ERMVP: Communication-Efficient and Collaboration-Robust Multi-Vehicle Perception in Challenging Environments

Jingyu Zhang[1] Kun Yang[1] Yilei Wang[1] Hanqi Wang[1] Peng Sun[2,*] Liang Song[1,*]
[1]Academy for Engineering and Technology, Fudan University [2]Duke Kunshan University
{jingyuzhang22, yileiwang23}@m.fudan.edu.cn, songl@m.fudan.edu.cn

Abstract

*Collaborative perception enhances perception performance by enabling autonomous vehicles to exchange complementary information. Despite its potential to revolutionize the mobile industry, challenges in various environments, such as communication bandwidth limitations, localization errors and information aggregation inefficiencies, hinder its implementation in practical applications. In this work, we propose ERMVP, a communication-**E**fficient and collaboration-**R**obust **M**ulti-**V**ehicle **P**erception method in challenging environments. Specifically, ERMVP has three distinct strengths: i) It utilizes the hierarchical feature sampling strategy to abstract a representative set of feature vectors, using less communication overhead for efficient communication; ii) It employs the sparse consensus features to execute precise spatial location calibrations, effectively mitigating the implications of vehicle localization errors; iii) A pioneering feature fusion and interaction paradigm is introduced to integrate holistic spatial semantics among different vehicles and data sources. To thoroughly validate our method, we conduct extensive experiments on real-world and simulated datasets. The results demonstrate that the proposed ERMVP is significantly superior to the state-of-the-art collaborative perception methods.*

1. Introduction

Autonomous vehicles are widely recognized as a valuable means to enhance road safety and traffic efficiency. Equipped with lidar, cameras, and other sensors, these vehicles are capable of accurately sensing their surroundings to ensure safe and reliable operation. However, the single-vehicle perception system has inevitable drawbacks [26, 39], such as a limited sensor field of view that can be easily obstructed and the challenge of detecting distant objects due to sparse and low-resolution data. Recently, the advances in vehicle-to-vehicle (V2V) communication technologies [12, 17, 36] and deep learning [4–6, 10, 16, 24, 45] have spurred innovation and progress in the collaborative perception technology. This technology allows connected autonomous vehicles (CAVs) to share sensory data, leading to more comprehensive environmental perception.

Although collaborative perception technology shows great potential in transforming the mobility industry, its practical application faces several challenges, including communication bandwidth limitations [15, 35], localization errors [11, 25] and information aggregation inefficiencies [38, 47]. In practical situations, wireless communication resource and reliability constraints severely hamper the efficacy of delay-sensitive collaborative perception. While recent works [15, 42] have achieved a balance between perception performance and communication bandwidth through well-designed mechanisms, these methods have their limitations because they primarily considering information compression over spatial redundancy. This narrow focus exacerbates performance degradation at high compression ratios.

Further, complex dynamic environments lead to localization errors, which result in inaccurate relative transform estimates and spatial feature misalignment. This relative pose noise produces misleading features that adversely affect the effectiveness of collaborative perception. Existing methods [25, 34] attempt to optimize the overall pose through intensive computation, but the high latency makes them unsuitable for real-time dynamic perception. Meanwhile, collaborative methods [3, 38, 41, 42, 46] only focus on aggregated information, but overlook the inherent perceptual strengths of ego vehicle. This paradigm is vulnerable to the perturbations introduced by collaborative noise, including asynchronous motion blur and inaccurate projections. Such drawback becomes a bottleneck for achieving optimal perceptual performance. In contrast, ego-centric features may contain locally accurate spatial location information that is not affected by collaborative noise. Therefore, a priority for establishing a pragmatic collaborative perception system is to effectively overcome the above challenges.

*Corresponding authors. Our code is available at https://github.com/Terry9a/ERMVP.

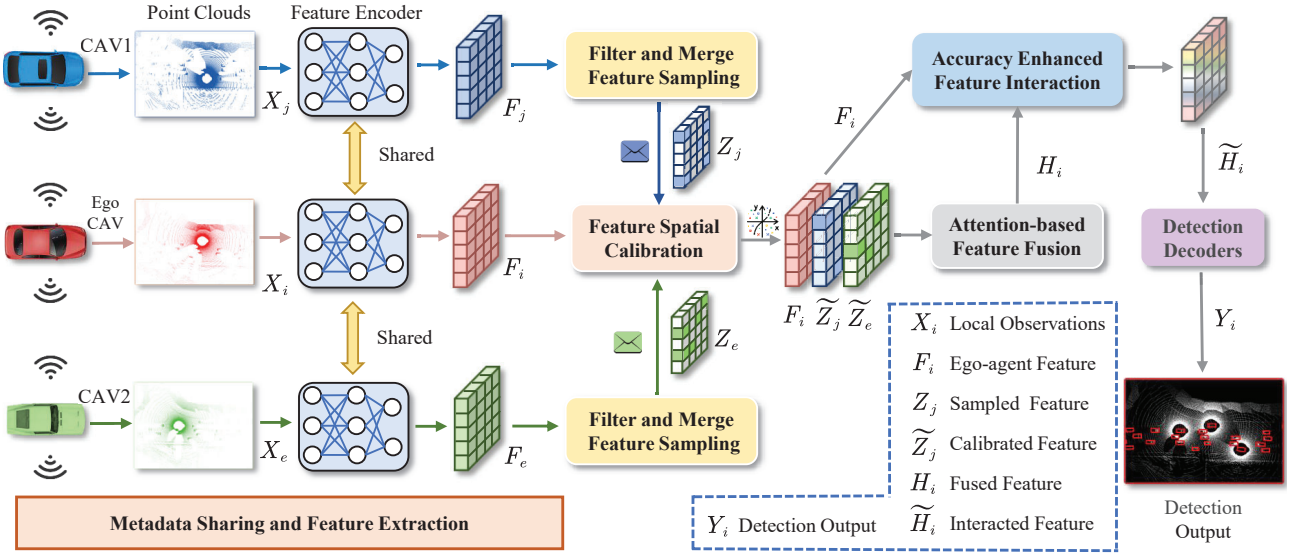


Figure 1. The overall architecture of the proposed framework. The framework consists of six phases: metadata sharing and feature extraction, filter and merge feature sampling, feature spatial calibration, attention-based feature fusion, accuracy enhanced feature interaction and detection decoders. The details of each individual component are illustrated in Section 3.

