR-V2X and V2V4Real while reducing the communication bandwidth to 1/281 and 1/264 compared to state-of-the-art methods. The code is available at https://github.com/CrazyShout/INSTINCT.*

1. Introduction

Perception is critical in autonomous driving, providing accurate inputs for decision-making and control. LiDAR has emerged as a key sensor for object detection, thanks to its exceptional 3D perception capabilities [11, 26, 27]. However, traditional single-agent perception systems are constrained by limited sensing range and inherent sensor deficiencies, leading to blind spots in obstructed scenarios and adverse weather conditions, posing significant safety risks.

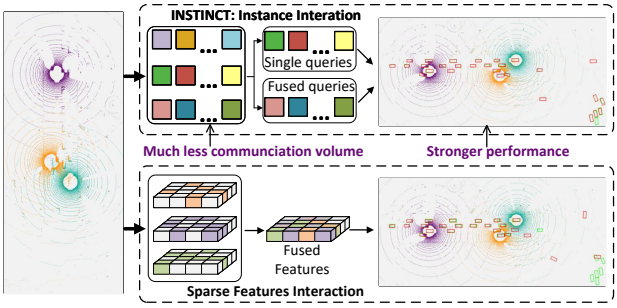

Figure 1. By implementing the innovative instance interaction architecture, INSTINCT demonstrates superior performance while operating under lower bandwidth constraints.

Multi-view and multi-agent collaborative perception addresses these limitations by utilizing distributed sensor networks. For example, infrastructure-based sensors, such as cameras and LiDAR, typically offer higher resolution, improved accuracy, and wider coverage, enabling effective detection of distant objects. Currently, collaborative perception methods are classified into early fusion, intermediate fusion, and late fusion, with intermediate fusion showing the greatest potential. Notably, extensive research has focused on Vehicle-to-Everything (V2X) systems [16, 28, 29], integrating vehicle and infrastructure data to significantly enhance perception performance.

However, as shown in Fig.1, the requirement for real-time perception ($\geq$10Hz) poses a severe challenge to communication bandwidth. Traditional intermediate fusion methods directly transmit complete feature maps [19, 24, 29], resulting in enormous bandwidth pressure. Where2comm [6] filters key information by constructing request and requirement graphs, and CodeFilling [7] further

*Corresponding author.

introduces a learning-based codebook compression mechanism, but the transmission of sparse feature maps still incurs high communication costs.

It is worth noting that query-based methods have been widely studied in single agent perception, and using query features for interaction in collaborative perception is also known as instance level fusion [3, 5, 9, 23]. It is obviously that sending instances features only is accurate and bandwidth friendly. However, current instance interaction methods are mainly focused on camera modality, and the unique representation of LiDAR point clouds limit their direct transfer. Although TransIFF [3] has achieved instance level collaborative perception of LiDAR for the first time, its architecture only supports vehicle-infrastructure (V2I) collaborative scenarios. In addition, TransIFF emphasizes communication bandwidth, but its performance still lags significantly behind advanced models.

To eliminate the problems we mentioned, we proposed INSTINCT, a novel **INST**ance-level **IN**tera**C**tion Archi**T**ecture based on LiDAR-V2X Systems. This architecture achieves breakthroughs through three innovative design: 1) An instance feature filtering mechanism based on selective transmission, effectively reducing communication bandwidth; 2) The dual branch detection architecture achieves collaborative/single agent detection decoupling, eliminating irrelevant instance interference; 3) The Cross-Agent Local Instance Fusion (CALIF) module achieves feature optimization of hybrid instances through domain adaption and Gaussian-distance-based local fusion.

We evaluate INSTINCT on multiple datasets, such as DAIR-V2X [34], V2XSet [29], and V2V4Real [31]. It shows that INSTINCT achieves state-of-the-art performance in all metrics while maintaining a low communication load of $2^{13} - 2^{14}$ bytes/frame, especially on the DAIR-V2X and V2V4Real real-world datasets. The main contributions can be summarized as follows:

- We propose INSTINCT, the first collaborative perception architecture enabling LiDAR-based instance-level interaction in V2X scenarios, achieving a superior balance between bandwidth usage and perception performance.
- We design a filtering mechanism to efficiently communicate instance information, combined with a dual-branch detection strategy for single-agent and collaborative detection, effectively avoiding irrelevant instance interference and improving system robustness.
- We introduce a Cross-Agent Local Instance Fusion module, effectively addressing domain gaps and leveraging local correlations among instances. INSTINCT achieves state-of-the-art performance across multiple datasets while maintaining a minimal bandwidth usage of only 2^{13}-2^{14} bytes.

2. Related Work

2.1. Collaborative Perception

Collaborative perception leverages multi-agent interactions to enhance individual vehicle sensing capabilities and driving safety. The availability of high-quality datasets like DAIR-V2X [34], V2XSet [29], V2V4Real [31], and OPV2V [30] has significantly advanced related research. Some studies focus on reducing communication bandwidth while maintaining performance. Where2comm [6] employs spatial confidence maps to eliminate irrelevant information, concentrating on perceptually critical areas. CodeFilling [7] further compresses data via a codebook-based representation, selectively transmitting essential information. Pose errors negatively impact perception performance. CoAlign [18] addresses this through pose graph modeling, while FreeAlign [12] integrates graph neural networks and multi-anchor subgraph search to enhance robustness against pose inaccuracies. To address spatiotemporal misalignment from communication latency, FFNet [35] introduces feature flow prediction for temporal alignment. CoBEVFlow and CoDynTrust [25, 32] propose BEV-flow-based motion compensation to align ROI features, ensuring effective spatiotemporal synchronization.

