Align before Collaborate: Mitigating Feature Misalignment for Robust Multi-Agent Perception

Kun Yang[1†], Dingkang Yang[1†], Ke Li[3], Dongling Xiao[3], Zedian Shao[4], Peng Sun[2], and Liang Song[1✉]

[1] Academy for Engineering and Technology, Fudan University
[2] Duke Kunshan University
[3] Tencent Youtu Lab
[4] Duke University
{kunyang20,dkyang20}@fudan.edu.cn

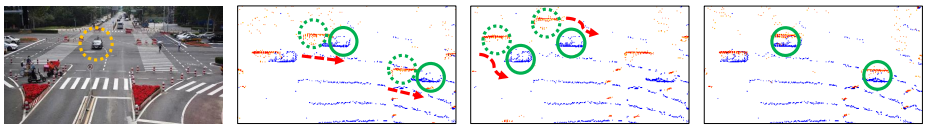
Abstract. Collaborative perception has received widespread attention recently since it enhances the perception ability of autonomous vehicles via inter-agent information sharing. However, the performance of existing systems is hindered by the unavoidable collaboration noises, which induce feature-level spatial misalignment over the collaborator-shared information. In this paper, we propose a model-agnostic and lightweight plugin to mitigate the feature-level misalignment issue, called dynamic feature alignment (NEAT). The merits of the NEAT plugin are threefold. First, we introduce an importance-guided query proposal to predict potential foreground regions with space-channel semantics and exclude environmental redundancies. On this basis, a deformable feature alignment is presented to explicitly align the collaborator-shared features through query-aware spatial associations, aggregating multi-grained visual clues with corrective mismatch properties. Ultimately, we perform a region cross-attention reinforcement to facilitate aligned representation diffusion and achieve global feature semantic enhancement. NEAT can be readily inserted into existing collaborative perception procedures and significantly improves the robustness of vanilla baselines against pose errors and transmission delay. Extensive experiments on four collaborative 3D object detection datasets under noisy settings confirm that NEAT provides consistent gains for most methods with distinct structures.

Keywords: Collaborative perception · Deformable feature alignment

1 Introduction

Perception is a fundamental capability of autonomous vehicles (AVs) to guarantee road safety in sophisticated driving scenarios [22]. Previous single-agent perception has been extensively explored in vision-oriented vehicular applications, including driver monitoring [39], object detection [24, 42], and instance segmentation [21, 52]. Nevertheless, the single-agent perception paradigm is generally

†Equal contribution. ✉Corresponding author.



(a) Infrastructure side image (b) Non-aligned point cloud fusion due to transmission delay (c) Non-aligned point cloud fusion due to localization errors (d) Aligned point cloud fusion with our NEAT aligner

Fig. 1: (a) shows the infrastructure side image where the ego vehicle is marked with the orange circle. (b) and (c) show the point cloud fusion mismatches due to transmission delay and pose errors, respectively. (d) Our NEAT plugin mitigates the feature spatial misalignment issue and produces the well-aligned point cloud fusion.

vulnerable in realistic conditions due to limited sensor ranges [51] and viewpoint occlusions [50]. To this end, collaborative perception has been proposed as a promising solution to alleviate the inadequate observations of individual agents. Based on the Vehicle-to-Vehicle/Everything (V2V/X) communication, existing studies [4,7,8,13,27–29,34–37,45,47,49] effectively improve the perception capabilities of the ego agent through information exchange and perspective complementation among heterogeneous agents (*e.g.*, AVs and infrastructures), resulting in a more holistic and precise understanding of surrounding environments.

Collaborative perception systems are categorized as early [4, 23], intermediate [13], and late [26, 33] fusion, where the feature-level intermediate fusion is the most favored due to its superior trade-off between perception performance and bandwidth cost. However, existing approaches [27, 34, 37] design intermediate collaboration schemes assuming that the spatial transformations between agents are perfect, which are vulnerable to real-world collaboration noises, including pose errors and transmission delay. Specifically, these realistic noises induce spatial misalignment at the feature level, thereby obfuscating the object locations and degrading the collaboration performance. Fig. 1 illustrates visual examples of the feature misalignment issues. We also conduct a toy comparison experiment on the OPV2V dataset [37] to evaluate the hazards posed by collaboration noises. From Fig. 2, the state-of-the-art (SOTA) methods CoBEVT [34], Where2comm [7], AttFuse [37], and F-Cooper [3] exhibit respectable detection precisions in the perfect setting regarding the Average Precision (AP)@IoU 0.7. Nevertheless, when confronted with the noisy setting in realistic scenarios, the potential feature misalignment results in unavoidable performance deterioration, even worse than the No Fusion baseline without collaboration. These findings confirm that the feature misalignment leads to severe performance bottlenecks in existing models. Accordingly, how to effectively mitigate these feature mismatches becomes the core of achieving robust collaborative perception.

Several solutions are proposed to mitigate feature-level misalignment in noisy conditions. For instance, some integrated frameworks leverage attention patterns with larger perceptive fields [36,41,45], multi-scale collaborator features [7,8,27], and historical information [28, 31, 48]. However, these approaches invariably lead to more complicated computations and larger bandwidth costs. Also, these model-specific solutions lack scalability and fail to be applied to other frame-

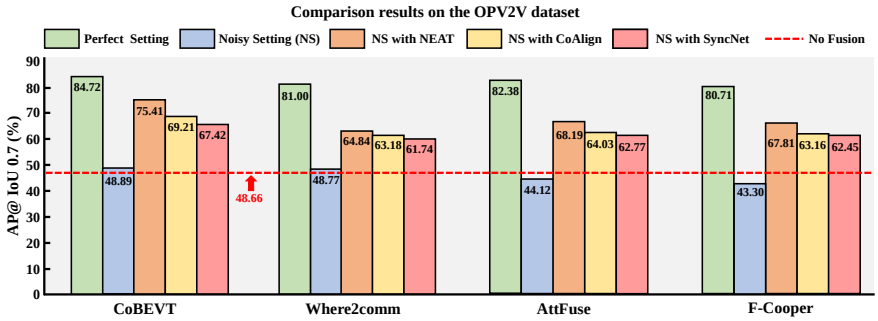


Fig. 2: We conduct a toy comparison experiment on the OPV2V dataset to evaluate the noise interference. The potential feature misalignment in the noisy setting cause the inevitable performance degradation of existing methods compared to the perfect setting. In this case, the NEAT plugin consistently improves the detection precision of the vanilla baselines. Under the same noisy conditions, NEAT brings more significant performance improvements for the baselines compared to SyncNet [11] and CoAlign [18].

works. In addition, various plugins enhance the robustness of existing approaches to specific noises, such as SyncNet [11] for transmission delay and CoAlign [18] for pose errors. Nevertheless, the history-based prediction module in SyncNet introduces significant computation and communication overheads, and the alignment module in CoAlign relies heavily on common visible objects between agents. Moreover, these plugins fail to cope with transmission delay and pose errors simultaneously since these two types of noises induce distinct mismatch patterns. From Fig. 1b, transmission delay causes temporal asynchrony and produces feature misalignment in the object motion directions, which are depicted by the direction-invariant red arrows. Moreover, pose errors bring about more irregular misalignment patterns due to random localization noises, as shown in Fig. 1c. To this end, we propose a lightweight plugin, NEAT, that can be readily integrated into existing frameworks and requires only the ego and collaborator features. As Fig. 2 shows, NEAT achieves more significant performance improvements than CoAlign [18] and SyncNet [11] under the same noisy conditions. Also, the extra parameter number brought from NEAT is only about 1MB (see Table 1).

