Multi-agent Collaborative Perception via Motion-aware Robust Communication Network

Shixin Hong[1], Yu Liu[2]†, Zhi Li[1], Shaohui Li[1], You He[2]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Department of Electronic Engineering, Tsinghua University
hsx22@mails.tsinghua.edu.cn, liuyu77360132@126.com,
{zhilizl, lishaohui}@sz.tsinghua.edu.cn, heyou@mail.tsinghua.edu.cn

Abstract

Collaborative perception allows for information sharing between multiple agents, such as vehicles and infrastructure, to obtain a comprehensive view of the environment through communication and fusion. Current research on multi-agent collaborative perception systems often assumes ideal communication and perception environments and neglects the effect of real-world noise such as pose noise, motion blur, and perception noise. To address this gap, in this paper, we propose a novel motion-aware robust communication network (MRCNet) that mitigates noise interference and achieves accurate and robust collaborative perception. MRCNet consists of two main components: multi-scale robust fusion (MRF) addresses pose noise by developing cross-semantic multi-scale enhanced aggregation to fuse features of different scales, while motion enhanced mechanism (MEM) captures motion context to compensate for information blurring caused by moving objects. Experimental results on popular collaborative 3D object detection datasets demonstrate that MRCNet outperforms competing methods in noisy scenarios with improved perception performance using less bandwidth. Our code will be released at https://github.com/IndigoChildren/collaborative-perception-MRCNet.

1. Introduction

Collaborative perception enables information sharing among multiple agents, including vehicles and infrastructure, to provide a comprehensive view of the environment through information communication and fusion [37]. This helps individual autonomous vehicles to tackle challenging perception scenarios, such as physical occlusion or long-range detection [27]. Collaborative perception technology has gained significant attention due to its advantages [20].

Previous works on multi-agent collaborative perception systems focus on improving perception performance by designing superior fusion modules [2, 14, 29, 35] and effective communication strategies [8, 24, 33]. For example, the CoBEVT [39] framework uses sparse transformers [32] to capture local and global spatial interactions across agents. The UMC [34] framework optimises communication by using a two-stage entropy-based selection mechanism. However, these frameworks assume ideal communication and perception environments [7], while real-world scenarios are affected by noise, such as pose noise in vehicle localization, perception noise in raw data [21], and the motion blur in fast-moving object detection. These noises cannot be ignored for real-world applications of multi-agent collaborative systems. Unfortunately, studies on robust collaborative perception with real-world noise are largely limited.

In fact, different forms of noise in the real world can affect the performance of collaborative perception systems in different ways. For example, when noise interferes with the positioning system, it affects the accuracy of pose information. Pose information is vital for fusing the perception data (*e.g.*, raw data, intermediate features, or detection outputs) of the collaborative agents in different coordinates [40]. Thus, pose noise can lead to information misalignment and inconsistency among agents [26] during data fusion, resulting in a reduced system performance. Perception noise introduces distortions into the sensor data, which affects the final performance of the perception system. In addition, the presence of motion blur due to rapid object movement affects the accuracy of target detection [1, 9, 15, 30].

However, the development of a robust collaborative perception system against real-world noise is complicated by the following challenges. First, a robust fusion strategy that adequately processes the multi-source perception data is required to deal with the perceptual misalignment caused by pose noise. Second, a feature selection strategy that filters out highly informative features is crucial to reduce bandwidth consumption and improve the efficiency of multi-agent collaboration [28]. Finally, it is desirable to find a sta-

†Corresponding author: Yu Liu.

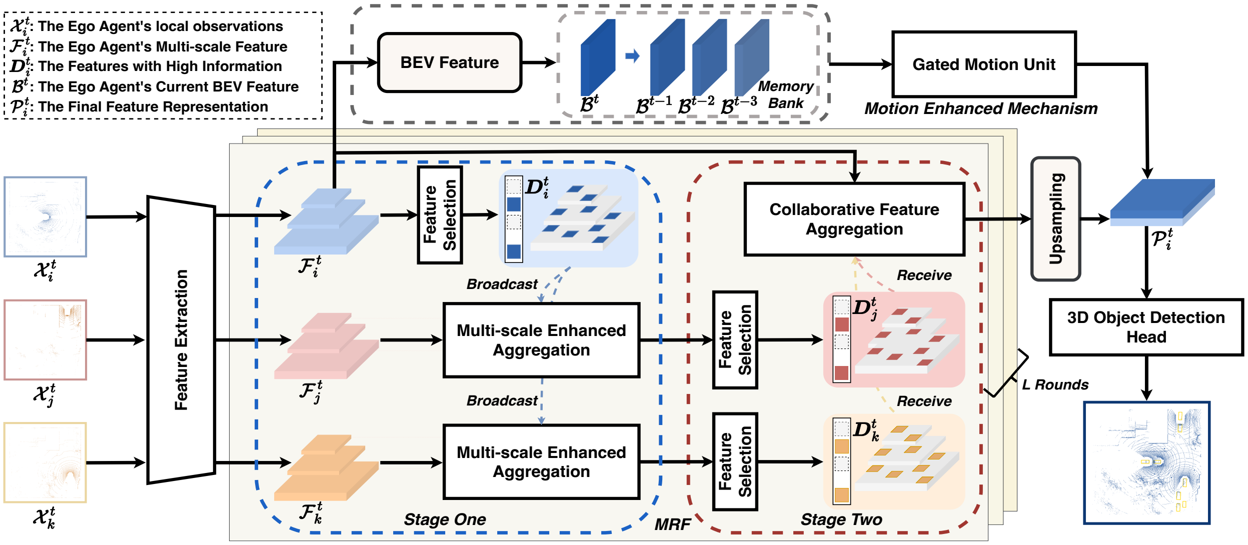


Figure 1. The overview of our proposed MRCNet framework. In MRCNet, agents exchange highly informative semantic features in two stages through multiple rounds of communication to mitigate the effect of noise. Meanwhile, the ego agent uses motion context extracted from historical sequences in the memory bank to reduce the effect of motion blur from moving objects.

ble and accurate feature extraction that eliminates the motion blur of rapidly moving objects.