Based on these observations, we propose ERMVP, a communication-efficient and collaboration-robust multi-vehicle perception method in challenging environments. From Figure 1, ERMVP uses four proposed core components to tackle existing challenges jointly. Specifically, (**i**) we first design an advanced filter and merge feature sampling strategy to tackle the limitations of wireless communication resources. This strategy considers both inter-class and intra-class redundancy relationships to abstract a refined set of feature vectors from the redundant features, using less communication overhead for efficient communication. (**ii**) Second, we introduce a plug-and-play feature spatial calibration module to mitigate the implications of vehicle localization errors. This module ingeniously utilizes consensus sparse foreground features to align the relative pose relationships between ego-vehicle and collaborators without any precise pose supervision. (**iii**) Furthermore, we present a pioneering feature fusion and interaction paradigm to integrate holistic spatial semantics. This paradigm comprises two key components: The first is an attention-based feature fusion module that alternates between local and global attention to fuse heterogeneous information from different vehicles. The second is an accuracy enhanced feature interaction strategy that leverages the accurate positional information inherent in ego-centric features to enhance the rich semantic information provided by fused features. Through these tailored components, ERMVP represents a significant advancement towards pragmatic collaborative perception. To validate the effectiveness of the ERMVP, we conduct extensive experiments on two collaborative 3D object detection datasets including V2V4Real [44] and OPV2V [43]. Comprehensive experimental results demonstrate that our method outperforms previous state-of-the-art methods under the bandwidth-limited noisy setting. The main contributions can be summarized as follows:

- We present ERMVP, a communication-efficient and collaboration-robust multi-vehicle perception method, which addresses the communication bandwidth limitations, localization errors and information aggregation efficiency challenges.
- We develop a filter and merge feature sampling strategy to enhance communication efficiency, a feature spatial calibration module for accurate spatial feature alignment, and two information aggregation components to optimize the fusion process.
- We conduct extensive experiments on both real-world and simulated datasets. The results show the superiority of our method and the necessity of the proposed components.

2. Related Work

2.1. Multi-Agent Communication

Communication has played a pivotal role in the development of robust multi-agent systems. Early multi-agent communication [20, 29, 32] relied on predefined protocols and heuristics to regulate communication between agents. However, these fixed methods are unsuitable in complex and dynamic settings. Advances in deep learning have inspired the development of advanced strategies for information exchange and collaboration among agents. For example, MAGIC [27] used a graph attention encoder to solve the

problem of when and to whom to send messages. TMC [49] utilized temporal smoothing to ensure efficient communication. EC-MARL [2] tackled high-dimensional continuous control and partially observable states by introducing a rapid communication protocol to resolve task dilemmas. In comparison, our research focuses on LiDAR-based collaborative 3D object detection tasks in complex driving scenarios. We propose a hierarchical feature sampling strategy to achieve efficient communication across agents.

2.2. Collaborative Perception

Factors including limited fields of view of sensors and physical obstructions in the environment can adversely impact the perception capabilities of individual agents. To address these challenges, collaborative perception within multi-agent systems has become a key technology. Utilizing newly available datasets [22, 43, 44], several novel collaborative perception methods have been proposed. For example, V2VNet [38] incorporated a graph neural network to fuse data from different agents. V2X-ViT [42] introduced a transformer architecture that combined information from vehicles and infrastructures. Where2comm [15] used feature spatial heterogeneity to reduce bandwidth utilization by transmitting sparse feature maps. CoBEVT [41] proposed the first multi-camera-based collaborative perception framework and designed the fused axis attention module to enable multi-view interactions. In this work, we propose an innovative feature calibration and interaction method that ensures robust vehicle collaboration in high-noise dynamic environments.

3. Method

This section introduces our efficient and robust collaborative perception method. Figure 1 shows the general procedure of our method, which is divided into six phases. We will detail each phase in the following introduction.

3.1. Metadata Sharing and Feature Extraction

In multi-vehicle collaboration scenarios, one of the CAVs, known as the ego vehicle, constructs a communication graph. In this graph, the ego vehicle serves as the requester, while other vehicles within its communication range act as supporters. The ego vehicle broadcasts its metadata information, including position, heading, speed, and more. Supporters, upon receiving this metadata, project their local point cloud observations into the ego vehicle's coordinate system. Each vehicle then encodes these transformed point clouds into bird's eye view (BEV) features, yielding a visual representation. Given the i-th vehicle local observations $\boldsymbol{X}_i$, the extracted features are represented as $\boldsymbol{F}_i = \Phi_{\text{enc}}(\boldsymbol{X}_i) \in \mathbb{R}^{H \times W \times C}$, where $\Phi_{\text{enc}}(\cdot)$ denotes the PointPillar [19] encoder shared by all vehicles and H, W, and C stand for the height, width, and channel of the feature map, respectively. Then, these extracted features are fed into the filter and merge feature sampling module.