In summary, we propose a *model-agnostic* and *lightweight* plugin, NEAT, to alleviate the feature misalignment dilemma and facilitate robust collaboration of multi-agent perception systems. Given ego and mismatched collaborator features, our NEAT plugin accomplishes dynamic feature alignment progressively through three tailored components. Specifically, we present the importance-guided query proposal (IQP) module to highlight potential foreground regions in collaborator features. IQP estimates the perceptually critical level of each location and introduces multi-scale views for robust query selection. Then, the deformable feature alignment (DFA) component is devised to integrate valuable visual clues for selected queries and facilitate local semantic enhancement in collaborator features. DFA introduces global context support for object-related to-

ken generation, then leverages the deformable cross-attention to construct query-aware spatial associations and aggregate semantically relevant representations. Ultimately, we propose the region cross-attention reinforcement (RCR) module to diffuse the locally enhanced features into the global representation, producing refined collaborator features with aligned properties for the subsequent feature fusion stage. Our contributions can be summarized as follows:

- The proposed NEAT model is the first dedicated plug-and-play design to address the feature misalignment issues caused by two types of collaboration noises, which can be readily integrated into most methods with diverse architectures and bring significant performance gains consistently.
- The proposed customized components progressively enhance perceptually critical semantics in collaborator-shared features and provide a universal solution for achieving the robust multi-agent perception system.
- Extensive experiments are conducted on multiple collaborative 3D object detection datasets. Comprehensive analysis in real-world and simulated scenarios under noisy settings shows the applicability and effectiveness of NEAT.

2 Related Work

2.1 Collaborative Perception

Collaborative perception aims to incorporate sensor observations from heterogeneous agents to improve the detection capability of the ego agent. Several intermediate collaboration efforts are presented to enhance robustness in noisy conditions, which can be categorized into integrated frameworks and scalable plugins. The former efforts employ framework-specific designs to mitigate noise. For instance, various works utilize attention patterns with larger receptive fields, including Swin Transformer in V2X-ViT [36], axial attention in CoBEVT [34], and deformable attention in SCOPE [45]. Moreover, FFNet [48] and CoBEVFlow [31] apply feature flow to predict the current frame to mitigate transmission delay. These prediction designs introduce significant computation and storage overheads and cause excessive cumulative errors. Multiscale solutions [7,8,17,28,30] employ multiscale collaborator features to overcome local noise interference. However, these methods invariably require sending multiple BEV features and depend on complicated feature fusion mechanisms. The latter efforts alleviate specific noises in a plug-and-play manner. Specifically, SyncNet [11] introduces historical information to predict features based on DiscoNet [13], and CoAlign [18] utilizes co-visible objects to calibrate the inter-agent transfer matrix. Nevertheless, these plugins lack scalability due to their dependence on specific conditions and noises, e.g., SyncNet's requirement for extensive stored collaborator features and CoAlign's dependence on co-visible objects. To summarize, the framework-specific methods fail to provide a unified solution for overcoming noise interference, and the existing plugins can only handle one specific noise. Accordingly, we propose the lightweight NEAT plugin to mitigate transmission delay and pose errors in a model-agnostic manner simultaneously.

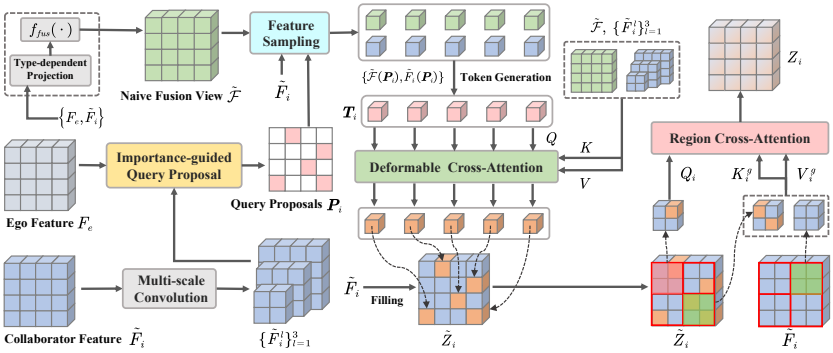


Fig. 3: NEAT plugin overview. Features F_e and $\tilde{F}_i$ are inputs. First, NEAT obtains the multi-scale views $\{\tilde{F}_i^l\}_{l=1}^3$ and utilizes the IQP module to produce the query proposals $\boldsymbol{P}_i$. The naive fusion view $\tilde{\mathcal{F}}_i$ is built via type-dependent projection and $f_{fus}(\cdot)$ to provide global object-related semantics. Afterward, we sample features from $\{\tilde{\mathcal{F}}, \tilde{F}_i\}$ based on the selected 2D positions in $\boldsymbol{P}_i$ and generate the token embedding $\boldsymbol{T}_i$. With deformable cross-attention and $\boldsymbol{T}_i$, NEAT enhances the selected queries and obtains initial aligned feature $\tilde{Z}_i$ via the filling operation. Ultimately, we take a partitioned region as an example to introduce region cross-attention, which generates globally enhanced aligned feature Z_i for the subsequent feature fusion stage.

2.2 Vision Attention for Object Detection

Benefiting from the development of learning-based technologies [38,40,43,44,46], attention mechanisms are widely exploited in vision tasks due to their outstanding contextual modeling capabilities, such as DETR [2] and AutoAlign [5] for object detection. However, vanilla global-wise attention invariably brings heavy computation and storage burdens. Numerous efforts are proposed to alleviate this issue via efficient attention designs, including [1,6,12,14,19,25,32,53]. For instance, deformable DETR [32] leverages the advantages of sparse spatial sampling. AutoAlignV2 [6] builds cross-modal correlations and accelerates feature integration through a deformable attention model. For LiDAR-based object detection, Centerformer [53] presents two sparse attention schemes to facilitate the extraction of meaningful object representations, including window-aware cross-attention and deformable cross-attention. In this paper, we focus on the collaborative 3D object detection tasks in autonomous driving scenarios, which expose more challenges due to real-world collaboration noises. Consequently, we design a deformable attention-based feature alignment component to mitigate error-prone feature associations between collaborators and the ego agent, improving the robustness of existing collaborative perception systems against noises.