2.2. Query-based Object Detection

DETR [2] pioneered an end-to-end detection paradigm by reconstructing the detection process with an attention mechanism and a query stream, significantly simplifying the complex procedures of traditional object detection. Subsequent studies have focused on its optimization. For instance, Deformable-DETR [38] achieves efficient local interactions through a deformable attention mechanism, dramatically reducing computation, and DN-DETR [13] introduces a denoising training strategy to accelerate model convergence. Building on the success of DETR-based method in 2D detection, its ideas have been extended to the 3D detection domain. Specifically, for camera modalities, BEVFormer [15] constructs a grid-based query structure in a bird's-eye view to enhance 3D spatial encoding. Sparse-BEV [27] introduces scale-adaptive attention along with a hybrid spatiotemporal sampling strategy to achieve fully sparse BEV detection. For LiDAR modalities, ConQueR [37] pioneered a query contrast strategy to eliminate dense false positives, and SEED [17] enhances detection accuracy by reconstructing the query selection and interaction modules.

3. Problem formulation

Consider N agents in the scene, where each agent is able to send and receive information to and from other agents. Let $\mathcal{X}_i$ be the raw input data for i-th agent, and $\mathcal{Y}_i$ is the

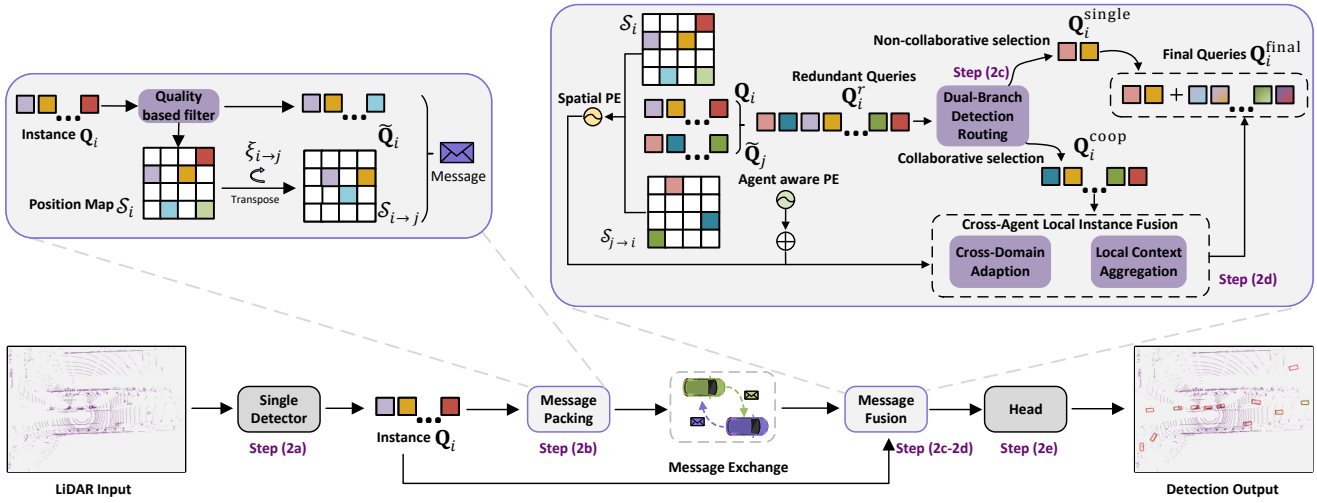

Figure 2. **System Overview.** Given LiDAR data for each intelligent agent, INSTINCT first extracts instance-level features using a single-agent detector. Next, it filters these features and attaches a spatial position map before transmitting them to other agents. Upon receiving messages from other agents, INSTINCT integrates its own features with the incoming features.

corresponding ground truth in the scene. The objective is to maximize the detection performance of all agents with total communication cost B; that is,

$$max \sum_{i}^{N} g(f_\theta(\mathcal{X}_i, \{\mathcal{P}_{j\to i}\}_{j=1}^{N}), \mathcal{Y}_i) \quad s.t. \sum_{i=1}^{N} |\mathcal{P}_{i\to j}| \le B, \tag{1}$$

where $g(\cdot,\cdot)$ is the detection evaluation metric, f is the collaborative perception network with trainable parameter θ, and $\mathcal{P}_{j\to i}$ denotes the message transmitted from j-th agent to i-th agent.

4. Methodology

4.1. Overall

For the i-th agent, the proposed collaborative works as follows:

$$\mathbf{Q}_i = f_{\text{single}}(\mathcal{X}_i) \tag{2a}$$

$$\tilde{\mathbf{Q}}_i, \mathcal{S}_{j\to i} = f_{\text{filter}}(\mathbf{Q}_i, \xi_{j\to i}) \tag{2b}$$

$$\mathbf{Q}_i^{\text{single}}, \mathbf{Q}_i^{\text{coop}} = f_{\text{DDR}}(\mathbf{Q}_i, \tilde{\mathbf{Q}}_j) \tag{2c}$$

$$\mathbf{Q}_i^{\text{final}} = f_{\text{CALIF}}(\mathbf{Q}_i^{\text{coop}}, \mathcal{S}_i, \mathcal{S}_{j\to i}) + \mathbf{Q}_i^{\text{single}} \tag{2d}$$

$$\hat{\mathcal{Y}}_i = f_{\text{head}}(\mathbf{Q}_i^{\text{final}}) \tag{2e}$$

As shown in Fig.2, in step (2a), $\mathbf{Q}_i \in \mathbb{R}^c$ denotes queries extracted from $\mathcal{X}_i$ by f_{single}. In step (2b), f_{filter} means Filtering Module that filter $\mathbf{Q}_i$. $\tilde{\mathbf{Q}}_i$ denotes filtered queries features and $S_{j\to i}$ denotes the spatial position map j-th agent coordinate to i-th agent, which will be discussed in 4.3. $\xi_{j\to i} \in \mathbb{R}^{4\times4}$ denotes spatial transform matrix.