3.2. Filter and Merge Feature Sampling

Previous works have utilized well-designed mechanisms such as information entropy communication selection [37] and spatial heterogeneity map [15, 35] to reduce the required transmission bandwidth. However, these methods primarily focus on the inter-class redundancy between foreground and background features, ignoring the intra-class redundancy among features, which results in sub-optimal compression. To address this gap, we introduce an advanced Filtering and Merging Feature Sampling strategy (FMS). This strategy considers both inter-class and intra-class redundancy relationships, efficiently extracting a concise and distinctive set of feature vectors from the original feature maps, thus reducing communication overheads more effectively. FMS is composed of two core components as follows.

Filter Sampler. In object detection, foreground areas containing objects are more significant than the background areas. Therefore, we implement the idea of reducing spatial redundancy into a feature filter sampler module, aiming to preserve perceptually important yet sparse sets of feature vectors. Since explicitly learning a binary sampler is infeasible, we develop a confidence filter strategy. Initially, a detection confidence map is generated for the feature map. It reflects the perceptual importance of different spatial areas, with higher levels indicating potential object areas and lower levels typically denoting redundant background areas. For the feature map $\boldsymbol{F}_i$ of the i-th vehicle, its confidence map $\boldsymbol{C}_i$ is defined as:

$$\boldsymbol{C}_i = \Phi_{\text{con\_gen}}(\boldsymbol{F}_i) \in [0,1]^{H \times W}, \tag{1}$$

where $\Phi_{\text{con\_gen}(\cdot)}$ represents the confidence generation network with detection decoder structure. Then the confidence map is thresholded, followed by non-maximum suppression, resulting in a binary mask B. Utilizing this binary mask, we proceed to preserve the sparse foreground features $\widetilde{\boldsymbol{F}}_i = B \odot \boldsymbol{F}_i$. In order to cope with the dynamically changing environmental conditions and fully ensure the robustness of the system, the filtering threshold can be varied with the sensor data and the network state. We set α as the filter rate and the number of remaining feature vectors is $L = \alpha \times HW$.

Merge Sampler. Upon extracting a detailed foreground feature vector set with the filter sampler, we employ the merge sampler for additional optimization, refining similar or repetitive foreground feature vectors through weighted merging. The process is divided into three stages: information-driven feature grouping, attention-inspired feature merging, and index-based feature reconstruction.

(a) Information-Driven Feature Grouping. Initially, a variant of the nearest neighbor clustering algorithm is applied to group the foreground feature vector set. Given a set of feature vector $\widetilde{\boldsymbol{F}}_i = [x_1, x_2, \ldots, x_L]^\top$ and cluster center X_c, we compute the indicator δ_i for each feature vector. δ_i is calculated as the minimum feature distance minus the average pixel distance to any other cluster center vector, expressed as:

$$\delta_i = \min_{j:x_j \in X_c} \left(\|x_i - x_j\|_2^2 - \gamma \, \|p(x_i), p(x_j)\|_2^2 \right), \quad (2)$$

where δ_i denotes to which cluster the feature vector x_i should belong. $p(\cdot)$ means getting the position of the vector and γ is a hyperparameter. Subsequently, we can divide all the feature vectors in $\widetilde{\boldsymbol{F}}_i$ into K clusters represented by $G = \{G_1, G_2, \ldots, G_K\}$. The total number of clusters is determined by dynamic clustering ratio β, and is calculated as $K = \beta \times L$.

(b) Attention-Inspired Feature Merging. A straightforward strategy for merging feature vectors is to average of each feature vector within a cluster. However, this scheme can be severely affected by outlier feature vectors. Drawing inspiration from attention mechanisms, we utilize the confidence score as a guide to quantify the significance of each feature. Thus, the merged feature vector $\widetilde{x}_i$ for the ith cluster G_i is computed as:

$$\widetilde{x}_i = \frac{\sum_{j \in G_i} c_j x_j}{\sum_{j \in G_i} c_j}, \quad (3)$$

where c_j and x_j represent the confidence score and the original feature vector, respectively. In the end, we obtain a final set of feature vectors $\boldsymbol{Z}_i = \{\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_K\}$ that is necessary for transmission.

(c) Index-Based Feature Reconstruction. During the feature grouping and merging process, each feature vector is allocated to a cluster, and every cluster is represented by a merged vector. We maintain a record of the index correspondence between original and merged feature vectors. Utilizing this index record, the ego-vehicle ensures that the merged feature vectors are mapped to their corresponding positions, leading to the reconstruction of feature maps.

3.3. Feature Spatial Calibration

Localization errors [30, 31] may lead to the misalignment of feature maps among vehicles. Such misalignment causes the ego-vehicle to misinterpret the object's location, resulting in sub-optimal perception, as depicted in the upper right corner of Figure 2. To address this challenge, we introduce the Feature Spatial Calibration module (FSC) to facilitate precise feature alignment, as shown in Figure 2. The core idea of FSC is that overlapping viewpoints allow multiple vehicles to detect partially the same objects. The process

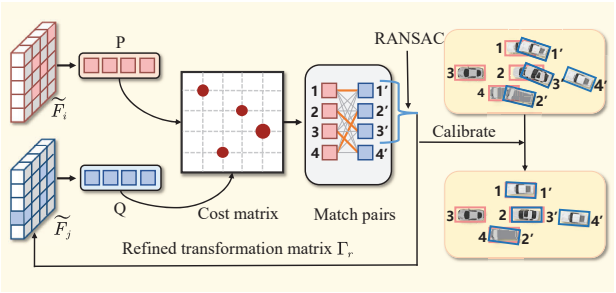


Figure 2. Illustration of the proposed FSC. By employing FSC, misaligned matching regions are corrected at the feature level, leading to more accurate predictions.

involves three phases: consensus matching, geometric verification, and error modulation.