To address the aforementioned challenges, we present MRCNet, a motion-aware robust communication network that mitigates noise interference and achieves more robust and accurate collaborative perception. As shown in Figure 1, MRCNet facilitates effective information collaboration among agents and improves the perception performance of the ego agent through two main components, *i.e.*, multi-scale robust fusion (MRF) and motion enhanced mechanism (MEM). Specifically, for the challenge of feature misalignment due to collaborators' pose noise, MRF improves the robustness and performance of collaborative perception through multi-scale semantic feature fusion. For perception noise, we filter out highly informative semantic features for multi-round communication based on channel-wise selection operations, while reducing communication bandwidth consumption. In addition, to address the problem of motion blur due to object movement, we introduce the MEM module, which is based on the gate recurrent unit (GRU) architecture and is designed to capture and incorporate the motion context of the ego agent. This mechanism aggregates historical information to refine the features of the current frame. To validate the effectiveness of our MRCNet framework, we carry out a comprehensive evaluation on three open-source collaborative 3D object detection datasets, including V2XSim [13], OPV2V [41], and V2XSet [40]. The experimental results show that MRCNet outperforms competing methods in noisy scenarios. Our main contributions can be summarised as follows:

- We present a focused study of the real-world noise problem in the collaborative perception task.
- We propose a motion-aware collaborative perception framework, *i.e.*, MRCNet, which presents a novel multi-scale feature communication and fusion strategy to improve the robustness of collaboration.
- Collaborative experiments are conducted on three large open-source datasets, proving that our proposed MRCNet can achieve the best perception performance in noisy scenarios with less bandwidth.

2. Related Work

2.1. Communication in Collaborative Perception

Previous works [8, 23, 24] focus on solving the communication problem, as excessive communication bandwidth consumption can lead to communication congestion and latency. To address this challenge, most research considers how to achieve a better trade-off between performance and bandwidth. There are two main methods to address the challenge of bandwidth constraints: compressing the communication volume or reducing the number of communicators. Several works reduce the communication volume by using learnable methods to facilitate efficient collaboration [18]. Where2comm [8] focuses on perceptually critical areas and transmits foreground features with high confidence. UMC [34] designs a two-stage communication strategy using extended entropy theory to filter out low quality regions. As for selecting of effective collaborators, Who2Com [24] proposes a multi-stage handshake communication mechanism that determines which agents to connect with, and Select2Col [19] introduces a collaborator selection method that selects contributing collaborators with low latency. However, there is still a lack of research on the problem of pose noise in collaborative communication. Therefore, we propose the multi-scale robust fusion module, which selects highly informative features from multi-scale features for communication and aggregation, thereby reducing the effect of multi-source noise in collaboration.

2.2. Information Fusion in Collaborative Perception

In addition to addressing communication issues, collaborative perception systems focus on designing collaboration modules to fuse information from multiple agents. The goal is to improve perception capabilities and thereby optimise performance in downstream tasks. Based on the stage of data sharing and collaboration, the collaborative perception can be broadly divided into three categories [14, 40], *i.e.*, early fusion [42], intermediate fusion [8, 23, 35, 41], late fusion [26, 31]. Early fusion involves sharing raw data for a comprehensive view, but requires significant bandwidth and overlooks contextual information. Intermediate fusion extracts intermediate features from each agent's observations before transmission, balancing perception performance and communication bandwidth usage. Late fusion shares detection outputs among agents, but produces suboptimal results due to individual inaccuracies or incomplete perceptions. Among them, intermediate fusion methods gain popularity due to their performance-bandwidth tradeoff [35, 40, 41]. V2VNet [35] proposes a spatially aware graph neural network to aggregate features from multiple agents. OPV2V [41] uses a global self-attention module for feature fusion. As an extension of intermediate fusion, our study introduces a novel cross-semantic fusion module, which is designed to aggregate features in the presence of feature misalignment caused by pose noise. Furthermore, we integrate motion context into the perception system to improve the understanding of the environment.

3. Method

In this paper, we focus on improving the robustness and accuracy of collaborative perception in complex collaborative scenarios. To achieve this, we propose a novel motion-aware robust communication network (MRCNet) that integrates multi-scale feature selection, effective information fusion, and motion-aware feature extraction into collaborative perception. As shown in Figure 1, MRCNet consists of four key components: feature extraction and selection, multi-scale robust collaboration, motion enhanced mechanism, and the 3D object detection head.

3.1. Problem Formulation

At time t, in a collaborative perception task involving N agents, the selected i-th agent is defined as the ego agent, and a communication graph is established among them. Other agents within the communication range serve as collaborators, providing complementary information and helping the ego agent to make a comprehensive perception. Let $\mathcal{X}_i^t$ be the point cloud observation and $\mathcal{Y}_i^t$ be the perception supervision of the ego agent. Each collaborator provides the ego agent with the information extracted from their local observations $\mathcal{X}_j^t$. Additionally, considering that we aim at modelling the motion-aware collaborations, we design a memory bank of the ego agent that stores K frames of historical point data, denoted as $\left\{\mathcal{X}_i^{t-k}\right\}_{k=1}^{K} = \left\{\mathcal{X}_i^{t-1}, \mathcal{X}_i^{t-2}, ..., \mathcal{X}_i^{t-K}\right\}$.

Thus, the motion-aware collaborative perception task can be defined as follows:

$$\hat{\mathcal{Y}}_i^t = \sum_{i=1}^{N} \left\{ \mathcal{Y}_i^t | \Omega\left(\mathcal{X}_i^t, \left\{\mathcal{X}_j^t\right\}_{j\neq i}^{N}, \left\{\mathcal{X}_i^{t-k}\right\}_{k=1}^{K} \right) \right\}, \quad (1)$$

where $\Omega\left(\cdot\right)$ is the proposed MRCNet. In this paper, we use the 3D object detection task to evaluate the performance of collaborative perception. Therefore, $\hat{\mathcal{Y}}_i^t$ are defined as the predicted detection boxes along this line.