After receiving $\tilde{\mathbf{Q}}_j$ sent from j-th agent, i-th agent start to collaborate. In step (2c), i-th agent send $\tilde{\mathbf{Q}}_\mathbf{j}$ and $\mathbf{Q}_i$ into the DDR (Dual-Branch Detection Routing) module, separately computing the collaborative queries and non-collaborative queries. Next, at step (2d), CALIF (Cross-Agent Local Instance Fusion module) aggregates collaborative features, performs cross-domain adaptation adjustments, and finally conducts local instance interactions. Then, at step (2e), all concatenated instance-level features $\mathbf{Q}_i^{\text{final}}$ are passed through a detection head to obtain the final detection result $\hat{\mathcal{Y}}_i$.

4.2. Single-Agent Detector

As the foundation of collaborative perception, single-agent detection provides essential perceptual information for the collaboration process. Inspired by ConQueR [37], we use BoxAttention [20] as a key component of both the Encoder and Decoder, where the Encoder is responsible for feature extraction and the Decoder corrects the reference boxes.

To better supervise training, we also modify the loss function for supervising the final output, which will be explained in Section 4.3. Additionally, a comparison with advanced single-vehicle models for direct post-fusion is listed in the appendix. In summary, following the single-agent detection process, each agent produces unfiltered outputs.

4.3. Quality-Aware Filtering

After the single-agent detection stage, the collaboration process determines the optimal moment for cooperation based on predefined rules, such as a simple distance-triggered condition. Specifically, when the distance between two agents falls below a preset threshold, collaboration is triggered, executing steps (2b) to (2e).

In step (2b), we perform the following filtering oper-

ation: 1) the preparation of high-quality collaborative instance features to avoid unnecessary interference during the collaboration; and 2) the preparation of a unified spatial position marker for each instance feature to ensure that, in subsequent steps, the system can understand the relative positions between all instances. Additionally, based on the unified spatial position markers, instance features outside the ego vehicle's perception range can be excluded from the communication process in advance. Both of these operations help filter unnecessary features, which further reduces the transmission bandwidth.

Recent studies [17, 36] have made corresponding improvements in query selection. Since confidence scores only reflect the credibility of classification but do not capture the reliability of box regression, it can lead to an unfair query selection process. To ensure that the quality of the instance features transmitted during information exchange is sufficiently high, we consider two methods: 1) predicting the IoU score with a separate branch, and 2) applying an IoU-related penalty to the classification loss. Ultimately, we chose the second approach. This decision was based on the different performance of the separate branch IoU prediction across both synthetic and real-world datasets, with related analysis presented in the appendix. Specifically, we apply an IoU-based penalty using the MAL [8] loss in the final layer of the single-object decoder. The loss formula is as follows:

$$\mathrm{MAL}(p,q,y) = \begin{cases} -q^{\gamma}\log(p) + (1-q^{\gamma})\log(1-p) & q > 0 \\ -p^{\gamma}\log(1-p) & q = 0 \end{cases} \tag{3}$$

Here, p and q, ranging within $[0,1]$, denote the foreground prediction probability and the IoU between the predicted and ground-truth boxes, respectively. The tunable parameter γ balances the gradient contributions from hard and easy samples. Notably, while the MAL loss was originally designed for 2D detection, its adaptation to 3D point cloud detection introduces heightened numerical sensitivity due to the precise alignment of eight corner points in 3D IoU (compared to four vertices in 2D IoU), often causing training instability. This issue is particularly pronounced when MAL is applied across multiple decoder layers. To address this, we apply MAL solely in the final layer and employ the computationally stable Bird's Eye View (BEV) IoU as the baseline for q.

To reduce the amount of communication bandwidth, it is a natural idea to only transmit the perception instances that inside the ego's perception range, which aligns with the idea of having a unified location marker based on a coordinate system. Therefore, we first label the position of each instance feature in its spatial domain, creating a sparse 2D relative position map $\mathcal{S}_i$. Then, we apply a spatial transformation matrix $\xi_{i\to j}$ to transform this position map. Based on the transformed position map, we obtain $\tilde{\mathbf{Q}}_i$ and the corresponding unified spatial position map $\mathcal{S}_{i\to j}$. Afterward, $\tilde{\mathbf{Q}}_i$ and $\mathcal{S}_{i\to j}$ are packaged and transmitted to other agents.

Overall, through QAF (Quality-Aware Filtering) module, we effectively select features that need to transmit, aligned coordinates spatially, reducing the transmission bandwidth and improving accuracy.