Consensus Matching. After reconstructing the received features, the ego vehicle generates sparse feature maps. Within these maps, areas rich in information signify potential target regions, which correspond to the proposed matching areas. We denote the proposed matching regions of the ego vehicle as P and those identified by the collaborative vehicle under noisy pose conditions as Q. Utilizing P and Q, a weighted bipartite graph is constructed, wherein the weight of each edge is determined by the distance between the nodes, encapsulated in a cost matrix. The matching process is then converted into a linear assignment task, with the objective of identifying a matching result with the lowest cumulative edge weight. This procedure yields the matching pairs M.

Geometric Verification. Invalid matches may occur due to objects located in exclusive zones and detection noise. To tackle this issue, we utilize RANSAC to filter and sift a consistent set of matches aligned with expected geometric transformations. Initially, a random matching subset $M_s \subseteq M$ is selected, and then the transformation matrix Γ_s is computed using singular value decomposition. When Γ_s is applied to all pairs within M_s, and if the post-transformation distance between these pairs remains below the threshold η, the set is considered correctly aligned. The threshold η reflects the allowable localization error within the original collaborative framework. This process is iteratively conducted to identify the optimal transformation matrix that correlates with the maximum number of correct matches. Ultimately, a optimal refined transformation matrix Γ_r is obtained and applied in subsequent spatial calibration operations, yielding the aligned feature $\widetilde{\boldsymbol{Z}}_j = \Gamma_r \boldsymbol{Z}_j$.

Error Modulation. To enhance the adaptability of the calibration method in various environments, we incorporate an error modulation strategy. This strategy aims to achieve a balance between localization errors and the estimation errors that emerge from the calibration process. It measures the overlap ratios between ego and collaborative features in both their adjusted and original states. Subsequently, fea-

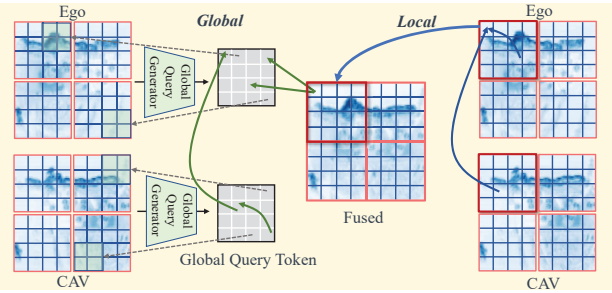


Figure 3. A example of local and global Attention in AFF. This shows how AFF processes both 3D local windows and the global query token for position-level and context-aware aggregation.

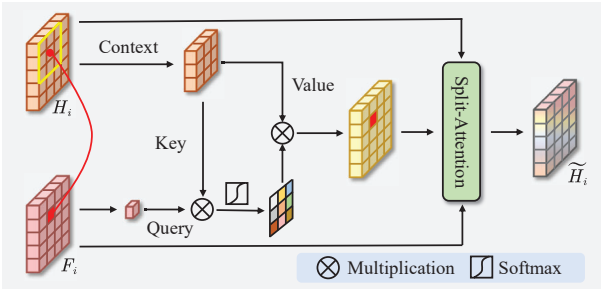


Figure 4. The architecture of the proposed AFI component.

tures demonstrating the most optimal overlap are selected as the final input.

3.4. Attention-based Feature Fusion

In multi-vehicle collaboration scenarios, vehicles are able to capture heterogeneous information from different spatial regions. In order to efficiently fuse perceptual features from multiple vehicles, we propose an Attention-based Feature Fusion method (AFF). As shown in Figure 3, AFF utilizes alternating local and global attention to achieve position-level accurate matching in occlusion-variable traffic scenes and to capture the global semantics attention of road topology and traffic states. Specifically, we stack all the vehicles' features to $S \in \mathbb{R}^{N\times H\times W\times C}$, where N is the number of vehicles. The features are divided into 3D non-overlapping windows, each of size $N \times P \times P$. The shape of the partition tensor is $(\frac{H}{P} \times \frac{W}{P}, N \times P^2, C)$. We perform 3D local self attention on the local window and implement region interaction within the window. In contrast, global attention goes beyond the limitations of the local view. It uses the extracted global query token and is shared between all windows to interact with the local key and value representations, thus helping to capture long-range dependencies of the features. Inspired by the efficiency of [13, 33], the global query token $\mathbf{g}$ is generated as follows:

$$\mathbf{g} = \mathcal{P}\left(\mathcal{C}_{1\times1}\left(\mathcal{SE}(\mathcal{G}(\mathcal{D}(\mathbf{x})))+\mathbf{x}\right)\right), \tag{4}$$

where $\mathcal{D}$, $\mathcal{G}$, $\mathcal{SE}$, $\mathcal{C}_{1\times1}$ and $\mathcal{P}$ denote the depth-wise separable convolution, GELU activation function, Squeeze-and-Excitation operation, 1×1 convolution and max pooling, respectively. These designs enable the global query token to provide advantages such as inductive bias and modeling of inter-channel dependencies. We then construct our proposed AFF module by combining this local and global attention with typical designs of Transformers, including LayerNorm [1], MLPs [9], and skip-connections [14]. This module allows the system to analyze spatial correlations from a local view while also capturing global feature responses, ensuring efficient and precise perception in dynamic, complex, and occlusion-variable traffic scenarios. Finally, we obtain the fused features $\boldsymbol{H}_i$.