3 Methodology

In this section, we formulate the collaborative perception procedure to introduce the working context of our NEAT plugin.

Feature Extraction and Agent Communication. Consider a driving scene with N agents, where e-th agent is identified as the ego agent, and X_i denotes the local point cloud of i-th agent. Each agent utilizes a feature encoder $\Phi_{enc}(\cdot)$ to transform the point cloud X_i into Bird's Eye View (BEV) features as follows:

$$F_i = \Phi_{enc}(X_i) \in \mathbb{R}^{C \times H \times W}, \tag{1}$$

where C, H, W stand for the channel, height, and width. To reduce the required bandwidth cost of F_i, the i-th collaborator adopts the communication mechanism $\Phi_{com}(\cdot)$ to obtain the compressed or sparse feature as $\tilde{F}_i = \Phi_{com}(F_i)$. Then, the i-th collaborator transmits feature $\tilde{F}_i$ and local pose information p_i to the ego agent for subsequent feature fusion stage.

Coordinate Transformation and Feature Fusion. Upon receiving the shared messages $\{\tilde{F}_i, p_i\}$ from the i-th collaborator, the ego agent obtain relative poses with $\{p_e, p_i\}$ and transform $\tilde{F}_i$ into the ego coordinate system via the coordinate transformation operator $\Gamma(\cdot)$ [36] as $\tilde{F}_i = \Gamma(\tilde{F}_i; p_e, p_i)$. After that, the ego agent produces the fused feature $\mathcal{F}_e$ for the final detection with the transformed features as follows:

$$\mathcal{F}_e = \Phi_{fus}(F_e, \{\tilde{F}_i\}_{i=1}^{N}) \in \mathbb{R}^{C \times H \times W}, \tag{2}$$

where $\Phi_{fus}(\cdot)$ denotes the cross-agent feature fusion models, such as attention schemes [7, 36, 37] or graph models [28, 29]. However, collaboration noises (*e.g.*, transmission delay and pose errors) cause feature-level spatial misalignment shown in Fig. 1 and hinder the feature fusion procedure, leading to the performance bottlenecks of existing frameworks under noisy conditions (see Fig. 2).

3.1 NEAT Overview

Unlike previous works [18, 28, 36, 45, 48] that superficially temper noise interference via integrated frameworks or plugins, we develop a *model-agnostic* and *lightweight* NEAT plugin to explicitly mitigate two types of collaboration noises and break performance bottlenecks of existing systems. As Fig. 3 shows, our NEAT plugin only requires the features F_e and $\tilde{F}_i$ as inputs. Then, three tailored components are progressively executed to produce the aligned feature Z_i for replacing $\tilde{F}_i$ in feature fusion. The summarized procedure includes: **(i)** $\tilde{F}_i$ is encoded into three scales with the same channel size using simple convolutions, and $\tilde{F}_i^l$ denotes the feature at l-th scale. With this basis, the importance-guided query proposal component produces the query proposals $\boldsymbol{P}_i \in \mathbb{R}^{N_p \times 2}$ with features $\{\tilde{F}_i^l\}_{l=1}^{3}$ and F_e, where N_p is the proposal number. **(ii)** The deformable feature alignment component first builds the naive fusion view $\tilde{\mathcal{F}}$ to provide global object-related information, then acquires the initial aligned feature $\tilde{Z}_i$ via critical semantics aggregation. **(iii)** The region cross-attention reinforcement component applies the features $\tilde{Z}_i$ and $\tilde{F}_i$ to perform feature refinement, generating the aligned feature Z_i for feature fusion. We will detail the implementation of the above three proposed components in Sections 3.2, 3.3, and 3.4.

3.2 Importance-guided Query Proposal

The query proposal phase aims to predict potential foreground regions, which will be queried in the feature alignment phase. We present an Importance-guided Query Proposal (IQP) component, which generates the importance map with space-channel semantics and applies multi-scale views to alleviate the noise interference for query selection. Unlike Where2comm [7], which relies on trained decoders for query selection, IQP employs space-channel saliency and multiscale solutions to accurately predict foreground regions under noisy conditions.

Importance Map Generation. To highlight object-related foreground regions within the features, we utilize a two-stage scheme to generate importance maps by exploring inter-pixel spatial interactions and intra-pixel channel semantics. First, a convolution-based decoder $f_{dec}(\cdot)$ is used to capture informative regions and yield the corresponding spatial confidence map M_i^s of $\tilde{F}_i$. Meanwhile, the channel-wise average pooling operation $\Psi_m(\cdot)$ integrates the visual semantics of different channels and produces the salient activation map M_i^c. The above schemes are formulated as:

$$M_i^s = \sigma(f_{dec}(\tilde{F}_i)),\ M_i^c = \sigma(\Psi_m(\tilde{F}_i)), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation. The importance map is obtained by these two maps as $M_i = M_i^s \odot M_i^c \in [0,1]^{H\times W}$, where $\odot$ denotes the element-wise multiplication. M_i reflects the perceptually critical level of each location within the collaborator feature $\tilde{F}_i$.

Query Proposal Selection. Since high-resolution perception views are sensitive to feature spatial misalignment [28], the single-scale importance maps inevitably induce erroneous foreground region predictions. To tackle this issue, we leverage multi-scale collaborator features to generate the importance maps $\{M_i^l\}_{l=1}^3$ with varying perception resolutions, which can correct initial importance estimations. Moreover, the corresponding importance map M_e of the ego feature F_e is also obtained to complement possible occlusions in collaborators' local observations. Given the above four importance maps, we synchronize their spatial dimensions by bilinear interpolation and obtain the corresponding pixel-wise average as $\tilde{M}_i$. Then, the heuristic selection approach $f_{sel}(\cdot;\delta)$ applies a predefined threshold δ to select the top $\delta\%$ of spatial locations from $\tilde{M}_i$ as query proposals $\boldsymbol{P}_i$, which is formulated as $\boldsymbol{P}_i = f_{sel}(\tilde{M}_i;\delta) \in \mathbb{R}^{N_p\times 2}$.

3.3 Deformable Feature Alignment

After the query selection phase, we introduce a Deformable Feature Alignment (DFA) component tailored for explicit feature alignment, which enhances the representation of query proposals with contextual visual clues and provides aligned properties. Considering the inability of static local attention [7,34] to cope with irregular feature misalignment and sparse BEV features, the deformable attention is adopted to build query-aware spatial associations and efficiently aggregate perceptually relevant semantics. DFA consists of the following two stages.