4.4. Dual-branch Detection Routing

It is important to note that step (2c) is based on the intuition that if an instance does not have any corresponding instances in the entire collaborative scene, it cannot gain any benefit from the collaboration process. Conversely, if this instance is not distinguished and is directly placed into the collaboration, it may negatively affect the collaboration. To address this, the following steps are executed:

1. First, all received instance features are concatenated into an instance vector table $\mathbf{Q}_i^r$, which may contain redundant element.
2. Next, $\mathbf{Q}_i^r$ is sent to a detection head, obtaining detection results. Here, the Head shares parameters from the final output layer of the single-agent detector.
3. Finally, we compute the IoU between all pairs of vectors in the vector table, forming an IoU matrix $\mathcal{M}_{iou}$. After removing the diagonal entries, we filter the rows of $\mathcal{M}_{iou}$ where all values are below a threshold λ, and retrieve the corresponding instances from $\mathbf{Q}_i^r$ to form $\mathbf{Q}_i^{single}$. The remaining instances constitute $\mathbf{Q}_i^{coop}$, where λ is a hyperparameter threshold.

Through the above steps, we divide all instance features into two categories: one as $\mathbf{Q}_i^{single}$ that do not require collaboration, and the other as $\mathbf{Q}_i^{coop}$ that are potentially need fusion for collaboration.

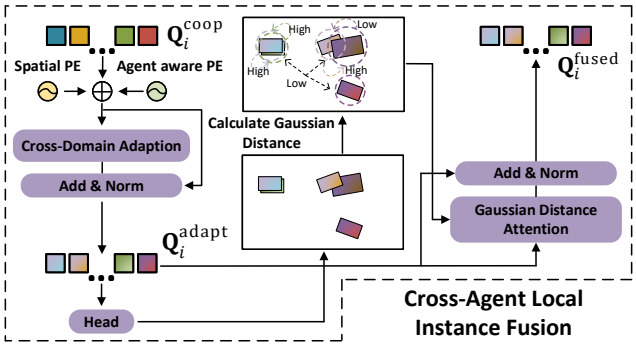

Figure 3. **Overall Structure of CALIF.** In this module, instance features are added with position encoding, then forwarded to the CDA module. Its results are used to compute Gaussian distances, identifying relevant features to obtain $\mathbf{Q}_i^{fused}$.

Model	Fusion Type	DAIR-V2X	V2XSet	V2V4Real	Comm(log2)
		AP@0.5/0.7 ↑	AP@0.5/0.7 ↑	AP@0.5/0.7 ↑	DAIR-V2X/V2XSet/V2V4Real
No Fusion	\	0.6349/0.4962	0.6515/0.5204	0.3980/0.2200	0
Late Fusion	\	0.6043/0.3746	0.7876/0.6801	0.5500/0.2670	8.39/9.60/9.79
V2VNet[24]	Intermidate	0.6344/0.4227	0.8274/0.6582	0.6470/0.3360	24.62/25.10/25.10
V2XViT[29]	Intermidate	0.7349/0.5675	0.8431/0.7018	0.6490/0.3690	24.62/25.10/25.10
DiscoNet[19]	Intermidate	0.7469/0.5920	0.8778/0.7207	0.6412/0.3451	24.62/25.10/25.10
CoBEVT[28]	Intermidate	0.6825/0.5787	0.8454/0.7119	0.5556/0.2708	24.62/25.10/25.10
Where2comm[6]	Intermidate	0.7901/0.6649	0.9259/0.8492	0.7021/0.3801	21.72/21.19/22.86
CoAlign[18]	Intermidate	0.7797/0.6547	**0.9293**/0.8466	0.7209/0.4656	24.62/25.10/25.10
Ours	Instance	**0.8191/0.7529**	0.9229/**0.8731**	**0.8088/0.6196**	**13.58/14.16/14.81**

Table 1. Different models' performance on the different datasets.

4.5. Cross-Agent Local Instance Fusion

Step (2d) addresses the fusion of collaborative mixed instance features $\mathbf{Q}_i^{coop}$, where two components are designed to resolve two practical challenges: 1) To mitigate inevitable domain gaps arising from heterogeneous hardware configurations across agents and environmental disparities, we need to introduce a cross-domain interaction mechanism. 2) Given the extreme sparsity of LiDAR data distributions, global interaction is computationally redundant and risks introducing extraneous learning tasks. Thus, we need local instance interaction mechanism.

Furthermore, for any two instance features p and q in $\mathbf{Q}_i^{coop}$, the mutual information exchange during agent collaboration exhibits inherent asymmetry. More specifically, the degree of interaction should depend on the differences between them in order to complement the completeness of each feature's information. Therefore, the interactions among the mixed features must be local and asymmetric.

For the first challenge, a self-attention mechanism is employed to bridge domain gaps, formulated as:

$$\begin{aligned}\mathbf{Q}_i^{adapt} &= \text{CDA}(\mathbf{Q}_i^{coop}, \mathbf{K}_i^{coop}, \mathbf{V}_i^{coop}) \\ &= \text{Softmax}(\frac{\mathbf{Q}_i^{coop}(\mathbf{K}_i^{coop})^T}{\sqrt{d_k}}) \cdot \mathbf{V}_i^{coop},\end{aligned} \tag{4}$$

where CDA denotes Cross-Domain Adaption, and $\mathbf{K}_i^{coop} = \mathbf{V}_i^{coop} = \mathbf{Q}_i^{coop}$. This operation adjusts the data distribution of each instance in the mixed samples, thereby bridging the domain gap. It is important to note that before cross-domain adaptation, dual position encodings were added to each instance. One of the position encoding is the unified spatial position coordinate map from the cooperative information, which serves as the spatial position encoding. The other one is the agent perception encoding, where we assign the ego agent the identifier 0, and the identifiers for the remaining agents are determined sequentially based on the timing of the cooperative process. Finally, we encode the perception identifiers to obtain the agent perception position encoding. These two position encodings enable the interaction module to identify the spatial location of an instance and the corresponding agent it belongs to.

