ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-1/W2-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

PosiFusion: A Vehicle-to-Everything Cooperative Perception Framework with Positional Prior Fusion

Huan Qiu[1], Youchen Tang[1], Jian Zhou[1*], Chengzhuo Xiong[3], Kai Liu[2], Fuxin Xie[4], Bijun Li[1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China-(huanqiu, youchentang, jianzhou, lee)@whu.edu.cn
[2] Electronic Information School, Wuhan University, China-kailiu@whu.edu.cn
[3] School of Remote Sensing and Information Engineering, Wuhan University, China-2021302131255@whu.edu.cn
[4] School of Electrical Engineering and Automation, Wuhan University, China-xiefuxin@whu.edu.cn
* Correspondence: jianzhou@whu.edu.cn

Keywords: V2X, Cooperative perception, Autonomous driving

Abstract

Collaborative perception technology improves perception performance by enabling agents to exchange complementary perceptual data through the sharing and fusion of multi-viewpoint information. However, existing collaborative perception methods in V2X scenarios face two main challenges. On one hand, they overly rely on sensor observation data for global perception, resulting in an exponential increase in communication bandwidth demands as the scene complexity grows. On the other hand, existing methods treat the trajectory and location information of traffic participants separately from sensor data, neglecting the fact that vehicles, as intelligent agents, serve both as sources and targets of perception. This limitation constrains the further improvement of collaborative perception performance. To address these issues, this paper proposes a novel position-prior-enhanced collaborative perception network, PosiFusion. In terms of communication, PosiFusion introduces a position-prior-based communication selection mechanism that uses prior location information to generate a confidence map of the global perception space. By selecting critical perceptual areas, it significantly reduces the communication bandwidth requirement. Regarding perception performance, PosiFusion incorporates a critical-area perception guidance module, which generates a guidance map of the global perception space using prior information. This guides the network to focus on the perception data from critical areas, thereby enhancing overall perception accuracy. To evaluate the effectiveness of PosiFusion, we conducted tests on two large-scale vehicular collaborative perception datasets, OPV2V and V2XSet. Experimental results demonstrate that PosiFusion outperforms existing state-of-the-art collaborative perception methods while ensuring minimal communication transmission costs.

1. Introduction

Collaborative perception technology improves perception performance by enabling different agents to exchange complementary perceptual data through the sharing and fusion of multi-viewpoint information. This technology is essential in fields such as autonomous driving(Wang et al., 2020, Huang et al., 2023), mobile mapping systems(Guo et al., 2025, Xiao et al., 2024), and multi-robot systems(Zaccaria et al., 2021), where it holds significant promise. To achieve collaborative perception, recent works have contributed several high-quality datasets(Xu et al., 2022c, Xu et al., 2022b, Yu et al., 2022, Huang et al., 2024, Li et al., 2022, Yu et al., 2023) and collaborative perception algorithms(Chen et al., 2019b, Chen et al., 2019a, Liu et al., 2020, Hu et al., 2022, Xu et al., 2022b, Zhao et al., 2023), greatly advancing the field. In the field of intelligent connected vehicles, multiple intelligent vehicles collaborate by sharing sensor data, as well as dynamic information such as their pose, speed, and other related data, in conjunction with perception information from other intelligent agents, such as infrastructure, to jointly build a global perception system. The goal is to achieve more precise environmental perception. However, current collaborative perception methods face two main issues: On one hand, these methods overly rely on sensor observation data to achieve global perception, leading to an exponential increase in communication bandwidth demands as scene complexity grows. On the other hand, existing methods typically overlook the integration of trajectory and location information of traffic participants with sensor data, failing to fully consider the dual role of vehicles in an intelligent perception system—as both perception sources and targets. This fragmented approach limits the performance improvement of collaborative perception systems. Particularly in intelligent connected environments, in addition to intelligent vehicles equipped with high-precision perception sensors, some connected vehicles can also share their real-time location information. This real-time location data efficiently reflects the perception distribution in the local space and provides precise spatial guidance for intelligent vehicles in selecting communication regions and enhancing perception features. However, current collaborative perception methods often fail to fully leverage the vehicles' location-prior information, and the lack of this spatially correlated information restricts further performance improvement of the system.