Object-related Token Generation. Previous deformable operations [32, 53] generate tokens with query features to learn sampling offsets and attention scores, which is not applicable to collaborative perception scenarios since single collaborator features fail to provide global context information. Accordingly, we propose to build a naive fusion view to provide global object-related semantics for token generation and mitigate feature misalignment in collaborators' local observations. Specifically, all features are first projected to the common space via type-dependent linear layers $\mathrm{LN}_t(\cdot)$. The intuition is that agent discrepancies in sensor characteristics worsen feature mismatch, whereas the type-dependent projection can capture agent-specific attributes and bridge feature heterogeneity [36]. Then, we employ a general fusion operation $f_{fus}(\cdot)$ (*e.g.*, convolution [27] and attention [37]) to generate the naive fusion view $\tilde{\mathcal{F}}$ as follows:

$$\tilde{\mathcal{F}} = f_{fus}(\mathrm{LN}_t(F_e),\ \mathrm{LN}_t(\tilde{F}_i)) \in \mathbb{R}^{C\times H\times W}. \tag{4}$$

As Fig. 3 shows, we extract the corresponding features of the query proposals $\boldsymbol{P}_i$ from $\tilde{\mathcal{F}}$ and $\tilde{F}_i$ and concatenate them along the channel dimension. The linear layer $\mathrm{LN}(\cdot)$ fuses the concatenated features to augment the object-related information within the tokens. The generation process of object-related tokens $\boldsymbol{T}_i$ is formulated as follows:

$$\boldsymbol{T}_i = \mathrm{LN}(\tilde{\mathcal{F}}(\boldsymbol{P}_i) \parallel \tilde{F}_i(\boldsymbol{P}_i)) \in \mathbb{R}^{N_p\times C}, \tag{5}$$

where $\parallel$ denotes the concatenation operation. $\boldsymbol{T}_i$ integrates the global semantics in the naive fusion view and the local context from the collaborators, facilitating subsequent attention learning to produce more robust aligned views.

Deformable Feature Aggregation. We first project $\boldsymbol{T}_i$ as positional encoding and obtain the token embedding as $\tilde{\boldsymbol{T}}_i = \mathrm{LN}(\boldsymbol{T}_i) + \boldsymbol{T}_i$. To aggregate comprehensive and multi-grained perceptual information, the multi-scale collaborator feature $\{\tilde{F}_i^l\}_{l=1}^3$ and naive fusion view $\tilde{\mathcal{F}}$ are utilized as attending features. Subsequently, the token embedding $\tilde{\boldsymbol{T}}_i$ is passed into a linear layer to learn the sampling offset map $\Delta\boldsymbol{P}_i$ for each attending feature, providing the 2D spatial offsets $\{\Delta p_m \mid 1 \le m \le N_m\}$ for each query proposal p, where N_m denotes the sampled key point number. We leverage the learned offset maps to sample key points on the four attending features and extract these key points' features as $\{\tilde{F}_i^l(\boldsymbol{P}_i + \Delta\boldsymbol{P}_i)\}_{l=1}^3$ and $\tilde{\mathcal{F}}(\boldsymbol{P}_i + \Delta\boldsymbol{P}_i)$, providing alignment properties for queries and alleviating local misalignment. The attention scores of the sampled key points are obtained as $\phi(\mathrm{LN}(\tilde{\boldsymbol{T}}_i))$, where $\phi(\cdot)$ denotes the softmax function.

Given the sampled features and learned attention scores, we formulate the enhanced feature of each query proposal p as follows:

$$\mathrm{DFA}(p) = \sum_{h=1}^{H} \boldsymbol{W}_h[\sum_{l=1}^{4}\sum_{m=1}^{N_m} \phi(\mathrm{LN}(\tilde{\boldsymbol{T}}_i(p)))\tilde{F}_i^l(p+\Delta p_m)], \tag{6}$$

where h indexes the attention head, $\boldsymbol{W}_h$ is the learnable weight, and $\tilde{\mathcal{F}}(p+\Delta p_m)$ is rewritten as $\tilde{F}_i^4(p+\Delta p_m)$ only for formula simplification. From Fig. 3, to produce the aligned feature $\tilde{Z}_i$, the augmented feature $\mathrm{DFA}(p)$ is filled into $\tilde{F}_i$ based

on the 2D location of query proposal p. Previous deformable attention-based efforts [41,45] learn offsets and attention scores only with the local information of the ego agent, which inevitably leads to sub-optimal feature aggregation. In contrast, NEAT optimizes the offset and attention learning process through tokens $\tilde{Z}_i$ with global semantics. Also, NEAT significantly alleviates the computation overhead by reducing the sampled key point number N_m (from 15 to 5).

3.4 Region Cross-attention Reinforcement

To generate the aligned feature Z_i through global representation refinement, we introduce a novel Region Cross-attention Reinforcement (RCR) component to diffuse locally-enhanced features in $\tilde{Z}_i$ to the surroundings and aggregate the relevant semantics from $\tilde{F}_i$. With the guidance of confidence priors, RCR filters out the irrelevant regions through region-wise correlation computation, reducing the attending range of global pixel-to-pixel attention adaptively. In contrast, global attention leads to excessive computation cost, while per-location attention [37] fails to achieve global feature refinement through a limited interaction range.

Specifically, the features $\tilde{Z}_i$ and $\tilde{F}_i$ are first concatenated as $Z_i^r = \tilde{Z}_i \| \tilde{F}_i \in \mathbb{R}^{C\times 2H\times W}$. Then, we partition $\tilde{Z}_i$ and Z_i^r into non-overlapping regions with size (S_h, S_w) and apply the 1×1 convolution $\omega_{1*1}(\cdot)$ to obtain the *query*, *key*, *value* embeddings as $Q_i = \omega_{1*1}(\tilde{Z}_i) \in \mathbb{R}^{hw\times S_h S_w \times C}$ and $K_i, V_i = \omega_{1*1}(Z_i^r) \in \mathbb{R}^{2hw\times S_h S_w \times C}$, where $h = \frac{H}{S_h}$ and $w = \frac{W}{S_w}$. The region-wise *query* and *key* embeddings are produced with the per-region average pooling operation $\Psi_a(\cdot)$:

$$Q_i^r = \Psi_a(Q_i \odot M_i^s),\ K_i^r = \Psi_a(K_i \odot M_i^r), \tag{7}$$

where the confidence maps $M_i^s/M_i^r = \sigma(f_{dec}(\tilde{Z}_i/Z_i^r))$ are employed as prior information to introduce object-related semantics during region selection. To select semantically related regions, we compute the correlation matrix by matrix multiplication between Q_i^r and transposed K_i^r and apply the selection function $f_{sel}(\cdot;\delta_r)$ to select the top δ_r regions. After that, as Fig. 3 shows, the features of the selected regions are gathered to form the *key* and *value* embeddings $\{K_i^g, V_i^g\}$. The above process is summarized as follows:

$$K_i^g, V_i^g = \text{gather}(f_{sel}(Q_i^r (K_i^r)^T; \delta_r)) \in \mathbb{R}^{hw\times \delta_r S_h S_w \times C}. \tag{8}$$