The above operation is not sufficient to optimize each instance; we also need to further calibrate the instances through stronger local relational interactions while masking irrelevant instances. Inspired by TransFusion [1], we designed a local attention mechanism based on Gaussian distance, as illustrated in Fig. 3. This is a heuristic approach in which all mixed instances are through a shared parameter Head, obtaining the corresponding detection boxes for each instance. Then,a circumcircle for each instance's bounding box is generated. For a given instance feature, we focus only on the distance from the center (x, y) of the bounding boxes of other instance features to the center of the circle. This distance guides to generate an attention weight, as shown in the following equation:

$$\mathcal{W}_{k,v} = \exp(-\frac{\sqrt{(x_k - x_v)^2 + (y_k - y_v)^2}}{\beta r_k^2}), \tag{5}$$

where (x_k, y_k) and (x_v, y_v) denotes the centers of two corresponding detection boxes in the mixed instances. Notice that $\mathcal{W}_{k,v} \in (0, 1]$, and β is a hyperparameter used to adjust the range of attention.

$$\begin{aligned}\mathbf{Q}_i^{fused} &= \text{GDA}(\mathbf{Q}_i^{adapt}, \mathbf{K}_i^{adapt}, \mathbf{V}_i^{adapt}) \\ &= \text{Softmax}(\frac{\mathbf{Q}_i^{adapt}(\mathbf{K}_i^{adapt})^T}{\sqrt{d_k}} \\ &\quad + \log(\mathcal{W})) \cdot \mathbf{V}_i^{adapt},\end{aligned} \tag{6}$$

where GDA denotes Gaussian Distance Attention, and $\mathbf{K}_i^{adapt} = \mathbf{V}_i^{adapt} = \mathbf{Q}_i^{adapt}$.

Through this attention mechanism, features will only focus on local features that are close to them, while those that are far away or have large prediction biases will be ignored. In the end, we combine $\mathbf{Q}_i^{fused}$ and $\mathbf{Q}_i^{single}$ into a single entity, $\mathbf{Q}_i^{final}$.

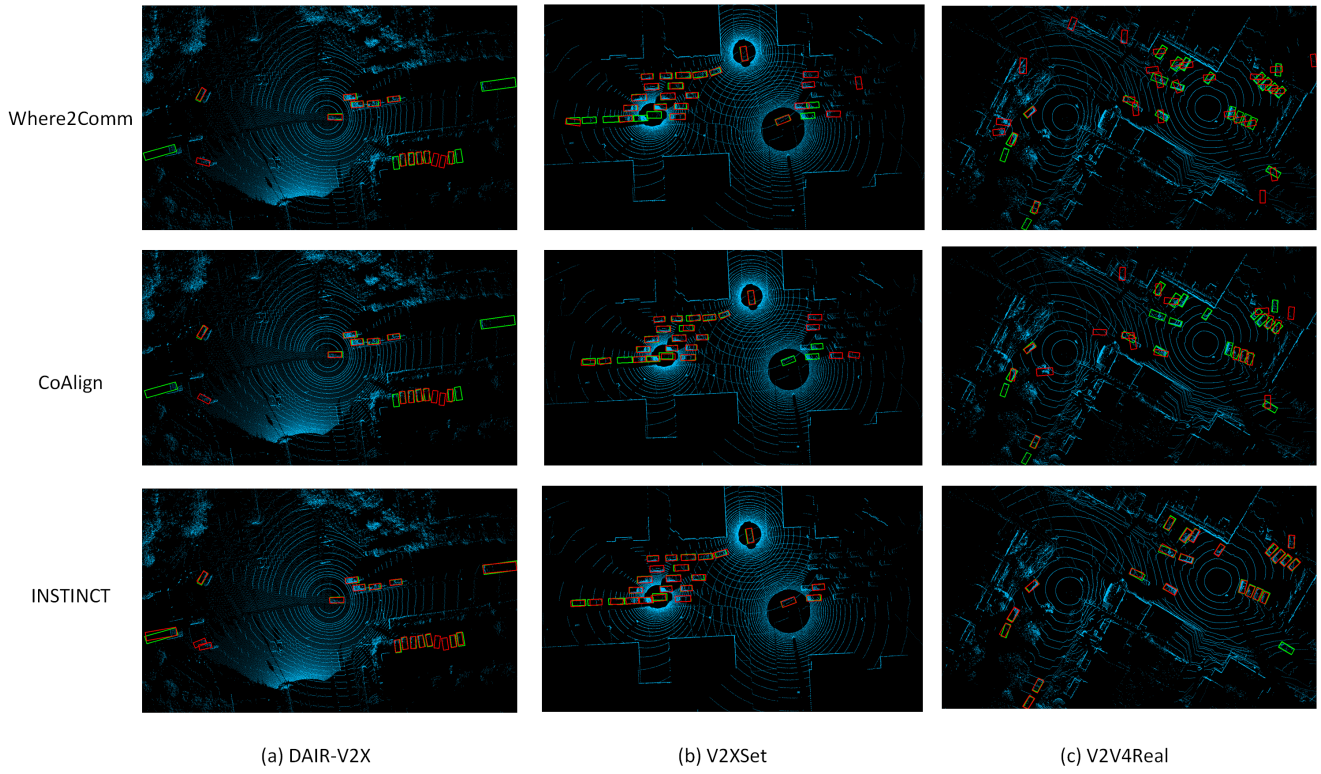

Figure 4. Visual comparison on different datasets. Green box denotes ground-truth, and red box denotes detection result.