3.2. Feature Extraction and Selection

Feature Extraction. After the ego agent constructs a vehicle-to-everything (V2X) collaboration graph connecting neighbouring agents at the current frame t, it exchanges pose data between agents for alignment purposes. The collaborators then transform their raw point cloud data into the ego agent's coordination using the transformation matrix $\mathbf{\Lambda}$ computed based on the agents' unique pose $\boldsymbol{\xi}^t$. Next, each agent uses a shared feature extractor consisting of S layers [41] to obtain multi-scale features from the raw data. This process yields a set of features $\mathcal{F}_i^t = \left\{\mathcal{F}_i^{s,t}\right\}_{s=1}^{S}$ for the i-th agent, where i ranges from 1 to N.

Feature Selection. In contrast to previous methods that transmit all features, we introduce a novel feature selection technique. Perception noise introduces distortions into the collaborative features, thereby misleading the collaborative system. Therefore, our feature selection technique selects highly informative features from the multi-scale features for communication, effectively reducing the effect of noise interference and conserving communication bandwidth.

The features extracted from the foreground regions are particularly important due to their rich semantic information [43,45]. To select highly informative features for effective communication, we use channel pooling operations to evaluate the semantic density within each feature. Initially, different scales of features are spatially flattened and then concatenated into a feature sequence $\boldsymbol{E}^t \in \mathbb{R}^{C\times L}$, where C is the number of channels of the feature, L is the length of the multi-scale feature after flatten. Subsequently, we perform channel-wise pooling operations on $\boldsymbol{E}^t$ to generate the semantic confidence map $\boldsymbol{m}$:

$$\boldsymbol{m} = \left(\mathrm{MaxPool}\left(\boldsymbol{E}^t\right) + \mathrm{AvgPool}\left(\boldsymbol{E}^t\right)\right) \in \mathbb{R}^{1\times L}, \quad (2)$$

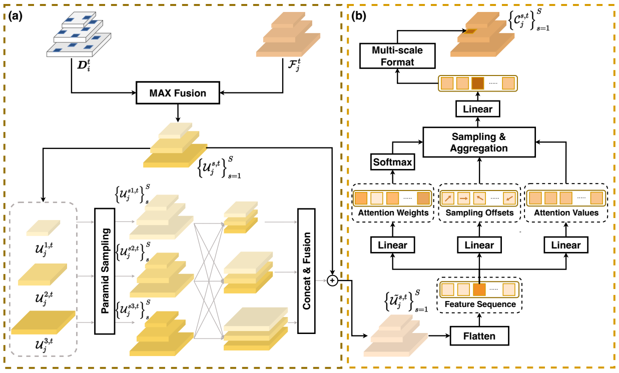


Figure 2. The architecture of the proposed MEA component. (a) shows the improved pyramid sampling aggregation module. (b) shows the multi-scale deformable self-attention module.

where MaxPool $(\cdot)$ and AvgPool $(\cdot)$ refer to the channel-wise max pooling and average pooling operations. The binary selection matrix for communication is defined as:

$$\boldsymbol{c} = \Phi_{\text{select}}\left(\boldsymbol{m}\right) \in \{0,1\}^{1\times L}, \tag{3}$$

where $\Phi_{\text{select}}\left(\cdot\right)$ is the selection function to select the most informative features for communication. The selected features with high information are then obtained as $\boldsymbol{D}^t = \left(\boldsymbol{c} \odot \boldsymbol{E}^t\right) \in \mathbb{R}^{C\times L}$, where $\odot$ is the element-wise multiplication. The selected feature map provides highly informative features with less bandwidth usage.

3.3. Multi-Scale Robust Fusion

Cross-agent collaboration improves the visual representation and perception performance of the ego agent by exploiting shared semantic information. Previous attention-based [39–41] approaches overlook to consider the robustness of the collaboration system to noise. Therefore, we introduce a novel multi-scale robust fusion (MRF) module, which is designed for the robust multi-scale feature fusion through two-stage communication. Multi-scale feature fusion combines information of different semantics to improve perception capability and robustness to pose noise. The two-stage communication module consists of multi-scale enhanced aggregation (MEA) and collaborative features aggregation (CFA). MEA uses cross-semantic fusion to overcome the challenge of feature misalignment caused by pose noise, and CFA enables the ego agent to maximise the perception information aggregation.

Multi-scale Enhanced Aggregation. In the first stage, the ego agent broadcasts highly informative features to the collaborators. For the sake of clarity, let's consider a communication process involving only two agents. Upon receiving highly informative features, the collaborator first converts these features from sequence to multi-scale representation, then performs a preliminary fusion using the max fusion operation to obtain the multi-scale feature $\mathcal{U}_j^t = \left\{\mathcal{U}_j^{s,t}\right\}_{s=1}^S$. To address the challenge of feature misalignment, MEA incorporates cross-semantic modules to fully exploit the strengths of different semantics [6]. As shown in Figure 2, MEA consists of two components: an improved pyramid sampling aggregation (IPSA) module and a multi-scale deformable self-attention (MDSA) module [46].

Low-level features capture fine-grained details, while high-level semantic features provide integrated contextual understanding. The pyramid sampling aggregation module [12, 25] uses simple convolution and channel-wise concatenation operations to fuse multi-scale features. This process gathers information through cross-semantic fusion, resulting in a more comprehensive and robust feature representation that effectively mitigates noise interference. We pyramid sample the feature $\mathcal{U}_j^{s,t}$ to obtain the new multi-scale feature $\left\{\mathcal{U}_j^{sl,t}\right\}_{l=1}^S$, where $\mathcal{U}_j^{sl,t}$ is the transformation of $\mathcal{U}_j^{s,t}$ from the s-th scale to the l-th scale, and the feature sizes of the new multi-scale feature correspond to $\mathcal{U}_j^t$.

