

Vehicle-infrastructure cooperative general object detection through feature flow and differentiable pose-based spatial alignment

Yanding Yang^{1,2}, Rujun Yan¹, Jinyu Miao¹, Zhiwei Zhao², Kewei Wang³, Kun Jiang^{1✉}, Diange Yang¹

 **Cite this article:** Yang Y, Yan R, Miao J, et al. *J Intel Connect Veh* 2026, 9(3):9210091. <https://doi.org/10.26599/JICV.2026.9210091>

ABSTRACT: Vehicle-infrastructure cooperative perception is critical for advanced autonomous driving but faces challenges: heavy reliance on ground truth (G.T.) vehicle-infrastructure poses for spatial synchronization and limitations of traditional object detection in dynamic scene modeling. This study proposes a differentiable pose-based spatial alignment (DPSA) module and explores occupancy flow output for cooperative perception. The DPSA module eliminates G.T. pose requirements by estimating 3-degree-of-freedom relative poses via feature fusion, spatial average pooling, and fully connected layers, balancing accuracy and practicality. Evaluations on DAIR V2X and V2X-Seq datasets show DPSA outperforms misaligned fusion methods in key metrics, maintains robustness under low-location precision scenarios, and reduces inference time. Notably, occupancy flow demonstrates superior dynamic modeling capability compared to traditional object detection by capturing spatial occupancy changes and motion flows (instead of bounding boxes), enabling high accuracy future occupancy prediction and occlusion resistance via multi-source temporal context. This work advances practical cooperative perception through an efficient synchronization solution and highlights occupancy flow advantages, bridging algorithmic innovation and engineering deployment.

KEYWORDS: vehicle-infrastructure cooperative; feature-level fusion; occupancy flow; pose estimation; spatial alignment

1 Introduction

Accurate three-dimensional (3D) object detection and motion prediction are critical tasks in autonomous driving, as they provide critical environmental context for planning and control by capturing the current positions and future states of surrounding obstacles. Traditional approaches rely heavily on onboard sensors with limited perception range, often resulting in performance degradation due to target distance exceeding sensor capabilities or inherent blind spots. To address these limitations, vehicle-infrastructure cooperative perception has emerged as a prominent research direction in autonomous driving (Gao et al. 2024). By integrating multi-source sensor data from extended spatial domains, this paradigm enhances perception robustness and safety margins of autonomous systems. This paper focuses on LiDAR-based vehicle-infrastructure cooperative object detection and motion prediction, aiming to improve collaborative target detection accuracy through innovative technical solutions.

Vehicle-infrastructure cooperative object detection inherently involves multi-sensor data fusion (Huang et al., 2025), presenting three primary challenges. First, effective fusion strategies must address heterogeneous data integration, with current research classifying approaches into early fusion (Chen et al., 2019), middle fusion (Chen et al., 2023; Li et al., 2021; Qiao and Zulkernine, 2023), and late fusion (Teufel et al., 2023). Practical deployment

faces critical bottlenecks from bandwidth constraints and stochastic communication delays, which significantly degrade real-time detection performance. Second, conventional 3D detection frameworks depend on bounding box outputs (Mao et al., 2023) as inputs for motion prediction modules, creating cascaded errors when front-end detectors fail to generalize to unseen categories in open-world scenarios, leading to critical missed detections. Third, spatial alignment errors arising from vehicle localization inaccuracies cause mismatched feature associations during fusion, directly impairing perception precision (Lei et al., 2024). These challenges reflect the inherent complexity of urban traffic systems and underscore the need for AI-empowered scenario planning toward smart and resilient cities (Gao et al., 2025; Hao et al., 2024).

To address category whitelisting constraints in object detection, we propose leveraging Bird's-Eye View (BEV) representation as an alternative to occupancy grid maps (OGMs). While OGMs encode spatial occupancy in two-dimensional (2D) grid cells, they typically lack semantic category annotations and fail to establish temporal correlations between consecutive frames, hindering motion prediction capabilities. In contrast, our BEV framework enriches each grid cell with height information alongside occupancy status, enabling simultaneous occupancy assessment and semantic classification. By reformulating 3D object detection

¹ School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. ² Dongfeng Motor Corporation Research & Development Institute, Wuhan 430100, China. ³ Dongfeng USharing Technology Co., Ltd., Wuhan 430058, China.