3.5. Accuracy Enhanced Feature Interaction

Previous works [15, 21, 38, 42, 43] have demonstrated that fused features can provide richer semantic information, thereby enhancing perceptual performance. However, they may be affected by collaborative noise, such as asynchronous motion blur and inaccurate projections, which can compromise accurate position information and become a bottleneck to the optimal realization of perceptual performance. Ego-centric features may contain locally critical spatial location information without being affected by collaborative noise. To this end, we propose a novel Accuracy Enhanced Feature Interaction (AEI) strategy that leverages the accurate positional information inherent in ego-centric features to enhance the rich semantic information provided by collaborative fused features. Firstly, we design a context cross attention module that is tailored for feature interaction, as shown in Figure 4. It considers the features from the ego-vehicle as queries, and obtains the context of the query location from the fused features as the key and value to achieve cross-attention. The computed output of the spatial location (i, j) is as follows:

$$y_{ij} = \sum_{a,b\in\mathcal{N}_k(i,j)} \text{softmax}\left(q_{ij}^{\top}k_{ab}\right)v_{ab}, \tag{5}$$

where $\mathcal{N}_k(i, j)$ is context position of (i, j) and q_{ij}, k_{ab}, v_{ab} represent the ego query, context key and value, respectively. As the relevant features of the same object are often located in similar locations across different feature maps, we prioritize the local context details of the query location. Additionally, considering that only a few positions within the sparse feature map hold significant information, focusing on these context-rich areas is computationally efficient. Following this, we employ the split attention [48] to adaptively fuse information from multiple branches, generating the enhanced output feature $\widetilde{\boldsymbol{H}}_i$.

Dataset	V2V4Real			OPV2V		
Noise Level σ_t/σ_r (m/°)	0.0/0.0	0.4/0.4	0.8/0.8	0.0/0.0	0.4/0.4	0.8/0.8
No Collaboration	41.64/23.43	41.64/23.43	41.64/23.43	73.25/58.22	73.25/58.22	73.25/58.22
Late Fusion	56.80/27.62	40.81/13.31	26.65/10.66	86.46/79.21	76.56/40.01	45.57/21.00
F-Cooper [3]	61.12/32.36	51.71/24.48	41.90/19.91	87.39/79.39	79.33/40.12	52.91/16.90
AttFuse [7]	64.41/34.32	58.42/30.02	49.37/25.44	90.50/81.80	83.54/52.43	68.37/40.96
Where2comm [15]	63.69/34.79	58.48/28.09	49.77/24.25	90.46/84.22	82.17/56.47	73.95/45.86
DiscoNet [21]	64.19/35.25	57.71/28.41	49.19/24.94	89.58/81.45	83.25/53.22	61.37/30.53
V2VNet [38]	65.70/35.38	62.11/30.63	54.73/25.57	91.35/82.43	85.43/54.16	70.43/31.28
CoAlign [25]	64.06/36.60	59.58/31.52	53.02/26.60	90.93/84.28	84.48/57.45	73.95/48.86
V2X-ViT [42]	66.42/37.34	63.33/32.38	57.15/28.22	91.74/83.31	82.91/54.73	61.70/26.07
CoBEVT [41]	66.01/37.36	58.63/29.09	49.22/23.66	91.71/**85.98**	85.49/60.64	64.63/29.59
ERMVP (Ours)	**67.66/42.97**	**65.01/40.45**	**60.88/35.01**	**92.18**/85.59	**85.67/66.32**	**77.13/59.15**

Table 1. Overall performance on V2V4Real and OPV2V datasets with pose noises. The results are reported in AP@0.5/0.7.

3.6. Decoder and Loss

Based on the final fused feature $\widetilde{\boldsymbol{H}}_i$, we use the detection decoder to generate the final prediction output $\boldsymbol{Y}_i = \Phi_{\text{dec}}(\widetilde{\boldsymbol{H}}_i)$. Each position of $\boldsymbol{Y}_i$ represents a rotated box with classes (c, x, y, h, w, cosα, sinα), corresponding to class confidence, position, size, and angle. These objects are the final output of the proposed collaborative perception system. Following existing work [19], we adopt the smooth L_1 loss for regression and focal loss [23] for classification.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We validate the proposed ERMVP in the task of LiDAR-based 3D object detection on two benchmark datasets including V2V4real [44] and OPV2V [43]. **V2V4real** [44] is the first large-scale real-world dataset on V2V perception, collected by two vehicles equipped with multi-modal sensors driving together in a variety of scenarios. It covers a driving area of 410 kilometers, including 20K frames of point clouds, with training/validation/test sets containing 14210/2000/3986 frames, respectively. **OPV2V** [43] is a large-scale vehicle-to-vehicle collaborative perception dataset, simulated by OpenCDA [40] and Carla [8]. It contains 73 different scenarios, in which different numbers (2 to 7) of collaborative vehicles appear together, each equipped with a lidar sensor and 4 cameras. The dataset contains a total of 11,464 frames of point clouds and RGB images. The training/validation/test sets contain 6374, 1980, and 2170 frames, respectively.

Evaluation Metrics. We adopt the Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 to evaluate the detection performance. The calculation format of communication volume in [15] is used to count the message size by byte in the log scale with base 2.

4.2. Implementation Details

We implement the proposed ERMVP and comparison model on the Pytorch toolbox [28], and train them on a NVIDIA GeForceRTX 3090 GPU using the Adam optimizer [18]. The initial learning rate is 2e-3, and it decays every 15 epochs with a factor of 0.1. All models are trained with 60 epochs and the batch size is set to 2. Early stopping is used to find the best epoch. We also add normal point cloud data augmentation for all experiments, including scaling, rotation, and flipping. All detection models are based on PointPillar [19] backbone to extract 2D features from point clouds and a 0.4 m width/length is used for each voxel. AFF component has 3 encoded layers and a window size of 8 for both local and global attention. The balance hyperparameter γ is 0.1 and the error tolerance threshold η is 0.25. Each vehicle has a communication range of 70 m based on [42], while vehicles outside of this broadcasting radius will be ignored. To simulate the localization and heading errors, we add Gaussian noise with a standard deviation of σ_t for localization errors and σ_r for heading errors.