Ultimately, we adopt region cross-attention to reinforce the representation of each pixel within $\tilde{Z}_i$ and produce the aligned feature Z_i as follows:

$$Z_i = \phi(\frac{Q_i (K_i^g)^T}{\sqrt{C}}) V_i^g \in \mathbb{R}^{hw\times S_h S_w \times C}. \tag{9}$$

We replace $\tilde{F}_i$ with the reshaped feature $Z_i \in \mathbb{R}^{C\times H\times W}$ and pass Z_i into the subsequent feature fusion stage. With the above RCR component, Z_i aggregates perceptually relevant semantics from the global context and mitigates the local spatial misalignment. Accordingly, Z_i provides more comprehensive and accurate spatial information and facilitates more robust feature fusion than feature $\tilde{F}_i$.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We conduct extensive experiments on four collaborative perception benchmark datasets, including V2XSet [36], OPV2V [37], OPV2V Culver City [37], and V2V4Real [35]. **V2XSet** is a large-scale simulation dataset supporting multi-agent V2X perception. This dataset contains 11,447 labeled point cloud frames, split into training/validation/testing sets with 6,694, 1,920, and 2,833 frames. The simulation dataset **OPV2V** provides raw sensor measurements from 2 to 7 autonomous vehicles in each scene. There are 10,914 frames of point clouds and RGB images, and the training, validation, and testing splits include 6,764, 1,981, and 2,169 frames. The testing dataset **OPV2V Culver City** contains approximately 600 point cloud frames with 3D annotations. **V2V4Real** is the first real-world dataset for V2V collaborative perception, where two self-driving vehicles are configured to record the surrounding environments with multi-modal sensors. To evaluate the object detection performance, we adopt the Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 as experimental metrics following [10].

Implementation Details. The proposed models are implemented using Pytorch toolbox [20] and trained with Nvidia Tesla V100 GPUs by Adam optimizer [9]. We set the default noisy setting to simulate the realistic noise levels, where the localization and heading errors are sampled from Gaussian distributions with a standard deviation of 0.2 m and 0.2°, and the transmission delay is 100 ms. The batch sizes on the V2XSet, OPV2V, and V2V4Real datasets are $\{3, 3, 5\}$, and epoch numbers are $\{15, 15, 10\}$. The initial learning rate is 2e-3, decaying every ten epochs with a factor of 0.1. We implement the regression and classification decoders using two convolutional layers to produce predictions and use smooth absolute error loss and focal loss [15] for objective optimization. $f_{dec}(\cdot)$ reuses the parameters of the classification decoder and utilizes a maximum pooling operation to reduce the channel dimension. The selection thresholds δ on the V2XSet, OPV2V, and V2V4Real datasets are $\{25, 25, 35\}$. We employ the per-location attention [37] to implement the fusion method $f_{fus}(\cdot)$. The sampled key point number N_m is 5, and the attention head H is 8. The RCR component sets the region size (S_h, S_w) and threshold δ_r as $(4, 8)$ and 2, respectively. We provide the evaluation results on the testing sets of the four datasets.

Model Zoo. To evaluate the effectiveness of NEAT, we select five representative models with distinct structures as the comparison baselines. Specifically, **F-Cooper** [3] employs a heuristic max-out strategy to fuse collaborator features. **CORE** [27] designs a feature-level reconstruction task to preserve critical semantics and designs an attention-aware collaboration component. **AttFuse** [37] uses the per-location attention to learn the attention weights of corresponding pixels across different feature maps. **Where2comm** [7] presents a confidence-aware per-location attention design to facilitate pragmatic collaboration. **CoBEVT** [34] applies axial attention to aggregate meaningful object representations from local windows and sparse global tokens.

Table 1: Comparisons of space-time complexity and performance on the four datasets. The complexity results are reported in the average parameter number (Para.) and MACs across datasets. The detection results are reported in AP@0.5/0.7.

Models	Para. (M)	MACs (G)	V2XSet		OPV2V		V2V4Real		OPV2V Culver City	
			AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Fusion	6.37	30.46	60.60	40.22	68.71	48.66	39.78	22.02	55.7	47.10
Late Fusion	6.37	31.19	54.93	30.74	76.24	56.14	49.41	22.05	60.18	39.23
When2com [16]	11.31	186.46	68.84	42.93	72.73	54.80	51.82	24.11	61.73	40.61
V2VNet [29]	14.45	435.62	79.11	49.28	77.58	62.27	57.16	26.73	71.46	51.50
DiscoNet [13]	9.63	147.85	79.83	54.16	80.66	63.82	55.94	26.61	70.62	50.36
V2X-ViT [36]	15.39	306.78	83.59	61.44	84.76	68.92	55.35	28.72	73.48	54.04
SCOPE [45]	19.71	355.45	84.26	62.17	85.02	71.13	56.28	28.04	72.27	53.66
FFNet [48]	13.74	251.02	78.76	55.49	83.26	65.58	57.02	28.55	70.51	52.83
CoBEVFlow [31]	15.56	295.27	82.94	58.35	85.18	69.24	59.37	30.63	74.06	54.37
F-Cooper [3]	8.21	160.25	71.47	46.92	79.28	58.46	53.19	25.93	68.68	48.02
F-Cooper + **NEAT**	9.50	165.41	**83.03**	**60.94**	**86.91**	**67.81**	**56.25**	**32.70**	**74.91**	**53.28**
CORE [27]	11.14	172.53	73.89	43.24	78.85	56.17	57.65	29.74	68.95	45.65
CORE + **NEAT**	12.51	178.08	**76.35**	**48.06**	**81.14**	**60.45**	**60.91**	**34.19**	**70.63**	**48.48**
AttFuse [37]	8.53	115.54	70.85	48.66	80.23	59.82	55.86	27.88	70.90	51.26
AttFuse + **NEAT**	9.85	120.98	**80.30**	**55.97**	**84.37**	**68.19**	**60.60**	**34.24**	**75.22**	**55.21**
Where2comm [7]	10.36	210.92	81.98	53.43	80.55	61.92	58.37	31.49	70.32	50.17
Where2comm + **NEAT**	11.64	216.67	**83.42**	**58.72**	**82.06**	**64.84**	**59.28**	**35.43**	**73.57**	**53.29**
CoBEVT [34]	12.03	238.14	81.12	54.30	81.32	63.19	58.20	28.72	71.54	52.33
CoBEVT + **NEAT**	13.37	244.75	**85.07**	**65.16**	**90.01**	**75.41**	**60.55**	**34.97**	**78.82**	**58.10**