To address this gap, we propose a collaborative perception strategy based on real-time positioning information embedding. The core idea is to use the real-time positioning information from connected vehicles to guide the communication region selection and perception feature enhancement for intelligent vehicles, thereby improving the communication efficiency and perception accuracy of the collaborative perception system. With the real-time positioning information shared by connected vehicles, intelligent agents can more accurately determine which spatial regions to transmit and enhance the perception data for those areas. Since vehicle movement is often gov-

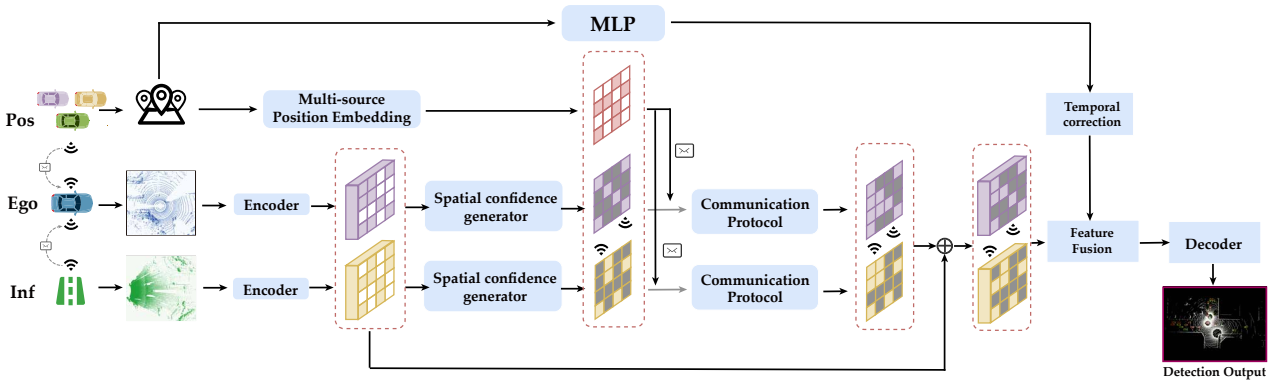


Figure 1. Overall framework of PosiFusion.

erned by road topology and traffic rules, vehicle position information explicitly reflects the driving conditions of surrounding vehicles. For example, vehicles are typically denser near stop lines and more sparse on straight roads.

Based on this idea, we propose a novel position-prior-enhanced collaborative perception network, PosiFusion. Its innovation lies in the dual optimization mechanism for communication efficiency and perception accuracy. In terms of communication, PosiFusion introduces a position-prior-based communication selection mechanism, which uses prior location information to generate a confidence map of the global perception space. By selecting critical perceptual areas, it significantly reduces the communication bandwidth requirement. Regarding perception performance, PosiFusion introduces a critical-area perception guidance module, which generates a guidance map of the global perception space using prior information. This guides the network to focus on the perception data from critical areas, thereby enhancing overall perception accuracy. To evaluate the effectiveness of PosiFusion, we conducted tests on two large-scale vehicular collaborative perception datasets, OPV2V and V2XSet. Experimental results show that, compared to the most advanced collaborative perception methods, PosiFusion achieves superior perception performance while ensuring minimal communication transmission costs. Posifusion saves 28.6% bandwidth compared to the previous method with the lowest communication cost, and averages 5.0% higher performance than the method with the best previous perception results.

2. Related work

2.1 V2X Feature-Level Communication

Feature-level communication strategies have emerged as a crucial research direction in collaborative perception systems, facilitating efficient information sharing among multiple agents. Prior studies have extensively investigated various techniques for compressing and encoding high-dimensional perception features to optimize bandwidth usage and maintain perceptual accuracy. For example, Wang et al.(Wang et al., 2020) proposed a neural encoding method specifically tailored to encode and compress intermediate feature maps, significantly reducing communication overhead while preserving crucial semantic information. Li et al.(Li et al., 2021) introduced convolutional-based compression to extract informative channels from high-dimensional representations, effectively mitigating redundancy. Hu et al.(Hu et al., 2022) introduced a communication strategy based on spatial confidence maps, which allows agents to filter and transmit only the most critical features from perceptually significant regions. Furthermore, recent methods(Yu et al., 2024, Wang et al., 2023) employed attention and entropy-guided frameworks to adaptively select or filter salient features, thereby achieving efficient feature transmission. However, despite these advancements, a challenge remains in achieving an effective balance between feature fidelity and communication efficiency. This issue becomes particularly pronounced when dealing with dense, high-dimensional perception data containing rich spatial and semantic cues.