After multiple rounds of communication, the fine details in the transmitted features tend to be compromised due to cross-semantic fusion during the pyramid sampling. Therefore, we improve the pyramid sampling aggregation module by adding original multi-scale features from collaborators to compensate for detailed features, which is the proposed IPSA module. The output can be represented as:

$$\tilde{\mathcal{U}}_j^{s,t} = \text{conv}_{1\times 1}\left(\left[\mathcal{U}_j^{s1,t}, \mathcal{U}_j^{s2,t}, \mathcal{U}_j^{s3,t}\right]\right) + \mathcal{U}_j^t, s \in [1,2,3], \tag{4}$$

where $[\cdot,\cdot,\cdot]$ is the channel-wise concatenation, and $\text{conv}_{1\times 1}\left(\cdot\right)$ is the 1×1 convolution operation.

Following the IPSA module, the collaborator applies the MDSA [46] module to further fuse multi-scale features and mitigate the effect of pose noise on feature distortion. The MSDA module consists of three linear layers $\boldsymbol{W}_a$, $\boldsymbol{W}_v$, $\boldsymbol{W}_o$, which separately compute the sampling offset Δq, the attention weight $\boldsymbol{A}$ and the value $\boldsymbol{V}$ at position q, where q is the two-dimensional reference point in $\tilde{\mathcal{U}}_j^{s,t}$. The output of MSDA at position q can be expressed as follows:

$$\begin{aligned} &\text{MSDA}\left(q\right) = \sum_{h=1}^{H} \boldsymbol{W}_h \cdot \\ &\left(\sum_{s=1}^{S}\sum_{p=1}^{P} \psi\left(\boldsymbol{W}_a \tilde{\mathcal{U}}_j^{s,t}\left(q\right)\right) \cdot \boldsymbol{W}_v\left(\tilde{\mathcal{U}}_j^{s,t}\left(q + \boldsymbol{W}_o \tilde{\mathcal{U}}_j^{s,t}\left(q\right)\right)\right)\right) \\ &= \sum_{h=1}^{H} \boldsymbol{W}_h \cdot \left(\sum_{s=1}^{S}\sum_{p=1}^{P} \boldsymbol{A}^{hsp}\left(q\right) \cdot \mathcal{V}^{hs}\left(q + \Delta^{hsp}\left(q\right)\right)\right), \end{aligned} \tag{5}$$

where H is the number of attention heads, S is the quantity of different feature scales, and P is the number of sampling points utilized in each scale. $\boldsymbol{W}_h$ is the projection matrix and the softmax function $\psi(\cdot)$ is used to determine the attention weight. Ultimately, we output the multi-scale format feature $\mathcal{C}_j^t = \left\{\mathcal{C}_j^{s,t}\right\}_{s=1}^{S}$ after MSDA.

Collaborative Features Aggregation. In the second stage of communication, the j-th collaborator uses the same feature selection technique to send highly informative features $\boldsymbol{D}_j^t = \left(\boldsymbol{m}_j \odot \boldsymbol{E}_j^t\right)$ back to the ego agent, where $\boldsymbol{E}_j^t$ is the feature sequence representation of $\mathcal{C}_j^t$. In order to maximise the aggregation of the semantic information provided by the collaborator, inspired by [29], the ego agent uses the spatial-wise adaptive fusion module on each scale of features. The ego agent first stacks the feature maps of the agents to form $\mathcal{A}_i^t = \left\{\mathcal{A}_i^{s,t}\right\}_{s=1}^{S}, \mathcal{A}_i^{s,t} \in \mathbb{R}^{n \times C \times H_s \times W_s}$, where s is the scale index and n is the maximum number of agents. Max and average pooling is applied to $\mathcal{A}_i^{s,t}$ over the first channel axis to create new feature maps $\mathcal{A}_{i,\text{max}}^{s,t} \in \mathbb{R}^{1 \times C \times H_s \times W_s}$ and $\mathcal{A}_{i,\text{avg}}^{s,t} \in \mathbb{R}^{1 \times C \times H_s \times W_s}$, where H_s and W_s represent the height and width of the feature. These pooled feature maps are concatenated and passed through a 3D convolutional layer with a ReLU activation function to perform further fusion. The aggregated features provide complementary information to the ego agent.

3.4. Motion Enhanced Mechanism

As mentioned above, the collaborative perception module increases the range of perception and provides compensation for occlusions [41], but it cannot directly address the motion blur introduced by moving objects. To combat this, we introduce the motion enhanced mechanism (MEM), which uses motion context from historical point cloud frames to achieve motion-aware perception.

When the ego agent initiates a collaboration request, it projects the last K frames of the point cloud stored in the memory bank to its current coordinate, producing a series of bird's eye view (BEV) feature maps $\left\{\mathcal{B}^{t-1}, ..., \mathcal{B}^{t-K}\right\}$. Inspired by [4, 22], we design a recurrent unit called the gated motion unit (GMU). This unit aggregates motion context to enhance frame-level features. Figure 3 illustrates the operation of GMU at timestamp t, showing that GMU takes three inputs stored in the ego agent's memory bank: the feature $\mathcal{B}^{t-1}$ of the previous frame, the feature $\mathcal{B}^t$ of the current frame, and the enhanced feature $\mathcal{H}^{t-1}$. Both $\mathcal{B}^0$ and H^0 are initialized as $\mathcal{B}^1$, and the output $\mathcal{H}^t$ represents the enhanced feature of the current frame after passing through GMU.