4.2 Detection Performance Comparison

Table 1 presents the average space-time complexity and detection performance comparison between various baselines and the NEAT-based models under default noisy settings across the four datasets. The complexity is measured by the mainstream Parameter Number (M) and MACs (G). No Fusion represents the ego-centered perception without collaboration. Late Fusion refers to aggregating all agents' output boxes and producing the results by non-maximum suppression. Moreover, we consider representative and reproducible state-of-the-art (SOTA) methods, including When2com [16], V2VNet [29], V2X-ViT [36], DiscoNet [13], SCOPE [45], FFNet [48] and CoBEVFlow [31]. The key observations are listed below. **(i)** NEAT significantly improves the five selected models on all metrics with very slight complexity costs. For instance, the NEAT-based F-Cooper, CORE, AttFuse, Where2comm, and CoBEVT achieve an average gain of 8.85%, 4.10%, 6.50%, 3.81%, and 8.78% compared to their vanilla counterparts across the four datasets concerning AP@0.7. **(ii)** Compared to V2X-ViT [36], SCOPE [45] and FFNet [48] that implicitly mitigate the feature misalignment by integrated frameworks, the models equipped with NEAT achieve better performance. Specifically, the NEAT-based CoBEVT enables the method CoBEVT to achieve SOTA performance on all four datasets with a slight extra computation cost. The reasonable explanations include: (1) The importance-aware NEAT alleviates misalignment conditions by aggregating semantically relevant object representations, which is applicable to methods with diverse structures; (2) our plugin can flexibly address varying feature mismatches in real-world (*e.g.*, V2V4Real) and simulated scenes via global feature refinement.

4.3 Robustness to Localization Error

Evaluating the collaboration performance under diverse pose error levels is essential in assessing the system robustness. Fig. 4 show the detection precision

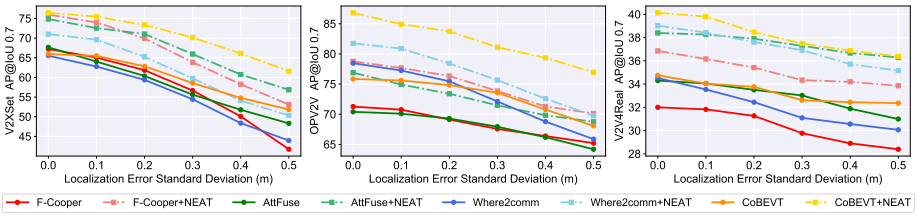


Fig. 4: Robustness assessment of the localization error on the three datasets.

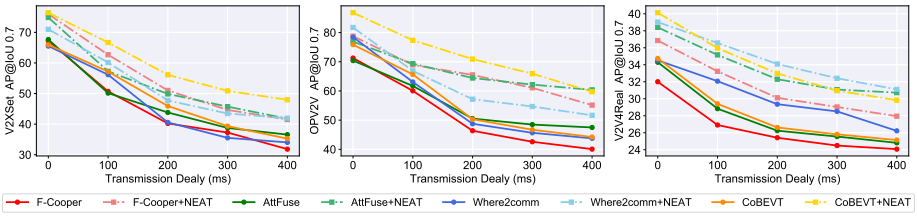


Fig. 5: Robustness assessment of the transmission delay on the three datasets.

of the four models before and after equipping the NEAT plugin under varying pose noises. The localization error is sampled from Gaussian distributions with a standard deviation of $\sigma_{xyz} \in [0, 0.5]$ m. **(i)** Intuitively, due to the flawed inter-agent spatial transformation caused by pose errors, the performance curves of all intermediate fusion methods show an evident downward trend with increasing error levels. In comparison, our NEAT plugin alleviates the performance degradation and significantly improves the vanilla baselines at all error levels. **(ii)** For instance, Where2comm [7] on V2XSet degrades to similar results as No Fusion under severe localization errors. In this case, NEAT mitigates the feature misalignment and enhances the corresponding performance favorably. These findings verify that NEAT can empower existing models with robust resistance to noise interference by aligning collaborator features.

4.4 Robustness to Transmission Delay

As Fig. 1b shows, temporal asynchrony caused by transmission delay also induces deleterious feature misalignment, resulting in subsequent two-side fusion errors among agents. Fig. 5 provides the robustness analysis of vanilla baselines and NEAT-based models against varying delays. Inevitably, performance deterioration spreads across all vanilla methods as transmission delay continuously increases. In particular, the AP@0.7 result of F-Cooper [3] on OPV2V drops substantially, even lower than the baseline No Fusion (AP@0.7 = 48.66%) when the delay exceeds 300 ms. Conversely, NEAT provides significant gains for different models at all delay levels. That is, the NEAT-based models maintain high detection precision compared to the vanilla counterparts, even under a se-

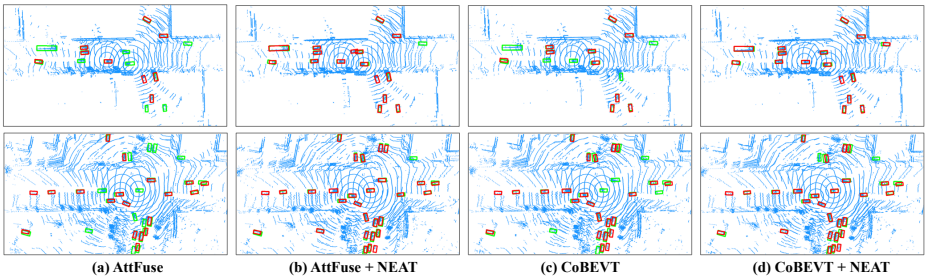


Fig. 6: Qualitative comparison results in real-world scenarios from the V2V4Real dataset [35]. Green and red boxes denote ground truths and detection results, respectively. Compared to the vanilla baselines, NEAT-based models significantly improve the detection results.

vere delay (*i.e.*, 400 ms). Our plugin's merit stems from the query-aware spatial associations to compensate for temporal asynchrony.