2.2 Collaborative Perception

Collaborative perception enhances the performance of autonomous systems by facilitating the exchange of information between vehicles and infrastructure, thereby improving both accuracy and robustness. The effectiveness of these systems is largely determined by the strategy employed for message sharing, which can be classified into early, intermediate, and late fusion. Early fusion, involving the sharing of raw sensor data, provides comprehensive information but requires high bandwidth, making it less suitable for real-time applications (Chen et al., 2019b). In contrast, late fusion aggregates detection results, such as bounding boxes or classifications, which reduces communication costs but lacks the contextual depth needed for complex scenarios (Rawashdeh and Wang, 2018). Intermediate fusion, which exchanges intermediate feature representations, offers a balance by preserving the richness of perceptual information while minimizing bandwidth usage. Recent works have predominantly focused on intermediate fusion to optimize the trade-off between perception accuracy and communication efficiency. Li et al. (Li et al., 2021) employed knowledge distillation to align feature representations, while Chen et al. (Chen et al., 2019a) introduced one of the earliest feature-level collaboration methods. Wang et al. (Wang et al., 2020) refined feature exchange with a spatially-aware message passing mechanism, and Liu et al. (Liu et al., 2020) utilized an attention-based approach to dynamically optimize bandwidth usage. Hu et al. (Hu et al., 2022) leveraged the sparsity of foreground information to prioritize the transmission of essential features, thereby reducing communication load. Xu et al. (Xu et al., 2022b) proposed a Transformer-based framework that unifies the fusion process across diverse V2X systems, while Xu et al. (Xu et al., 2022a) integrated multi-camera inputs for BEV map predictions through feature-level collaboration. Additionally, BM2CP (Zhao et al., 2023) and CodeFilling (Hu et al., 2024) introduced multi-modal frameworks that enhance

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W2-2025-115-2025 | © Author(s) 2025. CC BY 4.0 License.

cooperative perception, particularly in LiDAR-camera fusion.

However, none of the aforementioned methods incorporate the prior knowledge of vehicle positions in V2X environments, which could further enhance the collaborative perception performance.

3. Method

3.1 Position Prior Embedding Communication Strategy

Feature Extraction: To efficiently extract features from point cloud data in V2X scenarios, we adopt the PointPillar network, which transforms raw 3D points into structured 2D pseudo-images via pillar-based voxelization. This design significantly reduces memory consumption and computational complexity.

The process begins by organizing the raw point cloud collected by agent i at timestamp k, denoted as $P_i^{(k)}$, which consists of 3D points $\{p_1, p_2, ..., p_n\}$, where each p_i contains spatial coordinates (x_i, y_i, z_i) along with attributes such as intensity or reflectivity. The 3D space is discretized into vertical columns, referred to as pillars, along the horizontal plane. Each pillar is defined as:

$$V_{x,y} = \{p_i \mid p_i \in \text{Pillar}(x, y)\} \tag{1}$$

where $\text{Pillar}(x, y)$ denotes the spatial bin at grid cell (x, y). Features from all points within a pillar are aggregated into a fixed-length vector, typically using pooling operations such as mean or max pooling, resulting in a feature tensor $T_{x,y}$. These tensors are then stacked into a pseudo-image:

$$I_i^{(k)} \in \mathbb{R}^{H \times W \times C} \tag{2}$$

where H, W, and C represent the height, width, and number of channels, respectively. This 2D representation enables the use of standard convolutional neural networks for feature extraction.

A 2D CNN backbone, denoted as $\mathcal{F}(\cdot)$, processes the pseudo-image and outputs a high-level feature map:

$$F_i^{(k)} = \mathcal{F}(I_i^{(k)}) \in \mathbb{R}^{H' \times W' \times C'} \tag{3}$$

where H', W', and C' denote the height, width, and channel dimensions of the output feature. The resulting feature map $F_i^{(k)}$ is subsequently used in downstream V2X communication modules, including spatial confidence generation and target-driven communication.

Spatial Confidence Generator: The spatial confidence generator plays a crucial role in creating target-oriented composite confidence maps by fusing semantic confidence and positional priors. It first estimates semantic criticality from the base feature map $F_i^{(k)}$ to generate a base confidence map:

$$C_{\text{base}}^{(k)} = \Phi_{\text{det}}(F_i^{(k)}) \in [0, 1]^{H \times W} \tag{4}$$

This map shows the probability of each location in the feature map belonging to a target region.

To leverage vehicle positional priors, the module employs a *multi-source position embedding* strategy to aggregate connected vehicles' GPS/IMU data P_j^{V2X} into a unified feature heatmap representation:

$$H^{V2X} = \frac{1}{N} \sum_{j=1}^{N} \mathcal{N}(P_j^{V2X}, \Sigma) \in \mathbb{R}^{H \times W} \tag{5}$$

Simultaneously, intelligent vehicles process radar point cloud data with DBSCAN clustering to extract target positions P_m^{LiDAR} and generate target density maps:

$$H^{\text{obj}} = \sum_{m=1}^{M} \delta(P_m^{LiDAR}) * \mathcal{G}_\sigma \tag{6}$$

These positional priors are dynamically fused through a gating mechanism:

$$H_{\text{pos}}^{(k)} = \alpha \cdot H^{V2X} + (1 - \alpha) \cdot H^{\text{obj}} \tag{7}$$

where $\alpha = \sigma(\text{MLP}([F_i^{(k)}; H^{V2X}]))$ is dynamically adjusted using a multi-layer perceptron (MLP) and sigmoid function.