Between successive frames, object motion causes spatial transitions of features. Therefore, we use $\mathcal{M}^t = \mathcal{B}^t - \mathcal{B}^{t-1}$ to capture motion context, which can represent the motion clues of moving objects [17]. To reduce the effect of motion blur and to exploit historical information, we use a deformable convolution network [3]. The deformable convolution network can better capture the features of objects by adaptively adjusting the kernel size and shape. Specifically, the 3×3 convolutional layer is used on the motion context $\mathcal{M}_t$ to obtain the spatial offsets Δp_n^t. The convolutional operation is then applied to $\mathcal{B}^{t-1}$ to extract historical features at position p. The output feature map $\mathcal{I}^t$ at location p is:

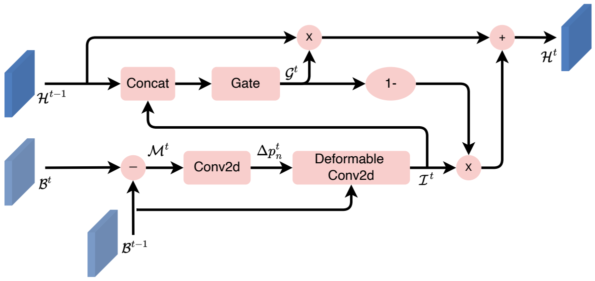


Figure 3. The architexture of the proposed GMU component.

$$\mathcal{I}^t(p) = \sum_{p_n \in \mathcal{R}} \boldsymbol{W}(p_n) \cdot \mathcal{B}^{t-1}\left(p + p_n + \Delta p_n^t\right), \tag{6}$$

where $\boldsymbol{W}$ is the convolution weight, $\mathcal{R}$ is the kernel grid and p_n enumerates the locations in $\mathcal{R}$.

To reduce the motion blur effect caused by the moving object between frames, it is necessary to determine whether the features at each position should be updated. Inspired by [36], we design an update gate $\mathcal{G}$ using the channel-spatial attention module. The channel-spatial attention module is applied to the stack feature of $\mathcal{I}^t$ and $\mathcal{H}^{t-1}$. After applying the sigmoid operation, the overall attention map is normalised to a range of values between 0 and 1, as formulated below:

$$\mathcal{G} = \sigma\left(\Gamma\left(\left[\mathcal{I}^t, \mathcal{H}^{t-1}\right]\right)\right) \in \mathbb{R}^{C \times H \times W}, \tag{7}$$

where $[\cdot, \cdot]$ is the channel-wise concatenation, $\Gamma(\cdot)$ is the channel-spatial attention module, and $\sigma(\cdot)$ is the sigmoid function for value normalisation. The value in $\mathcal{G}$ indicates the probability that the corresponding position of $\mathcal{H}^{t-1}$ is reserved. Then the improved feature $\mathcal{H}^t$ is refined with the previous improved feature $\mathcal{H}^{t-1}$ as follows:

$$\mathcal{H}^t = \mathcal{G} \odot \mathcal{H}^{t-1} + (1 - \mathcal{G}) \odot \mathcal{I}^t, \tag{8}$$

where $\mathcal{H}^t \in \mathbb{R}^{C \times H \times W}$ is the final output of the GMU.

3.5. 3D Object Detection Head

After obtaining the fused output after MRF, the multi-scale feature is encoded to the same size and concatenated to $\mathcal{Z}_i^t \in \mathbb{R}^{C \times H \times W}$. Then, we concatenate $\mathcal{Z}_i^t$ and $\mathcal{H}^t$ along the channel dimension to derive the final representation $\mathcal{P}_i^t$. To generate the final object detection results, we then use a dual-branch convolutional module on $\mathcal{P}_i^t$ to perform classification and regression tasks for each detection anchor. Similar to the previous study [44], the formulation for the object detection loss L_{det} is expressed as:

$$L_{\text{det}} = L_{\text{cls}} + \beta L_{\text{reg}}, \tag{9}$$

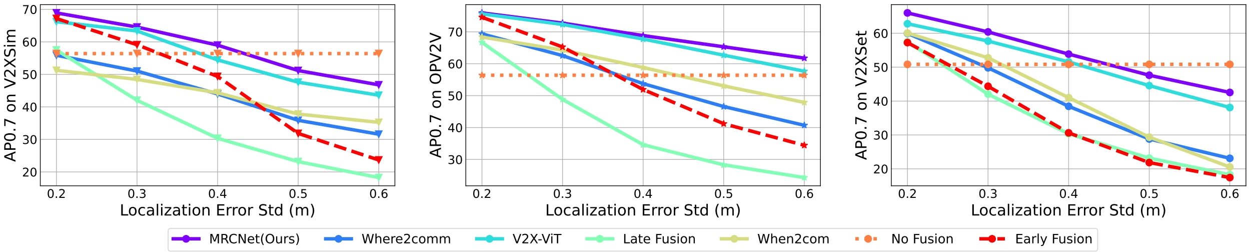


Figure 4. Robustness of our proposed method and compared benchmark to varying levels of localization noise.

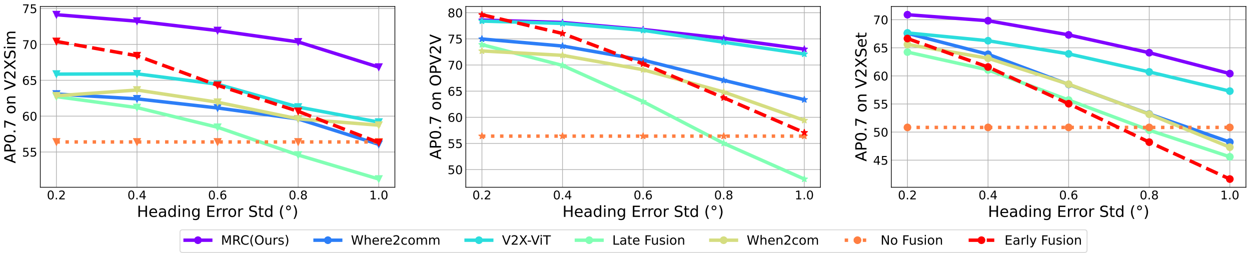


Figure 5. Robustness of our proposed method and compared benchmark to varying levels of heading noise.

where L_{cls} is the cross-entropy loss for object classification, L_{reg} is the smooth L1 loss for box regression [16], and β is the balancing hyperparameter.