4.5 Qualitative Evaluation

To intuitively evaluate the merits of NEAT, we select two challenging scenarios on V2V4Real to present the detection visualizations of vanilla baselines and NEAT-based models. In Fig. 6, AttFuse [37] and CoBEVT [34] exhibit vulnerable performance in noisy settings due to massive missing predicted bounding boxes. Through the proposed plugin, the NEAT-based models produce more detection results that are well aligned with the ground truths. These observations confirm that NEAT improves the detection robustness of collaborative perception systems under noisy conditions.

4.6 Comparison of Plugin Methods

We compare the existing plugins SyncNet [11] and CoAlign [18] on V2XSet, OPV2V, and V2V4Real datasets to explore the effectiveness of NEAT. As shown in Table 2, four models are selected as baselines to observe the detection performance changes, including CoBEVT [34], AttFuse [37], F-Cooper [3], and CORE [27]. NEAT provides more significant and consistent performance gains across different baselines by jointly addressing the pose errors and transmission delay. In contrast, CoAlign and SyncNet offer sub-optimal solutions due to their noise-specific designs and inadequate improvements. In practice, our NEAT introduces extremely low spatio-temporal complexity overhead (*i.e.*, Para. = 1.31 M and MACs = 5.11 G) compared to the existing plugins, including CoAlign and SyncNet. Specifically, these two plugins bring average parameter numbers of 2.94/3.80M and MACs of 26.51/31.02G across datasets, respectively.

Table 2: Comparison results of plugin methods on the V2XSet, OPV2V, and V2V4Real datasets. We provide four models for different plugins as baselines for use in assembly, including CoBEVT, AttFuse, F-Cooper, and CORE. NEAT brings more significant gains to vanilla models on all metrics.

Models	V2XSet	OPV2V	V2V4Real
CoBEVT	81.12/54.30	81.32/63.19	58.20/28.72
CoBEVT + SyncNet	82.28/57.75	83.77/67.42	58.74/31.75
CoBEVT + CoAlign	83.15/60.20	85.65/69.21	59.05/31.22
CoBEVT + NEAT	**85.07/65.16**	**90.01/75.41**	**60.55/34.97**
AttFuse	70.85/48.66	80.23/59.82	55.86/27.88
AttFuse + SyncNet	75.07/52.34	81.92/62.77	58.14/30.79
AttFuse + CoAlign	76.82/51.90	82.53/64.03	57.32/31.37
AttFuse + NEAT	**80.30/55.97**	**84.37/68.19**	**60.60/34.24**
F-Cooper	71.47/46.92	79.28/58.46	53.19/25.93
F-Cooper + SyncNet	77.29/52.76	83.16/62.45	54.78/27.76
F-Cooper + CoAlign	76.73/53.42	82.83/63.16	55.02/29.64
F-Cooper + NEAT	**83.03/60.94**	**86.91/67.81**	**56.25/32.70**
CORE	73.89/43.24	78.85/56.17	57.65/29.74
CORE + SyncNet	75.23/45.53	79.82/56.96	58.37/31.82
CORE + CoAlign	74.41/45.77	80.07/57.54	58.56/31.41
CORE + NEAT	**76.35/48.06**	**81.14/60.45**	**60.91/34.19**

Table 3: Ablation study results of the proposed components on the V2XSet, OPV2V, and V2V4Real datasets. **FA**: deformable feature aggregation; **TG**: object-related token generation; **MG**: importance map generation; **PS**: query proposal selection; **RCR**: region cross-attention reinforcement.

CoBEVT [34] + NEAT					V2XSet	OPV2V	V2V4Real
FA	TG	MG	PS	RCR			
					81.12/54.30	81.32/63.19	58.20/28.72
✔					82.69/58.24	84.41/67.05	58.95/31.81
✔	✔				83.10/59.97	85.74/69.30	59.43/32.45
✔	✔	✔			83.56/62.08	87.26/72.16	59.87/33.50
✔	✔	✔	✔		83.94/63.25	88.33/73.58	60.02/34.33
✔	✔	✔	✔	✔	**85.07/65.16**	**90.01/75.41**	**60.55/34.97**

AttFuse [37] + NEAT					V2XSet	OPV2V	V2V4Real
FA	TG	MG	PS	RCR			
					70.85/48.66	80.23/59.82	55.86/27.88
✔					74.24/51.48	81.95/62.66	57.44/29.96
✔	✔				76.27/52.41	82.36/64.05	58.35/31.47
✔	✔	✔			77.35/54.02	82.95/65.51	59.23/32.64
✔	✔	✔	✔		78.64/54.63	83.14/66.38	59.82/33.22
✔	✔	✔	✔	✔	**80.30/55.97**	**84.37/68.19**	**60.60/34.24**

4.7 Ablation Studies

We select two representative models to perform thorough ablation studies on V2XSet, OPV2V, and V2V4Real to verify the necessity of the proposed components. As Table 3 shows, the vanilla baselines without any components in NEAT are employed as the performance reference. On this basis, we progressively add (1) FA, (2) TG, (3) MG, (4) PS, and (5) RCR and present the corresponding AP@0.5/0.7 results. **(i)** The continuously improved detection performance over the three datasets reveals that each component provides indispensable contributions. **(ii)** Particularly, deformable feature aggregation brings noteworthy gains by aggregating vital context semantics with corrective mismatch attributes, while object-related token generation facilitates attention learning by introducing global object-related semantics. **(iii)** Furthermore, region cross-attention reinforcement efficiently diffuses the locally enhanced representations with region-wise attention, achieving global feature refinement and significant improvements.

5 Conclusion

This paper proposes NEAT, a lightweight plugin, to address the feature misalignment issue induced by collaboration noises in multi-agent perception systems. NEAT accomplishes the explicit alignment of collaborator-shared features with customized components, ensuring subsequent high-quality collaboration and perception. Extensive experiments show that NEAT can be readily integrated into existing methods to provide a generic solution for improving system robustness.