Finally, spatial modulation refines the confidence map:

$$C_{\text{final}}^{(k)} = C_{\text{base}}^{(k)} \odot \exp(H_{\text{pos}}^{(k)}) + \lambda \cdot \nabla^2 C_{\text{base}}^{(k)} \tag{8}$$

This process enhances key region responses, strengthens edge responses with a Laplacian operator, and adjusts λ based on channel capacity to ensure robustness and accuracy.

Target-Driven Communication Protocol: The target-driven communication protocol optimizes communication strategies to reduce background information interference and improve efficiency. It uses a dual-threshold masking mechanism to identify critical perception regions:

$$M_{\text{com}} = \mathbb{I}(C_{\text{final}}^{(k)} > \tau) \cap \mathbb{I}(H_{\text{pos}}^{(k)} > \mu) \tag{9}$$

Here, τ and μ set thresholds for semantic and positional confidence, effectively filtering out non-target regions and significantly reducing redundant data transmission.

Communication packages contain compressed feature maps with target details preserved through multi-scale pooling. Position residuals, compressed via wavelet transforms, and metadata (including timestamps and coordinate transformation matrices) are also included, which are crucial for temporal-spatial alignment and data fusion at the receiving end.

3.2 Key Region Perception Guidance Module

Construction of Global Perception Space: In collaborative perception networks, to fully utilize vehicle positioning prior information and guide intelligent vehicles to focus on key perception areas, this paper designs a spatial coding network. This network fuses the GPS/IMU data and LiDAR point cloud data of agents to generate a Gaussian mixture position heatmap. For agent i, its positional prior information is represented as:

$$H_i^{pos} = \frac{1}{N} \sum_{j=1}^{N} \mathcal{N}(P_j^{V2X}, \Sigma) \tag{10}$$

where P_j^{V2X} denotes the V2X positioning data of the j-th agent, $\mathcal{N}$ represents the Gaussian distribution, and Σ is the covariance matrix. Furthermore, to enhance the expressiveness of positional priors, we introduce a positional encoding vector E_i^{pos}, generated as follows:

$$E_i^{pos} = \text{MLP}(H_i^{pos}) \tag{11}$$

where MLP denotes a multilayer perceptron that transforms the position heatmap into a high-dimensional feature vector.

To integrate spatial positional prior information with perception features, this paper constructs a global perception space guidance map. Generated via an attention mechanism, this guidance map highlights key regions and suppresses background areas. Specifically, the global perception space guidance map C^{guide} is defined as:

$$C^{guide} = \text{Attention}(E_i^{pos}, F_i^{(k)}) \tag{12}$$

where $F_i^{(k)}$ represents the feature map of the i-th agent during the k-th communication round, and Attention denotes the attention mechanism, calculated as:

$$\text{Attention}(E_i^{pos}, F_i^{(k)}) = \text{Softmax}\left(\frac{E_i^{pos} \cdot F_i^{(k)T}}{\sqrt{d}}\right) \tag{13}$$

where d is the feature dimension. Through this method, spatial positional prior information is effectively incorporated into the perception features, enhancing the model's focus on key regions.

Temporal Correction: Given the potential inaccuracies in spatial positional prior information, such as sensor latency and GNSS errors, this paper proposes a temporal linear enhancement network to correct these issues. This network dynamically adjusts positional priors primarily based on the agent's speed and latency information. Specifically, the input to the temporal linear enhancement network includes the agent's current speed v and latency τ, with the output being the corrected positional prior information $H_{corrected}^{pos}$. The mathematical expression is:

$$H_{corrected}^{pos} = H^{pos} + \Delta H \tag{14}$$

where ΔH is the positional correction term, computed by the temporal linear enhancement network based on speed and latency. The specific calculation for ΔH is:

$$\Delta H = W_v v + W_\tau \tau + b \tag{15}$$

where W_v and W_τ are weight matrices, and b is a bias term. These parameters are learned through the following optimization problem:

$$\min_{W_v, W_\tau, b} \sum_{i=1}^{N} \|H_{corrected}^{pos} - H_i^{gt}\|^2 \tag{16}$$

where H_i^{gt} represents the ground-truth positional information of the i-th agent. This network compensates for positional priors using the agent's kinematic information through linear combination, thereby improving the accuracy of positional information.