4.6. Adaptive Adjustment of GT Sampling

While the proposed design enables effective local instance fusion, viewpoint diversity and scene sparsity often yield limited samples after dual-stage filtering and DDR, compromising model training. To resolve this, we develop Co-GT Sampling specifically designed for cooperative scenarios. In fact, DI-V2X [14] is the first work to apply this method in cooperative perception scenarios, but it uses it solely to bridge the domain gap between the teacher and student models, focusing primarily on domain adaptation in knowledge distillation.

Our implementation involves: 1) Object-Centric Database: We built an object-centered point cloud database by point cloud cropping guided by ground-truth labels and cross-agent mixing. 2) Ego-Centric Sampling: Samples are drawn from the aforementioned database, then we validate them via spatial consistency verification to ensure their validity. Specifically, the sampled labels must neither overlap with each other nor with existing ground-truth annotations. 3) Scenario Synchronization: To maintain scene consistency, we iterates other agents in the collaborative scenario, projects their point clouds into the ego vehicle's coordinate system, and determines whether any point cloud objects should be inserted. Eligible point clouds are then incorporated into the ego point cloud, and their label information is synchronized accordingly. 4) Coordinate Restoration: The processed agent point clouds and the augmented labels are inverse-transformed back to the agent's original coordinate system for single-agent detection supervision.

4.7. Detection Head

The obtained $\mathbf{Q}_i^{final}$ will be fed into the final output head in Step. 2d. The parameters of this output head are independent, and it uses a separate FFN to perform predictions for classification and regression tasks. It is worth noting that we also employ the improved MAL loss for the classification task.

5. Experiments

5.1. Setup

Dataset. We conducted comprehensive experiments on 3 datasets to evaluate our proposed method, including DAIR-V2X [34], V2XSet [29] and V2V4Real [31]. **Dair-V2X** is a public real-world dataset for research on V2X autonomous driving. It includes 71,254 LiDAR frames and 71,254 Camera frames. Each sample scenario includes a vehicle, an infrastructure and its corresponding annotations. **V2XSet** is a large-scale V2X perception dataset created by CARLA [4] and OpenCDA [30]. It is consists with 11,447 frames, including V2V cooperation and V2I cooperation. **V2V4Real**

is the first large-scale real-world dataset for vehicle-to-vehicle collaborative perception. Comprising 410 kilometers of driving coverage, V2V4Real captures multimodal sensor data including 20K LiDAR point clouds, 40K high-resolution RGB images, and 240K precisely annotated 3D bounding boxes, establishing a comprehensive benchmark for V2V cooperative perception research.

Evaluation metrics. Following [3, 6], we use Average Precisions (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 to evaluate perception performance. The communication bandwidth is counted in bytes and represented using a logarithm with base 2.

Implementation details. All experiments with INSTINCT were conducted on one NVIDIA A100 GPU, using SECOND [33] as the 3D backbone. The grid resolution is set to 0.1 m in length, width, and height (with the height adjusted to 0.2 m for V2V4Real). We utilized ConQueR [37] to generate single-agent instances. For classification, we employed a combination of Focal Loss and MAL Loss, while L1 loss was used for regression. We use Adam optimizer [10] with an initial learning rate of 0.001, and the learning rate is adjusted following the OneCycle [21] strategy. Finally, as training approached convergence, we use the Fade strategy [22] to obtain a data distribution that better reflects real-world conditions.

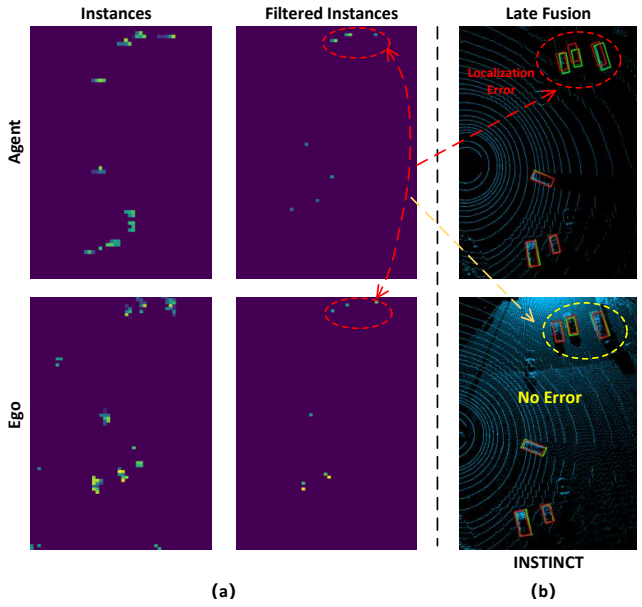

Figure 5. **Visualization of Late Fusion (ConQueR-based) vs. INSTINCT.** (a) Instance feature filtering: agent vs. ego; (b) Detection results: Late Fusion vs. INSTINCT.