Acknowledgement. This work is supported in part by the China Mobile Research Fund of the Chinese Ministry of Education under Grant KEH2310029 and in part by the Specific Research Fund of the innovation Platform for Academicians of Hainan Province under Grant YSPTZX202314. This work is also supported in part by the Shanghai Key Research Laboratory of NSAI, the Joint Laboratory on Networked Al Edge Computing, Fudan University-Changan, the National Natural Science Foundation of China (Grant No. 62250410368), and the Kunshan Government Research Fund 24KKSGR024.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1090–1099 (2022) 5
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020) 5
3. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: 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. pp. 88–100 (2019) 2, 10, 11, 12, 13
4. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems. pp. 514–524. IEEE (2019) 2
5. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F., Zhou, B., Zhao, H.: Autoalign: pixel-instance feature aggregation for multi-modal 3d object detection. arXiv preprint arXiv:2201.06493 (2022) 5
6. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F.: Deformable feature aggregation for dynamic multi-modal 3d object detection. In: European Conference on Computer Vision. pp. 628–644. Springer (2022) 5
7. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in Neural Information Processing Systems **35**, 4874–4886 (2022) 2, 4, 6, 7, 10, 11, 12
8. Hu, Y., Lu, Y., Xu, R., Xie, W., Chen, S., Wang, Y.: Collaboration helps camera overtake lidar in 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2023) 2, 4
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015) 10
10. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019) 10
11. Lei, Z., Ren, S., Hu, Y., Zhang, W., Chen, S.: Latency-aware collaborative perception. In: European Conference on Computer Vision. pp. 316–332. Springer (2022) 3, 4, 13
12. Li, F., Zeng, A., Liu, S., Zhang, H., Li, H., Zhang, L., Ni, L.M.: Lite detr: An interleaved multi-scale encoder for efficient detr. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 18558–18567 (2023) 5
13. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021) 2, 4, 11
14. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision. pp. 1–18. Springer (2022) 5
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017) 10
16. Liu, Y.C., Tian, J., Glaser, N., Kira, Z.: When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4106–4115 (2020) 11
17. Lu, Y., Hu, Y., Zhong, Y., Wang, D., Chen, S., Wang, Y.: An extensible framework for open heterogeneous collaborative perception. arXiv preprint arXiv:2401.13964 (2024) 4
18. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023) 3, 4, 6, 13
19. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021) 5
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019) 10
21. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017) 1
22. Ram, T., Chand, K.: Effect of drivers' risk perception and perception of driving tasks on road safety attitude. Transportation Research Part F: Traffic Psychology and Behaviour **42**, 162–176 (2016) 1
23. Rawashdeh, Z.Y., Wang, Z.: Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In: 2018 21st International Conference on Intelligent Transportation Systems. pp. 3961–3966. IEEE (2018) 2
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016) 1
25. Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.S., Zhao, M.J.: Improving 3d object detection with channel-wise transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2743–2752 (2021) 5
26. Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., Ouyang, Z.: Vips: Real-time perception fusion for infrastructure-assisted autonomous driving. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. pp. 133–146 (2022) 2

27. Wang, B., Zhang, L., Wang, Z., Zhao, Y., Zhou, T.: Core: Cooperative reconstruction for multi-agent perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8710–8720 (October 2023) 2, 8, 10, 11, 13
28. Wang, T., Chen, G., Chen, K., Liu, Z., Zhang, B., Knoll, A., Jiang, C.: Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8187–8196 (October 2023) 2, 4, 6, 7
29. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: Proceedings of the European Conference on Computer Vision. pp. 605–621. Springer (2020) 2, 6, 11
30. Wang, Z., Fan, S., Huo, X., Xu, T., Wang, Y., Liu, J., Chen, Y., Zhang, Y.Q.: Vimi: Vehicle-infrastructure multi-view intermediate fusion for camera-based 3d object detection. arXiv preprint arXiv:2303.10975 (2023) 4
31. Wei, S., Wei, Y., Hu, Y., Lu, Y., Zhong, Y., Chen, S., Zhang, Y.: Robust asynchronous collaborative 3d detection via bird's eye view flow. arXiv preprint arXiv:2309.16940 (2023) 2, 4, 11
32. Xizhou, Z., Weijie, S., Lewei, L., Bin, L., Xiaogang, W., Jifeng, D.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021) 5, 8
33. Xu, R., Chen, W., Xiang, H., Xia, X., Liu, L., Ma, J.: Model-agnostic multi-agent perception framework. In: 2023 IEEE International Conference on Robotics and Automation. pp. 1471–1478. IEEE (2023) 2
34. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In: Conference on Robot Learning (2022) 2, 4, 7, 10, 11, 13, 14
35. Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., Meng, Z., Xiang, H., Dong, X., Song, R., 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. pp. 13712–13722 (2023) 2, 10, 13
36. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: Proceedings of the European Conference on Computer Vision. pp. 107–124. Springer (2022) 2, 4, 6, 8, 10, 11
37. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: Proceedings of the International Conference on Robotics and Automation. pp. 2583–2589. IEEE (2022) 2, 6, 8, 9, 10, 11, 13, 14
38. Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L.: Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). pp. 1642–1651 (2022) 5
39. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., Wang, Y., Liu, J., Zhang, P., Zhai, P., Zhang, L.: Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20459–20470 (October 2023) 1
40. Yang, D., Yang, K., Li, M., Wang, S., Wang, S., Zhang, L.: Robust emotion recognition in context debiasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12447–12457 (2024) 5

41. Yang, D., Yang, K., Wang, Y., Liu, J., Xu, Z., Yin, R., Zhai, P., Zhang, L.: How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. Advances in Neural Information Processing Systems **36** (2024) 2, 9
42. Yang, K., Liu, J., Yang, D., Wang, H., Sun, P., Zhang, Y., Liu, Y., Song, L.: A novel efficient multi-view traffic-related object detection framework. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2023) 1
43. Yang, K., Sun, P., Lin, J., Boukerche, A., Song, L.: A novel distributed task scheduling framework for supporting vehicular edge intelligence. In: 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS). pp. 972–982. IEEE (2022) 5
44. Yang, K., Sun, P., Yang, D., Lin, J., Boukerche, A., Song, L.: A novel hierarchical distributed vehicular edge computing framework for supporting intelligent driving. Ad Hoc Networks **153**, 103343 (2024) 5
45. Yang, K., Yang, D., Zhang, J., Li, M., Liu, Y., Liu, J., Wang, H., Sun, P., Song, L.: Spatio-temporal domain awareness for multi-agent collaborative perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23383–23392 (October 2023) 2, 4, 6, 9, 11
46. Yang, K., Yang, D., Zhang, J., Wang, H., Sun, P., Song, L.: What2comm: Towards communication-efficient collaborative perception via feature decoupling. In: Proceedings of the 31th ACM International Conference on Multimedia (ACM MM). p. 7686–7695 (2023) 5
47. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., 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. pp. 21361–21370 (2022) 2
48. Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., Nie, Z.: Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. Advances in Neural Information Processing Systems **36** (2024) 2, 4, 6, 11
49. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., et al.: V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5486–5495 (2023) 2
50. Yuan, X., Kortylewski, A., Sun, Y., Yuille, A.: Robust instance segmentation through reasoning about multi-object occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11141–11150 (2021) 2
51. Yuan, Z., Song, X., Bai, L., Wang, Z., Ouyang, W.: Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2068–2078 (2021) 2
52. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017) 1
53. Zhou, Z., Zhao, X., Wang, Y., Wang, P., Foroosh, H.: Centerformer: Center-based transformer for 3d object detection. In: European Conference on Computer Vision. pp. 496–513. Springer (2022) 5, 8