Feature Fusion: The perception enhancement strategy is implemented through the following steps: First, the spatial coding network is used to generate spatial positional prior information and combined with an attention mechanism to produce a global perception guidance map. This ensures the model focuses on key regions while suppressing background noise.

Second, a temporal linear enhancement network corrects the global perception guidance map based on the agent's speed and latency, generating a more accurate guidance map. This addresses positional inaccuracies caused by sensor latency and GNSS errors. The corrected global perception guidance map is calculated as:

$$C_{guide}^{corrected} = C^{guide} + \Delta C \tag{17}$$

where ΔC is the correction term computed by the temporal linear enhancement network based on speed and latency.

Then, the corrected global perception guidance map is fused with the perception features to update the feature representation, enhancing detection accuracy and robustness. This fusion enables the model to integrate spatial positional information and perception features. The fusion of the corrected global perception guidance map with perception features is:

$$F_i^{fused} = F_i^{(k)} \odot C_{guide}^{corrected} \tag{18}$$

where $\odot$ denotes element-wise multiplication.

Finally, a feed-forward neural network (FFN) further processes the fused features to obtain the final updated features:

$$F_i^{updated} = \text{FFN}(F_i^{fused}) \tag{19}$$

Through this key region perception guidance module, the proposed collaborative perception network can more effectively perform target detection and recognition in complex environments, enhancing overall performance.

3.3 Detection Decoder

The detection decoder consists of two parallel heads: a *regression head* for 3D bounding box parameter prediction and a *classification head* for foreground object confidence estimation. Given the feature map $\mathcal{F}_i$ from agent i, the detection decoder $\Phi_{\text{dec}}(\cdot)$ predicts the object detection outputs as $\hat{\mathcal{O}}_i = \Phi_{\text{dec}}(\mathcal{F}_i)$, where $\hat{\mathcal{O}}_i \in \mathbb{R}^{H \times W \times (1+7)}$ includes both the classification confidence and the 3D box regression outputs at each spatial location. Specifically, for each grid location, the regression head predicts a 3D bounding box encoded as $(x, y, z, l, w, h, \theta)$, representing the center location, size, and orientation of the object, while the classification head outputs a confidence score indicating the probability that the location corresponds to a valid foreground object. The final object list is obtained by applying post-processing steps such as thresholding and non-maximum suppression (NMS). Note that $\hat{\mathcal{O}}_i^{(0)}$ denotes the detection results of agent i without collaborative fusion.

3.4 Training Details and Loss Functions

To train the entire system, we supervise two tasks: spatial confidence generation and object detection. The spatial confidence generator shares parameters with the detection decoder to improve parameter efficiency. The detection decoder decodes feature maps into object detection results, including classification

Table 1. Comparison to different fusion methods on the OPV2V and V2XSet, PosiFusion consistently outperforms all other fusion approaches

Model	Fusion	OPV2V		V2XSet		AB (Average Byte) ↓
		AP0.5 ↑	AP0.7 ↑	AP0.5 ↑	AP0.7 ↑	
PointPillars	VehOnly	0.687	0.487	0.606	0.402	–
Late Fusion	Late	0.824	0.658	0.727	0.620	$\mathbf{8.40 \times 10^2}$
F-Cooper	Middle	0.878	0.732	0.840	0.680	5.09×10^6
V2VNet	Middle	0.858	0.733	0.845	0.677	1.53×10^7
Where2Comm	Middle	0.889	0.751	0.855	0.654	4.96×10^5
V2X-ViT	Middle	0.867	0.749	**0.882**	0.712	5.09×10^6
PosiFusion	Middle	**0.890**	**0.820**	0.869	**0.772**	3.53×10^5

and regression components. The detection loss comprises classification and regression losses, with focal loss for classification to handle class imbalance and smooth L1 loss for bounding box coordinate regression.

The overall loss function for the system is defined as:

$$L = L_{\text{det}}\left(\hat{\mathcal{O}}_i^{(0)}, \mathcal{O}_i\right) \tag{20}$$

where $\mathcal{O}_i$ represents the ground-truth objects for the i-th agent, $\hat{\mathcal{O}}_i^{(0)}$ is the detection result from the observation encoder without collaboration, and L_{det} is the detection loss combining classification and regression losses.

This simplified training strategy focuses on single-round communication, enabling efficient model training without handling the complexity of multiple communication rounds. It maintains the robustness of the perception system while reducing computational overhead and improving training efficiency.