5.2. Quantitative Results

Comparison of performance. To evaluate the performance of the INSTINCT model, we conduct multi-model comparative experiments on three benchmark datasets: DAIR-V2X, V2XSet, and V2V4Real. The experiments includes Three types of baselines: a No Fusion baseline without collaborative perception, a Late Fusion method that performs post-fusion based on detection results and six advanced intermediate fusion models including V2VNet, V2XViT, DiscoNet, Where2comm and CoAlign as comparative benchmarks. Detailed experimental data can be found in Tab. 1. The results indicate that INSTINCT demonstrates significant performance advantages across all three datasets. On the real-world datasets DAIR-V2X and V2V4Real, the AP@0.7 metric surpasses the current best models by absolute margins of 13.23% and 33.08%, respectively. And on the simulated dataset V2XSet, a 2.81% performance improvement is maintained, albeit with a relatively limited gain. We attribute these differences to the following factors: (1) Data characteristic disparities. As a simulated dataset, V2XSet exhibits relatively homogeneous scene distributions and object structures, enabling most comparative models to reach a high performance ceiling. However, it is noteworthy that INSTINCT also has limitations when detecting targets in overly sparse point clouds; (2) Validation of architectural advantages. In contrast to conventional feature-level interaction methods, the instance-level interaction mechanism employed by INSTINCT shows stronger adaptability in complex real-world scenarios. It is substantiated by the significant performance leaps observed on DAIR-V2X and V2V4Real. For a more in-depth analysis of the model's performance, supplementary analysis is provided in the Appendix.

Comparison of Communication Bandwidth. Tab. 1 also summarizes the communication bandwidth of each model. The results indicate that while maintaining optimal performance, INSTINCT requires significantly lower communication bandwidth compared to the competing models. Specifically, 1) although the Late Fusion method achieves minimal communication overhead by transmitting low-dimensional bounding box data (averaging of 0.8 KB per frame), it relies on post-NMS fusion mechanism to prevent the calibration of overlapping detection results, leading to a notable performance degradation; 2) in contrast, INSTINCT utilizes medium-dimensional instance feature transmission (averaging 16 KB per frame) and achieves detection calibration through high-dimensional feature fusion, thereby striking the best balance between communication efficiency and detection accuracy.

Robustness to Pose Errors. To evaluate the model's stability under realistic noise environments, we assessed its performance at varying pose noise intensities on the DAIR-V2X dataset (see Tab. 2). To simulate pose noise, Gaussian noise was added during inference to each agent's 2D position (with $\mu_t = 0$ m and $\sigma_t \in [0\text{ m}, 0.5\text{ m}]$) and yaw angle (with $\mu_r = 0$ and $\sigma_r \in [0°, 0.5°]$), emulating real-world lo-

calization errors. The results shows that INSTINCT consistently maintains optimal detection accuracy across different noise levels, demonstrating exceptional environmental robustness.

5.3. Qualitative Results

Detection Performance Visualization. Fig. 4 illustrates a comparison between the detection results of INSTINCT and the current state-of-the-art models, including Where2comm [6], CoAlign [18] *etc*. Visual analysis indicates that INSTINCT consistently achieves lower false positive rates across all three datasets. Moreover, as measured by IoU, the alignment between its predicted bounding boxes and GT is significantly superior to that of the competing model—especially in scenarios with complex occlusions. It demonstrates enhanced object discrimination capability.

Late Fusion vs. INSTINCT: A Visual Analysis. The exceptional performance of INSTINCT stems from its core component—the ConQueR instance generator. As an advanced 3D object detector, we conducted an in-depth comparison of the performance differences between ConQueR-based late fusion and INSTINCT by visualizing the instance distribution in feature maps. As shown in Fig. 5, a clear comparison between subfigures (a) and (b) reveals that in long-range object detection, the ConQueR-based late fusion method exhibits significant localization deviations. In contrast, INSTINCT, leveraging its unique collaborative calibration mechanism, effectively integrates detection results from both the agent and ego vehicles, achieving precise localization alignment. Further details regarding this section are provided in the appendix.

Noise Level $\sigma_t/\sigma_r(m/\circ)$	0.0/0.0	0.1/0.1	0.2/0.2	0.3/0.3	0.4/0.4
Method/Metric	AP@0.5 ↑				
V2VNet	0.6572	0.6476	0.6336	0.6144	0.5974
V2XViT	0.7349	0.7313	0.7190	0.7021	0.6914
DiscoNet	0.7469	0.7436	0.7326	0.7205	0.7071
Where2comm	0.7901	0.7847	0.7619	0.7301	0.7040
CoAlign	0.7797	0.7758	0.7508	0.7220	0.6989
Ours	**0.8191**	**0.8160**	**0.7782**	**0.7327**	**0.7144**
Method/Metric	AP@0.7 ↑				
V2VNet	0.4982	0.4879	0.4705	0.4624	0.4512
V2XViT	0.5675	0.5642	0.5550	0.5445	0.5363
DiscoNet	0.5920	0.5891	0.5812	0.5744	0.5680
Where2comm	0.6649	0.6403	0.6021	0.5718	0.5587
CoAlign	0.6547	0.6327	0.5933	0.5727	0.5604
Ours	**0.7529**	**0.7238**	**0.6451**	**0.6105**	**0.5985**

Table 2. Performance on DAIR-V2X dataset with pose noise.

5.4. Ablation Study

To validate the effectiveness of each module, we conducted systematic ablation experiments on the DAIR-V2X dataset (see Tab. 3). The results are as follows: 1) When all INSTINCT structures are removed, the model degrades to a ConQueR-based late fusion baseline, resulting in a significant performance drop. 2) Enabling the quality-aware filtering module reduces the communication demand to 1/17 of the baseline (a reduction of approximately 94.1%). 3) Introducing a DDR module to separate collaboration-relevant and irrelevant instances not only recovers the baseline performance while maintaining low bandwidth but also further increases the AP@0.7 by 3.43 percentage points compared to baseline. 4) With the CDA module enabled, the interaction of mixed instances combined with DDR to eliminate interference boosts the AP@0.7 by 11.23 percentage points compared to baseline. 5) After adding the CALIF mechanism, the AP@0.7 further increases by 14.16 percentage points compared to baseline. The Co-GT Sampling strategy, by increasing the diversity of positive samples, drives an overall AP@0.7 improvement of 15.51 percentage points compared to baseline.