4. Experiments

To assess the robustness of our module in noisy collaboration scenarios, we conduct experiments on three large open-source datasets. In particular, our experiments focus on LiDAR-based 3D object detection. We measure detection performance using average precision (AP) at intersection-over-union (IoU) thresholds of 0.50 and 0.70, and note that the AP values are given as percentages. To ensure a fair comparison, we exclude data and feature compression operations in the compared methods.

4.1. Datasets and Experimental Settings

Datasets. V2XSim [13] is a V2X collaborative perception dataset co-simulated by SUMO [10] and CARLA [5]. It contains 10000 frames of LiDAR point clouds and 501K annotated 3D boxes. There are 8000/1000/1000 frames for the training/validation/testing, respectively. The detection range of the agent is $x \in [-32m, 32m], y \in [-32m, 32m]$. **OPV2V** [41] is a vehicle-to-vehicle collaborative perception dataset co-simulated by OpenCDA [38] and CARLA [5], including 11464 frames of 3D point clouds with 230K annotated 3D boxes, split into training/validation/testing of 6764/1981/2719 frames. The detection range is $x \in [-140m, 140m], y \in [-40m, 40m]$. **V2XSet** [40] is the first open dataset that includes both V2X cooperation and realistic noise simulation, which is co-simulated by OpenCDA [38] and CARLA [5]. There are a total of 11447 frames in the dataset, and the train/validation/testing splits are 6694/1920/2833, respectively. The detection range setting is the same as OPV2V.

Implementation Details. In our implementation, we use PointPillars [11] as the multi-scale feature extractor, with the grid size of $(0.4m, 0.4m)$ for the point cloud discretisation. And we get intermediate features in three scales, which is $S = 3$ as we define in the method. The MSDA module uses 8 attention heads and 4 sampling points per communication round. For V2XSim, we extract information from three historical frames, while for V2XSet and OPV2V, we use two historical frames. To train the detection model, we set the hyperparameter in the loss function to $\beta = 2$ according to [11]. We use the Adam optimizer with an initial learning rate of $\{1e-3, 5e-4, 3e-4\}$ for the OPV2V, V2XSet and V2XSim datasets. The learning rate is gradually reduced at epochs 35 and 45 on the V2XSim dataset using a decay factor of 0.1. For the OPV2V and V2XSet datasets, the decay occurs at epochs 30 and 40. All models are trained on 2 NVIDIA RTX 3090 GPUs with a batch size of 2 for a total of 50 epochs.

Benchmark Comparison. Our considering benchmark includes early fusion, intermediate fusion, and late fusion techniques as well as single agent perception for comparison. In addition, we compare with six state-of-the-art (SOTA) intermediate fusion models: When2Com [23], V2VNet [35], AttFuse [41], V2X-ViT [40], DiscoNet [14], CoBEVT [39], and Where2Comm [8].

4.2. Quantitative Results

The pose of the agent can be represented by a 6D vector. Since the agent only has yaw angle to measure rotation,

Table 1. Performance comparison on the V2XSim, OPV2V, and V2XSet datasets. The results are reported in AP0.5/0.7.

Model	V2XSim		OPV2V		V2XSet	
	AP0.5	AP0.7	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	65.73	52.57	69.38	56.40	64.88	50.83
Late Fusion	71.22	56.99	82.76	61.64	76.31	53.33
Early Fusion	84.94	65.97	89.27	72.96	83.40	51.71
When2com [23]	73.36	52.52	82.78	67.66	74.73	40.62
V2VNet [35]	82.70	65.31	88.69	74.93	84.79	64.73
AttFuse [41]	81.70	66.24	88.54	72.91	84.37	66.27
V2X-ViT [40]	82.32	64.41	86.74	75.70	82.42	63.14
DiscoNet [14]	83.56	66.12	88.05	72.07	82.34	64.79
CoBEVT [39]	81.00	65.06	88.99	72.80	84.84	65.14
Where2comm [8]	77.57	57.02	86.58	68.97	80.83	58.72
MRCNet	**85.33**	**69.82**	**89.77**	**76.12**	**85.00**	**66.31**

we simplify the pose to $\boldsymbol{\xi} = (x, y, \theta)$ in 2D space [26], where x, y and θ represent the 2D center position and the yaw angle of the agent's accurate global position, respectively. To simulate pose noise in real-world scenarios, we add Gaussian noise with a standard deviation of $0.2m$ and $0.2°$ to the collaborators' poses in the datasets. We compare the performance of 3D object detection with other methods on the datasets, as shown in Table 1. According to the experimental results, despite the introduction of pose noise, our proposed MRCNet module shows superior performance compared to the benchmark on all datasets.

In addition, to investigate the effect of different components of pose noise on the collaborative perception system, we conduct experiments to evaluate the robustness of the models to localization noise and heading noise, separately. Experiments are also performed to assess the effect of various levels of localization and heading noise. Specifically, we add Gaussian noise with standard $\sigma_{xy} \in [0.2m, 0.6m]$, $\sigma_\theta \in [0.2°, 1.0°]$ separately to the collaborators' poses. As shown in the Figure 4 and the Figure 5, an increase in Gaussian noise leads to a rapid decrease in AP0.7 values for the other methods. This suggests that misaligned spatial features from collaborators can mislead the collaboration system. However, our method maintains superior performance even under significant pose noise, demonstrating its ability to effectively select and fuse multi-level semantic features while achieving robustness to pose noise.

4.3. Component Analysis

Contribution of Major Components. To evaluate the effectiveness of our proposed component, we progressively remove (i) MEM, (ii) FS (feature selection), (iii) IPSA and (iii) MSDA and present the detection precision. Table 2 shows the results of our ablation studies, with each innovative component contributing to the performance gains.

Table 2. Results of the ablation study of the proposed core components on the V2XSim, OPV2V and V2XSet datasets.