4. Experiments

4.1 Dataset

V2XSet: The V2XSet dataset (Xu et al., 2022b)is a large-scale V2X perception dataset built upon the CARLA and OpenCDA simulators. Unlike previous datasets, V2XSet incorporates the simulation of localization errors and communication delays, providing a more realistic representation of real-world conditions. The dataset contains a total of 11,447 frames, with the train, validation, and test splits comprising 6,694, 1,920, and 2,833 frames, respectively. V2XSet spans five types of roadway environments: straight segments, curvy segments, midblocks, entrance ramps, and intersections. Each scene features between 2 to 7 intelligent agents engaged in collaborative perception tasks.

OPV2V: The OPV2V dataset (Xu et al., 2022c)is a collaborative vehicle-to-vehicle perception dataset developed through co-simulation using OpenCDA and CARLA. It consists of two primary subsets: the default CARLA township and Culver City. The default CARLA township subset includes 6,765 training samples, 1,980 validation samples, and 2,170 test samples. The Culver City subset, specifically designed for evaluating domain adaptation capabilities, contains 550 samples. The dataset features 12,000 frames, consisting of both 3D point clouds and RGB images, with annotations for 230,000 3D bounding boxes.

4.2 Experimental Setup

Implementation details: In the training phase, a random autonomous vehicle (AV) is selected as the ego vehicle, while during evaluation, a fixed ego vehicle is used for all compared models. For the PointPillar backbone, the voxel resolution is set to 0.4 meters for both height and width. The default compression rate for all intermediate fusion methods is configured to 32. We employ the Adam optimizer, starting with an initial learning rate of 0.001, which is decayed by a factor of 0.1 every 10 epochs. The evaluation range is defined as [-140, 140] meters in the x-direction and [-40, 40] meters in the y-direction. The spatial confidence threshold is set to 0.3 in all experiments. For the position prior, we use the positioning information of connected vehicles located more than 40 meters away from the ego vehicle. All experiments are conducted on an RTX 3090 GPU.