QAF	DDR	CDA	GDA	GT	AP@0.5/0.7 ↑	Comm (log2)
					0.7302/0.5978	17.67
✓					0.6962/0.6041	13.58
✓	✓				0.7198/0.6321	13.58
✓	✓	✓			0.7898/0.7101	13.58
✓	✓	✓	✓		0.8109/0.7394	13.58
✓	✓	✓	✓	✓	**0.8191/0.7529**	13.58

Table 3. Ablation study results of our proposed core components on DAIR-V2X dataset. **QAF** : Quality-Aware Filtering in 4.3; **DDR** : Dual-Branch Detection Routing in 4.4; **CDA** : Cross-Domain Adaption in 4.5; **GDA** : Gaussian Distance Attention in 4.5; **GT** : Co-GT Sampling in 4.6.

6. Conclusion

We propose INSTINCT, an instance-level interaction architecture based on LiDAR-V2X Systems. To ensure high-quality instances, we design a quality-aware filtering module. INSTINCT further employs a dual-branch detection strategy that decouple collaborative related and unrelated instances, thereby preventing interference. Moreover, during the fusion of hybrid instances, we introduce a cross-agent instance interaction mechanism that incorporates both domain gap compensation and a local interaction attention mechanism based on Gaussian distances. INSTINCT effectively integrates instances transmitted by multiple agents, and experimental results shows that it achieves state-of-the-art performance on benchmark datasets while maintaining an extremely low communication overhead.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62072321), the Science and Technology Program of Jiangsu Province (BZ2024062), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (22KJA520007), Suzhou Planning Project of Science and Technology (2023ss03).

References

[1] 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*, pages 1090–1099, 2022. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Ziming Chen, Yifeng Shi, and Jinrang Jia. Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18205–18214, 2023. 2, 7

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 6

[5] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Quest: Query stream for practical cooperative perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18436–18442. IEEE, 2024. 2

[6] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022. 1, 2, 5, 7, 8

[7] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15481–15490, 2024. 1, 2

[8] Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. *arXiv preprint arXiv:2412.04234*, 2024. 4

[9] Suozhi Huang, Juexiao Zhang, Yiming Li, and Chen Feng. Actformer: Scalable collaborative perception via active queries. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14716–14723. IEEE, 2024. 2

[10] Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*. San Diego, California;, 2015. 7

[11] 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. 1

[12] Zixing Lei, Zhenyang Ni, Ruize Han, Shuo Tang, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative perception without external localization and clock devices. *arXiv preprint arXiv:2405.02965*, 2024. 2

[13] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 2

[14] Xiang Li, Junbo Yin, Wei Li, Chengzhong Xu, Ruigang Yang, and Jianbing Shen. Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3215, 2024. 6

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[16] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4106–4115, 2020. 1

[17] Zhe Liu, Jinghua Hou, Xiaoqing Ye, Tong Wang, Jingdong Wang, and Xiang Bai. Seed: A simple and effective 3d detr in point clouds. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 2, 4

[18] 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. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4812–4818. IEEE, 2023. 2, 5, 8

[19] Eloi Mehr, Ariane Jourdan, Nicolas Thome, Matthieu Cord, and Vincent Guitteny. Disconet: Shapes learning on disconnected manifolds for 3d editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3474–3483, 2019. 1, 5

[20] Duy-Kien Nguyen, Jihong Ju, Olaf Booij, Martin R Oswald, and Cees GM Snoek. Boxer: Box-attention for 2d and 3d transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4773–4782, 2022. 3

[21] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 7

[22] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11794–11803, 2021. 7

[23] Shaohong Wang, Lu Bin, Xinyu Xiao, Zhiyu Xiang, Hangguan Shan, and Eryun Liu. Iftr: An instance-level fusion transformer for visual collaborative perception. In *European Conference on Computer Vision*, pages 124–141. Springer, 2024. 2

[24] 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 *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 605–621. Springer, 2020. 1, 5

[25] Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-robust collaborative perception via bird's eye view flow. In *Advances in Neural Information Processing Systems*, 2024. 2

[26] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2795–2802, 2023. 1

[27] 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*, pages 5418–5427, 2022. 1, 2

[28] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022. 1, 5

[29] 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 *European conference on computer vision*, pages 107–124. Springer, 2022. 1, 2, 5, 6

[30] 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 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2, 6

[31] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. 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*, pages 13712–13722, 2023. 2, 6

[32] Yunjiang Xu, Lingzhi Li, Jin Wang, Benyuan Yang, Zhiwen Wu, Xinhong Chen, and Jianping Wang. Codyntrust: Robust asynchronous collaborative perception via dynamic feature trust modulus. In *arXiv preprint arXiv:2502.08169*, 2025. 2

[33] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 7

[34] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 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*, pages 21361–21370, 2022. 2, 6

[35] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[36] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 4

[37] Benjin Zhu, Zhe Wang, Shaoshuai Shi, Hang Xu, Lanqing Hong, and Hongsheng Li. Conquer: Query contrast voxel-detr for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2023. 2, 3, 7

[38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2