MEM	FS	ISPA	MSDA	V2XSim		OPV2V		V2XSet	
				AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
				80.77	64.28	85.78	74.01	83.12	63.88
			✓	81.46	67.64	86.29	74.63	83.44	64.71
		✓	✓	82.87	69.15	87.32	75.31	84.21	65.31
	✓	✓	✓	83.60	69.21	88.11	75.77	84.43	65.47
✓	✓	✓	✓	**85.33**	**69.82**	**89.77**	**76.12**	**85.00**	**66.31**

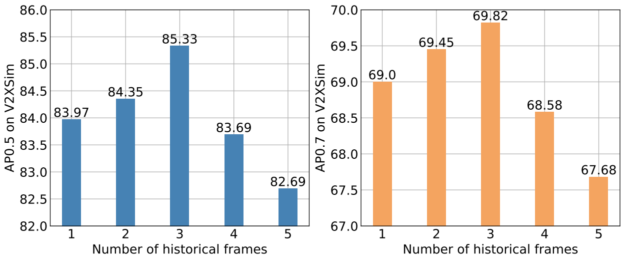
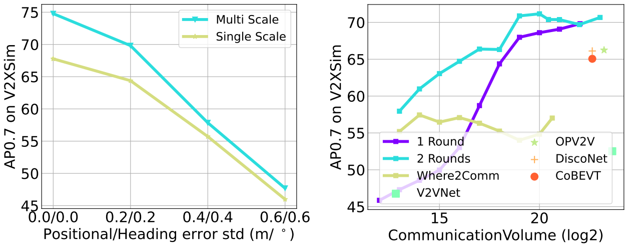


Figure 6. Collaborative Performance with varying number of history frames. The MEM module works best when fusing three history frames on the V2XSim dataset, and two frames on the OPV2V and V2XSet datasets.



(a)

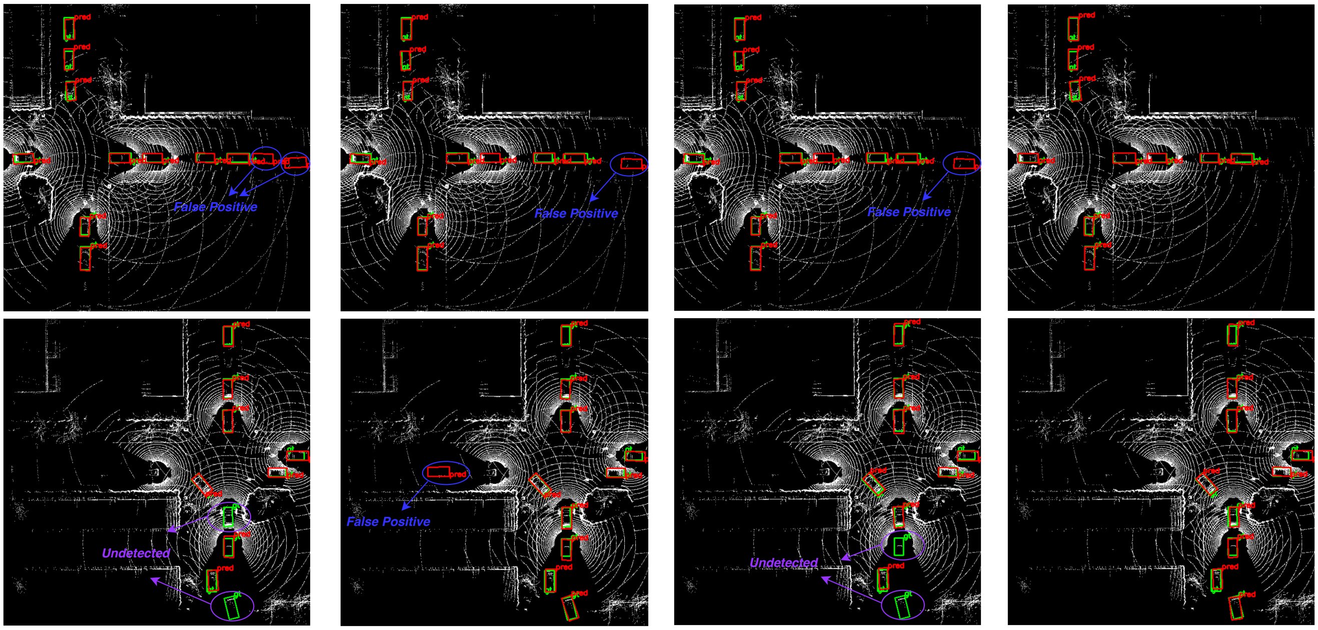
(b)

Figure 7. (a) shows the perception performance comparison when using the proposed multi-scale method versus the single-scale approach. (b) is the comparison of collaborative perception performance of our proposed method and other baseline methods with different communication volume.

Varying Number of History Frames. We evaluate the effect of integrating different numbers of history frames from the ego agent. As shown in Figure 6, integrating multiple frames into the MEM module leads to an improvement in the detection accuracy in the V2XSim [13] dataset. However, once the number of history frames exceeds three, no further improvement is observed, as MEM faces challenges in effectively capturing features with large changes.

Effect of Multi-Scale Collaboration. To evaluate the effectiveness of multi-scale feature collaboration in complex scene perception, we separately evaluate the detection accuracy achieved by single-scale and multi-scale feature collaboration. In our implementation, the single-scale feature collaboration is modified by using only a single-scale deformable self-attention module [46] within MEA. The results of this comparison are shown in Figure 7a. By using multi-scale feature fusion, our method exploits the advantages of different semantics for better noise robustness and achieves an improved detection performance.

Comparison of Communication Volume. Figure 7b shows the performance-bandwidth trade-off of the proposed



(a) V2X-ViT (b) DiscoNet (c) CoBEVT (d) MRCNet (Ours)

Figure 8. In the qualitative comparison on the V2XSim dataset under pose noise, green and red boxes indicate the ground truths and the detection outputs, respectively. Our method outperforms previous SOTA models by achieving more accurate detection results.