4.3. Quantitative Evaluation

Comparison of Detection Performance. Table 1 shows the performance comparison results of 3D detection on two datasets. We use the No Collaboration method as a baseline, which only uses ego vehicle's point clouds data without collaboration. Late Fusion allows vehicles to exchange detected outputs and utilizes non-maximum suppression to produce the final result. For intermediate fusion strategies, we evaluate the existing state-of-the-art (SOTA) methods: AttFuse [43], F-Cooper [3], V2VNet [38], DiscoNet [21], V2X-ViT [42], CoAlign (only fusion) [25], CoBEVT [41] and Where2comm [15]. The proposed ERMVP achieves average improvements of 12.48%/18.47% and 6.18%/10.02% on two datasets in AP@0.5/0.7 compared with No Collaboration and Late Fusion methods, demonstrating the superiority of our collaborative method.

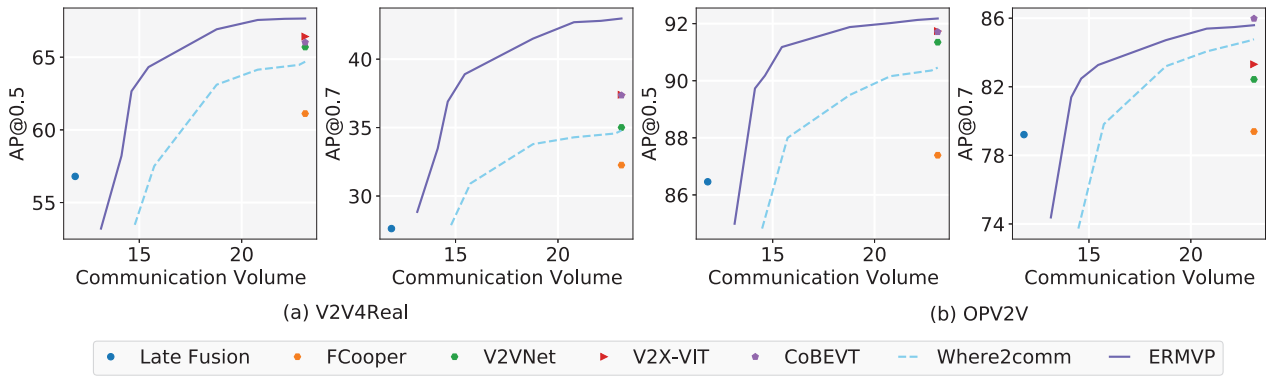


Figure 5. Collaborative perception performance comparison on the V2V4Real and OPV2V with varying communication volume.

Meanwhile, it is superior to the SOTA collaborative perception method in both real-world and simulated scenarios: improves the SOTA collaborative performance by 1.24% and 5.61% on V2V4Real datasets in AP@0.5/0.7, and by 0.44% on OPV2V datasets in AP@0.5. These results fully demonstrate the superiority of the ERMVP collaboration paradigm.

Dataset	V2V4Real		OPV2V	
Noise Level σ_t(m)	0.5	1.0	0.5	1.0
Late Fusion	11.02	10.13	53.88	38.80
F-Cooper [3]	24.60	18.41	32.24	12.71
AttFuse [43]	28.10	24.36	42.95	35.53
DiscoNet [21]	28.36	24.05	45.36	24.96
V2VNet [38]	29.57	23.67	47.48	25.82
Where2comm [15]	28.02	23.53	46.59	45.59
CoAlign [21]	29.84	25.82	49.09	46.66
V2X-ViT [42]	30.99	23.67	44.83	21.77
CoBEVT [41]	24.67	22.01	41.91	26.05
ERMVP (Ours)	**39.26**	**33.94**	**77.25**	**76.09**

Table 2. Detection AP@0.7 on V2V4Real and OPV2V datasets with localization error.

Robustness to Localization and Heading Errors. To evaluate the sensitivity of existing methods to localization and heading errors, we use the same noise setting as [42] and conduct extensive experiments on two datasets. As shown in Tables 1 and 2, under ideal settings, Late Fusion and some advanced intermediate fusion methods can detect object vehicles with high accuracy, but their detection accuracy rapidly decreases as the standard deviation of errors increases. For example, when the localization and heading errors of OPV2V dataset are over 0.4 m and 0.4°, the Disconet [21] and F-Cooper [3] methods fail and perform even worse than No Collaboration. Our method exceeds previous SOTA models at all noise levels and consistently outperforms the No Collaboration baseline, clearly demonstrating the robustness of ERMVP to pose errors. It is noteworthy that for localization errors, ERMVP can maintain consistent accuracy levels in high noise environments, showing superior robustness. When the localization error is 1.0 m, it outperforms the second-best method by 29.43% and 8.08% in AP@0.7 on OPV2V and V2V4Real datasets. The reasonable explanations are: (i) feature spatial calibration aligns mismatched collaborative features; (ii) accuracy enhanced feature interaction reduces performance degradation caused by collaborative noise.