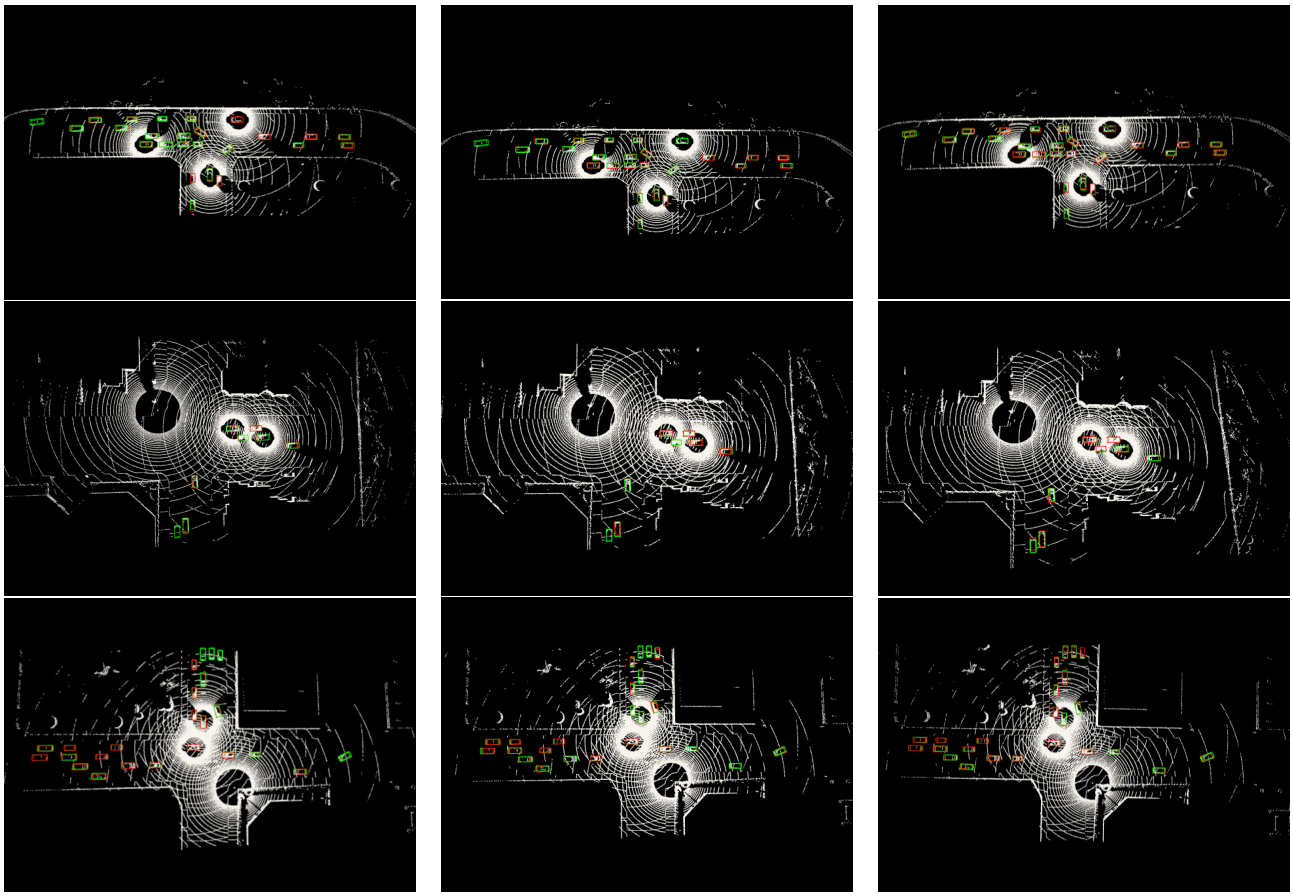
Evaluation metrics

Detection Performance: For the evaluation of 3D object detection, Average Precision (AP) was used to assess detection performance at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7. For this evaluation, vehicles detected by at least one connected LiDAR were considered. The performance assessment is conducted under the assumption of sufficient communication bandwidth and absence of localization noise.

Communication Bandwidth: To evaluate transmission costs, we used AB (average Byte) as the metric, excluding calibration files and timestamps. The overall transmission cost was assessed based on the transmission of raw data, detection results, or feature tensors, with the bandwidth consumption quantified on a per-frame basis. The evaluation is conducted under consistent compression rates and transmitted feature dimensions aligned with those of the respective baseline methods.

4.3 Quantitative Evaluation

The experimental results clearly demonstrate the superiority of PosiFusion in achieving high detection performance while maintaining exceptional communication efficiency across both the OPV2V and V2XSet datasets. As summarized in Table 1, PosiFusion achieves the best detection accuracy among all compared methods.Specifically, on the OPV2V dataset, PosiFusion achieves an AP@0.5 of 0.890 and AP@0.7 of 0.820, exhibiting consistent performance advantages over other advanced middle fusion baselines. Compared to Where2comm, PosiFusion improves the AP@0.7 by 6.9%, while maintaining a comparable AP@0.5. It also surpasses V2X-ViT by 2.3% in



(a) V2XVit (b) Where2Comm (c) PosiFusion

Figure 2. The detection visualization results of V2XVit, Where2Comm and PosiFusion across several challenging scenarios.

AP@0.5 and 7.1% in AP@0.7. On the V2XSet dataset, PosiFusion continues to demonstrate its superiority with an AP@0.5 of 0.869 and AP@0.7 of 0.772, achieving a 2.4% and 9.5% improvement over V2VNet in AP@0.5 and AP@0.7, respectively, and outperforming F-Cooper by 2.9% in AP@0.5 and 9.2% in AP@0.7. These substantial gains confirm PosiFusion's capability to effectively exploit positional and semantic priors for enhanced collaborative perception. By providing spatial context during feature fusion, PosiFusion enhances the network's ability to reason about object locations and scene topology, thereby contributing significantly to the observed improvements in detection performance. In addition to its strong detection performance, PosiFusion also achieves remarkable communication efficiency. It consumes only 3.53×10^5 bytes per frame on average, which is less than 1/14th of V2X-ViT and over 40 times lower than traditional middle fusion baselines like F-Cooper and V2VNet. Notably, even when compared to the middle fusion scheme Where2Comm, which transmits only minimal information, PosiFusion achieves both higher accuracy (e.g., +11.8% AP@0.7 on V2XSet) and better bandwidth efficiency (roughly 28.6% lower transmission cost). These results demonstrate that the integration of positional prior embedding not only improves spatial alignment but also assists in identifying critic

4.4 Qualitative evaluation

Detection visualization:Fig. 2 presents the detection visualizations of PosiFusion in comparison with V2X-ViT and Where2Comm across three challenging intersection scenarios. PosiFusion consistently exhibits superior capability in detecting occluded and distant targets. In the first scenario, a T-junction with a wide field of view, PosiFusion is the only method that successfully detects a distant vehicle located at the far-left edge of the scene, which remains undetected by both baselines. In the second crossroad scenario, a vehicle occluded by a lead car in the lower region is accurately identified only by PosiFusion, while both V2X-ViT and Where2Comm fail to perceive it. The third scenario further highlights PosiFusion's strength, where three target vehicles partially occluded by a preceding car in the upper region are all correctly detected solely by PosiFusion. These consistent improvements underscore the model's robustness in complex urban settings where occlusions and long-range perception are prevalent challenges.