MRCNet and the benchmark at different levels of communication bandwidth consumption. No fusion serves as the baseline, as it does not involve any communication. Our analysis reveals several key trends: (i) Where2Comm [8] initially improves detection performance by transmitting the most critical information, but struggles to filter out noisy data as the transmission volume increases. When Where2Comm [8] transmits all feature information, the global attention mechanism captures the spatial information and improves its effectiveness. (ii) Our proposed MRCNet method consistently shows better detection performance at different bandwidth levels, achieving superior results, and multi-round communication can improve the performance of collaborative perception, but excessive misaligned feature transmission can also cause degradation.

4.4. Case Study

Detection Results Visualization. Figure 8 provides a visual comparison of the detection results of different models in real-world scenarios. Our proposed MRCNet outperforms SOTA methods, with well-aligned bounding boxes that closely match the ground truth labels. In contrast, other models such as V2X-ViT [40], DiscoNet [14], and CoBEVT [39] fail to accurately detect fast moving objects accurately, resulting in missed objects or falsely-detected predictions. Our proposed MRCNet exploits the multi-level semantic features of collaborative systems and demonstrates improved robustness against noise interference. In addition, the proposed motion-aware fusion model extracts motion information derived from moving objects, which can effectively mitigate the negative effects of motion blur in complex scenes, thus improving detection accuracy.

5. Conclusion

In this paper, we study the real-world noise problem in multi-agent collaborative perception systems. We propose a novel framework called motion-aware robust communication network (MRCNet) to mitigate noise interference and achieve robust collaborative perception. MRCNet uses two-stage communication to enable the selection, communication and fusion of multi-level semantic features among agents. To address the problem of motion blur caused by fast-moving objects, a GRU-based module is proposed to capture the motion context and to refine the current BEV feature. Comprehensive experiments on three open-source datasets show that the proposed MRCNet outperforms existing competing methods in noisy scenarios. The further consideration of communication issues and model heterogeneity will be a future direction.

Acknowledgements. This work was partially supported by grants from the National Natural Science Foundation of China (No.62293544 and 62301299), and China Postdoctoral Science Foundation (No.2022M720077). We would also like to thank Yue Hu from Shanghai Jiao Tong University for sharing the collaborative perception datasets.

References

[1] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiuhua Zhao. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *arXiv preprint arXiv:2303.17099*, 2023. 1

[2] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019. 1

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

[4] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 5

[5] 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

[6] Tianrui Guan, Jun Wang, Shiyi Lan, Rohan Chandra, Zuxuan Wu, Larry Davis, and Dinesh Manocha. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 772–782, 2022. 4

[7] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets and challenges. *arXiv preprint arXiv:2301.06262*, 2023. 1

[8] 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, 3, 6, 7, 8

[9] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 1

[10] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012. 6

[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. 6

[12] Junhyung Lee, Junho Koh, Youngwoo Lee, and Jun Won Choi. D-align: Dual query co-attention network for 3d object detection based on multi-frame point cloud sequence. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9238–9244. IEEE, 2023. 4

[13] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2, 6, 7

[14] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 1, 3, 6, 7, 8

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[17] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788–2802, 2017. 5

[18] Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*, 2023. 2

[19] Yuntao Liu, Qian Huang, Rongpeng Li, Xianfu Chen, Zhifeng Zhao, Shuyuan Zhao, Yongdong Zhu, and Honggang Zhang. Rethinking collaborative perception from the spatial-temporal importance of semantic information. *arXiv preprint arXiv:2307.16517*, 2023. 2

[20] Yu Liu, Zhi Li, Zhizhuo Jiang, and You He. Prospects for multi-agent collaboration and gaming: challenge, technology, and application. *Frontiers of Information Technology & Electronic Engineering*, 23(7):1002–1009, 2022. 1

[21] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 1

[22] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8786–8793, 2019. 5

[23] 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. 2, 3, 6, 7

[24] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative

perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883. IEEE, 2020. 1, 2

[25] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. Tanet: Robust 3d object detection from point clouds with triple attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11677–11684, 2020. 4

[26] 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. 1, 3, 7

[27] Guiyang Luo, Hui Zhang, Quan Yuan, and Jinglin Li. Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3578–3586, 2022. 1

[28] Hangyu Mao, Zhibo Gong, Zhengchao Zhang, Zhen Xiao, and Yan Ni. Learning multi-agent communication under limited-bandwidth restriction for internet packet routing. *arXiv preprint arXiv:1903.05561*, 2019. 1

[29] Donghao Qiao and Farhana Zulkernine. Adaptive feature fusion for cooperative perception using lidar point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1186–1195, 2023. 1, 5

[30] Yao Rong, Xiangyu Wei, Tianwei Lin, Yueyu Wang, and Enkelejda Kasneci. Dynstatf: An efficient feature fusion strategy for lidar 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3237–3246, 2023. 1

[31] Zhiying Song, Fuxi Wen, Hailiang Zhang, and Jun Li. A cooperative perception system robust to localization errors. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–6. IEEE, 2023. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[33] Junyong Wang, Yuan Zeng, and Yi Gong. Collaborative 3d object detection for automatic vehicle systems via learnable communications. *arXiv preprint arXiv:2205.11849*, 2022. 1

[34] Tianhang Wang, Guang Chen, Kai Chen, Zhengfa Liu, Bo Zhang, Alois Knoll, and Changjun Jiang. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. *arXiv preprint arXiv:2303.12400*, 2023. 1, 2

[35] 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, 3, 6, 7

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[37] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3584–3591. IEEE, 2023. 1

[38] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162. IEEE, 2021. 6

[39] 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, 4, 6, 7, 8

[40] 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, 3, 4, 6, 7, 8

[41] 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, 3, 4, 5, 6, 7

[42] Yunshuang Yuan, Hao Cheng, and Monika Sester. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):3054–3061, 2022. 3

[43] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3085–3094, 2019. 3

[44] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5

[45] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018. 3

[46] 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. 4, 7