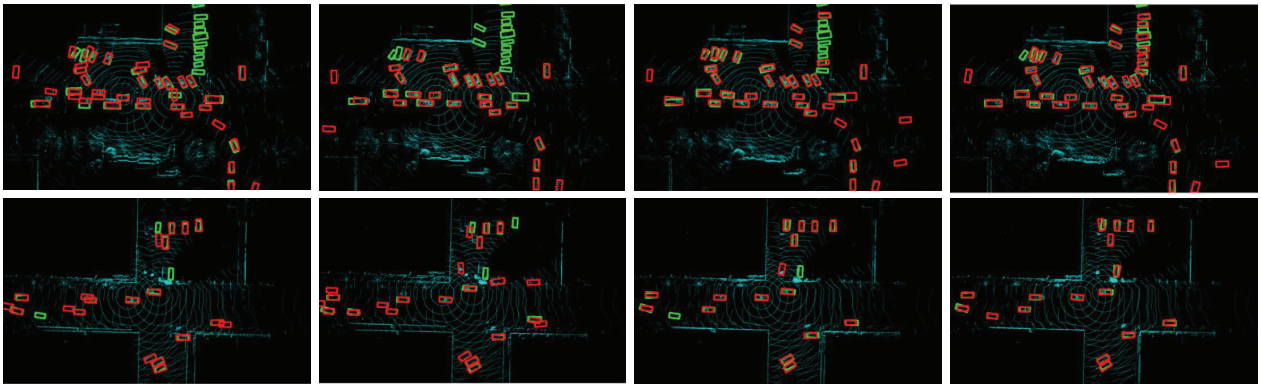
Comparison of Communication Volume. To achieve pragmatic collaborative perception, it is crucial to evaluate the perceptual performance under different communication volumes. Figure 5 shows the results of the performance comparison under different bandwidth consumption conditions. Note that we do not consider any additional model/data/feature compression for a fair comparison. The results indicate that ERMVP achieves excellent perceptual performance and bandwidth consumption trade-off under all communication bandwidth conditions, consistently outperforming Where2comm [15] . At the same time, it achieves the same detection performance as the previous model [3, 38, 41, 42] with less communication volume on both real-world and simulated datasets.

4.4. Ablation Studies

AFF	FSC	AEI	V2V4Real	OPV2V
			51.71/24.48	79.33/40.12
✓			62.24/37.70	84.30/61.43
✓	✓		64.36/39.59	84.98/62.67
✓	✓	✓	**65.01/40.45**	**85.67/66.32**

Table 3. Ablation study results of the proposed core components on the both datasets with noise level of 0.4/0.4. AFF: Attention-based Feature Fusion; FSC: Feature Spatial Calibration; AEI: Accuracy Enhanced Feature Interaction.

Effect of Core Components. Table 3 details the contribution of each core components in our ERMVP framework. The base model is the naive position-wise maximum fusion method. We then assess the impact of each component by sequentially introducing i) attention-based feature



(a) V2VNet (b) CoBEVT (c) V2X-ViT (d) ERMVP

Figure 6. Qualitative comparison results in real-world scenarios from the V2V4Real dataset with noise level of 0.4/0.4. Green and red boxes denote the ground truths and detection results, respectively.

fusion (AFF), ii) feature spatial calibration (FSC), iii) accuracy enhanced feature interaction (AEF). The consistent rise in detection results over both datasets demonstrates the effectiveness of each introduced component. Notably, integrating all three components boosts detection performance by 13.3% and 15.97% on the V2V4Real dataset for AP@0.5 and AP@0.7, respectively.

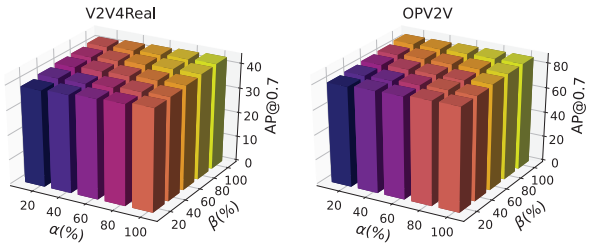


Figure 7. Hyperparameter analysis of α and β

Hyperparameters α and β. We explore the impact of the hyperparameters α and β (as described in Section 3.2), which jointly determine the communication volume. Specifically, the feature compression ratio is represented as $\alpha \times \beta$. The communication volume, in turn, is expressed as the product of the original feature size and the compression rate. By modulating the values of α and β within the interval of 10% to 80%, we chart the resultant object detection performance (measured by AP@0.7) in Figure 7. As expected, a drop in either α or β will impair ERMVP's efficacy. Yet, it's worth noting that ERMVP's performance generally exhibits robustness to these hyperparameters. The performance decline remains fairly minimal even under significant compression, such as a factor of 100. The results demonstrate the effectiveness of our proposed filter and merge feature sampling module. Therefore, ERMVP establishes an appropriate and effective communication paradigm among vehicles, ensuring an optimal trade-off between perception and communication.

4.5. Qualitative Evaluation

Figure 6 displays the detection results of V2VNet [38], CoBEVT [41], V2X-ViT [42], and ERMVP in two scenarios. Clearly, ERMVP offers more precise and comprehensive detection than the prior SOTA methods. To begin with, ERMVP produces a greater number of predicted bounding boxes that are well aligned with the ground truths, while other methods show significant discrepancies. This demonstrates ERMVP's robustness, particularly in high-noise environments. Moreover, ERMVP detects more dynamic objects (more ground truth bounding boxes find matches). This suggests that ERMVP can effectively combine the inputs from nearby vehicles, leading to a thorough scene representation. Overall, these qualitative assessments confirm ERMVP's strengths in delivering accurate and comprehensive perception, especially in challenging conditions.