This performance advantage is largely attributed to the integration of positional prior embedding, which enables the model to incorporate global spatial context into the feature representation. By leveraging spatially-informed priors during feature fusion, PosiFusion achieves more complete scene understanding, thus facilitating the detection of targets that are distant, occluded, or otherwise difficult to perceive. These qualitative results demonstrate the effectiveness of positional priors in improving the robustness of collaborative perception.

5. Conclusion

In this paper, we presented PosiFusion, a position-prior-enhanced collaborative perception framework tailored for V2X scenarios, aiming to simultaneously optimize perception performance and communication efficiency. By incorporating real-

time positional priors from connected vehicles, PosiFusion introduces a communication selection mechanism and a critical-area guidance module that jointly enable spatially informed feature transmission and attention. This design allows the system to effectively focus on key perception regions, thereby reducing redundant data exchange while enhancing the accuracy of object detection, especially under challenging conditions such as occlusions and long-range perception. Experimental results on two large-scale collaborative perception datasets, OPV2V and V2XSet, demonstrate that PosiFusion achieves superior detection accuracy compared to state-of-the-art methods, with substantial reductions in communication bandwidth—achieving high-performance perception at a fraction of the transmission cost. These findings validate the effectiveness of positional priors in collaborative perception and indicate the strong potential of PosiFusion for deployment in intelligent transportation systems with limited communication resources. However, this work does not explicitly analyze the impact of localization noise in shared positional data or the effects of the number and distance of connected vehicles on perception performance. In future work, we plan to investigate these factors in depth and develop robust adaptation mechanisms to enhance the reliability and scalability of PosiFusion in diverse and dynamic V2X environments.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42471480 and in part by the Hubei Provincial Natural Science Foundation of China under Grant 2024AFB778. The authors also acknowledge all editors and reviewers for their suggestions.

References

Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S., 2019a. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 88–100.

Chen, Q., Tang, S., Yang, Q., Fu, S., 2019b. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 514–524.

Guo, Y., Zhou, J., Dong, Q., Li, B., Xiao, J., Li, Z., 2025. Refined high-definition map model for roadside rest area. *Transportation Research Part A: Policy and Practice*, 195, 104463.

Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S., 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35, 4874–4886.

Hu, Y., Peng, J., Liu, S., Ge, J., Liu, S., Chen, S., 2024. Communication-efficient collaborative perception via information filling with codebook. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15481–15490.

Huang, X., Wang, J., Xia, Q., Chen, S., Yang, B., Wang, C., Wen, C., 2024. V2X-R: Cooperative LiDAR-4D Radar Fusion for 3D Object Detection with Denoising Diffusion. *arXiv preprint arXiv:2411.08402*.

Huang, Y., Zhou, J., Li, X., Dong, Z., Xiao, J., Wang, S., Zhang, H., 2023. MENet: Map-enhanced 3D object detection in bird's-eye view for LiDAR point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 120, 103337.

Li, Y., Ma, D., An, Z., Wang, Z., Zhong, Y., Chen, S., Feng, C., 2022. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4), 10914–10921.

Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W., 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34, 29541–29552.

Liu, Y.-C., Tian, J., Glaser, N., Kira, Z., 2020. When2com: Multi-agent perception via communication graph grouping. *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 4106–4115.

Rawashdeh, Z. Y., Wang, Z., 2018. Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 3961–3966.

Wang, T., Chen, G., Chen, K., Liu, Z., Zhang, B., Knoll, A., Jiang, C., 2023. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8187–8196.

Wang, T.-H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R., 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 605–621.

Xiao, J., Wang, S., Zhou, J., Tian, Z., Zhang, H., Wang, Y.-F., 2024. MIM: High-Definition Maps Incorporated Multi-View 3D Object Detection. *IEEE Transactions on Intelligent Transportation Systems*.

Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J., 2022a. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*.

Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.-H., Ma, J., 2022b. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *European conference on computer vision*, Springer, 107–124.

Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J., 2022c. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2583–2589.

Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J. et al., 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21361–21370.

Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., Nie, Z., 2024. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36.

Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N. et al., 2023. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5486–5495.

Zaccaria, M., Giorgini, M., Monica, R., Aleotti, J., 2021. Multi-robot multiple camera people detection and tracking in automated warehouses. *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, IEEE, 1–6.

Zhao, B., ZHANG, W., Zou, Z., 2023. Bm2cp: Efficient collaborative perception with lidar-camera modalities. *7th Annual Conference on Robot Learning*.

https://doi.org/10.5194/isprs-annals-X-1-W2-2025-115-2025