5. Conclusion

In this paper, we present ERMVP, a communication-efficient and collaboration-robust multi-vehicle perception method in challenging environments. We introduce a filter and merge feature sampling strategy to efficient communication, a feature spatial calibration module to align features in fine-grid and two spatial information aggregation components to solve the bottleneck of the fusion method. Extensive experiments prove the superiority of ERMVP and the effectiveness of our components. The feasibility of our approach for other modalities will be explored in future work.

Acknowledgments

This work is supported by the Shanghai Key Research Laboratory of NSAI, the Innovation Platform for Academicians of Hainan Province, Haikou, 570228, the Specific Research Fund of the Innovation Platform for Academicians of Hainan Province under Grant YSPTZX202314, and NFSC under grant 62250410368.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Marwa Chafii, Salmane Naoumi, Reda Alami, Ebtesam Almazrouei, Mehdi Bennis, and Merouane Debbah. Emergent communication in multi-agent reinforcement learning for future wireless networks. *arXiv preprint arXiv:2309.06021*, 2023. 3

[3] 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, 6, 7

[4] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *European conference on computer vision*, pages 529–548. Springer, 2022. 1

[5] Zhaoyu Chen, Bo Li, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Query-efficient decision-based black-box patch attack. *IEEE Transactions on Information Forensics and Security*, 2023.

[6] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021. 6

[8] 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

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1

[11] Jiaming Gu, Jingyu Zhang, Muyang Zhang, Weiliang Meng, Shibiao Xu, Jiguang Zhang, and Xiaopeng Zhang. Feaco: Reaching robust feature-level consensus in noisy pose conditions. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3628–3636, New York, NY, USA, 2023. Association for Computing Machinery. 1

[12] Monowar Hasan, Sibin Mohan, Takayuki Shimizu, and Hongsheng Lu. Securing vehicle-to-everything (v2x) communication platforms. *IEEE Transactions on Intelligent Vehicles*, 5(4):693–713, 2020. 1

[13] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Yue Hu, Shaoheng Fang snd Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 5, 6, 7

[16] Kaixun Jiang, Zhaoyu Chen, Hao Huang, Jiafeng Wang, Dingkang Yang, Bo Li, Yan Wang, and Wenqiang Zhang. Efficient decision-based black-box patch attacks on video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4379–4389, 2023. 1

[17] Hamza Khan, Sumudu Samarakoon, and Mehdi Bennis. Enhancing video streaming in vehicular networks via resource slicing. *IEEE Transactions on Vehicular Technology*, 69(4): 3513–3522, 2020. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[19] 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. 3, 6

[20] Yiming Li, Bir Bhanu, and Wei Lin. Auction protocol for camera active control. In *2010 IEEE International Conference on Image Processing*, pages 4325–4328. IEEE, 2010. 2

[21] 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. 5, 6, 7

[22] Yiming Li, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: A virtual collaborative perception dataset for autonomous driving. *arXiv preprint arXiv:2202.08449*, 2022. 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[24] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. Ampnet: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*, 20(2):2843–2855, 2024. 1

[25] 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, 6

[26] Zonglin Meng, Xin Xia, Runsheng Xu, Wei Liu, and Jiaqi Ma. Hydro-3d: Hybrid object detection and tracking for cooperative perception using 3d lidar. *IEEE Transactions on Intelligent Vehicles*, 2023. 1

[27] Yaru Niu, Rohan R Paleja, and Matthew C Gombolay. Multi-agent graph-attention communication and teaming. In *AAMAS*, pages 964–973, 2021. 2

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[29] Faisal Qureshi and Demetri Terzopoulos. Smart camera networks in virtual reality. *Proceedings of the IEEE*, 96(10): 1640–1656, 2008. 2

[30] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021. 4

[31] Zhiying Song, Fuxi Wen, Hailiang Zhang, and Jun Li. An efficient and robust object-level cooperative perception framework for connected and automated driving. *arXiv preprint arXiv:2210.06289*, 2022. 4

[32] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993. 2

[33] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 5

[34] Nicholas Vadivelu, Mengye Ren, James Tu, Jingkang Wang, and Raquel Urtasun. Learning to communicate and correct pose errors. In *Conference on Robot Learning*, pages 1195–1210. PMLR, 2021. 1

[35] Binglu Wang, Lei Zhang, Zhaozhong Wang, Yongqiang Zhao, and Tianfei Zhou. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8710–8720, 2023. 1, 3

[36] Jian Wang, Yameng Shao, Yuming Ge, and Rundong Yu. A survey of vehicle to everything (v2x) testing. *Sensors*, 19(2): 334, 2019. 1

[37] 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. 3

[38] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*, pages 605–621. Springer, 2020. 1, 3, 5, 6, 7, 8

[39] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21764–21774, 2023. 1

[40] 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

[41] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers, 2022. 1, 3, 6, 7, 8

[42] 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*, page 107–124. Springer, 2022. 1, 3, 5, 6, 7, 8

[43] 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, 5, 6, 7

[44] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 2, 3, 6

[45] Kun Yang, Peng Sun, Jieyu Lin, Azzedine Boukerche, and Liang Song. A novel distributed task scheduling framework for supporting vehicular edge intelligence. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pages 972–982. IEEE, 2022. 1

[46] Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-temporal domain awareness for multi-agent collaborative perception, 2023. 1

[47] Kun Yang, Dingkang Yang, Jingyu Zhang, Hanqi Wang, Peng Sun, and Liang Song. What2comm: Towards communication-efficient collaborative perception via feature decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 7686–7695, New York, NY, USA, 2023. Association for Computing Machinery. 1

[48] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2736–2746, 2022. 5

[49] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. *Advances in Neural Information Processing Systems*, 33:17271–17282, 2020. 3