✉ Corresponding author. E-mail: jiangkun@tsinghua.edu.cn

Received: January 7, 2026; Revised: May 4, 2026; Accepted: July 24, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

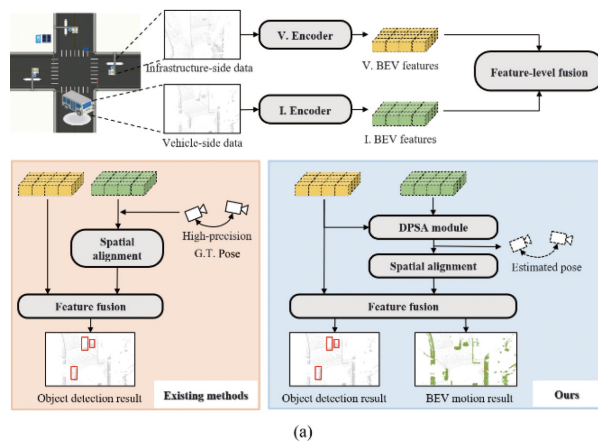


Fig. 1 Comparison of the proposed DPSANet and SOTA methods. (a) Network architecture; (b) BEV AP and communication cost.

and motion prediction as a spatio-temporal occupancy forecasting task, this approach mitigates missed detection issues inherent to traditional category-constrained methods, enhancing generalization to open-world scenarios.

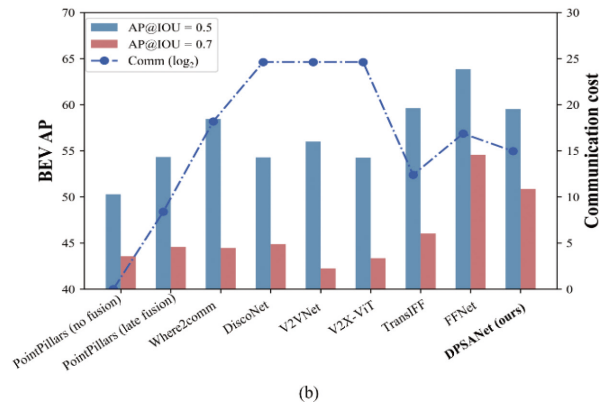
To mitigate performance degradation caused by stochastic communication delays, we adapt the flow-based feature fusion network (FFNet) framework (Yu et al., 2023) to generate compressed feature streams from continuous infrastructure sensor data. These streams function as temporal prediction models, explicitly aligning infrastructure data features with ego-vehicle observations across temporal offsets to resolve asynchronous data fusion errors. To address bandwidth limitations, attention-based masking mechanisms are employed during transmission, prioritizing critical spatial features while reducing data payload (Hu et al., 2022).

To resolve feature misalignment caused by vehicle localization errors, this paper introduces the differentiable pose-based spatial alignment (DPSA) module network, inspired by recurrent all-pairs field transforms (RAFT) (Teed and Deng, 2020), for precise spatial alignment between vehicle and infrastructure features. Operating on BEV representations of point cloud data from both agents, which estimates 3-degree-of-freedom (3-DoF) relative pose parameters and computes optical flow fields to register heterogeneous sensor features. This process enhances spatial coherence of multi-source data, directly improving detection accuracy by minimizing mismatches during fusion.

Comprehensive evaluations are conducted on the DAIR-V2X and DAIR-Seq dataset (Yu et al., 2022), with benchmarking against representative state-of-the-art approaches, as previewed in Fig. 1. Ablation experiments further validate the contribution of each proposed component. Experimental results demonstrate that our framework achieves substantial improvements in cooperative object detection accuracy and motion prediction performance under realistic operating conditions.

Specifically, this paper makes three key contributions:

- We propose a novel LiDAR-based vehicle-infrastructure cooperative framework that explicitly reformulates 3D object detection and motion prediction as a spatio-temporal occupancy forecasting task using BEV representation. This paradigm mitigates annotation overhead and enhances adaptability to downstream planning and control tasks.
- The proposed framework ensures robust spatio-temporal consistency between vehicle and infrastructure features through integrated feature stream alignment and the differentiable pose-based spatial alignment (DPSA) module, directly improving perception precision without relying on prior relative pose information.



- Extensive experiments on public datasets demonstrate state-of-the-art detection accuracy, with particular robustness in low-precision localization scenarios, validating its practical applicability for real-world deployment.

2 Related works

Vehicle-infrastructure cooperative perception, as a core scenario of multi-sensor fusion for autonomous driving, has received sustained attention from top transportation and intelligent vehicle journals in recent years. Following the research vein of related works, this section reviews three closely related directions: LiDAR object detection, vehicle-infrastructure cooperative object detection, and optical flow applications.

LiDAR object detection. Before being fed into detectors, the geometric fidelity of raw LiDAR measurements is degraded by sensor intrinsic errors, e.g., the eccentric structure of mechanical LiDAR, which can be compensated by intrinsic-parameter modeling with angle and distance correction (Zhang et al., 2021). According to the representation form of data, LiDAR object detection can be classified into view-based, voxel-based, and point-based methods (Song et al. 2024). View-based approaches transform LiDAR point clouds into pseudo-images and then conduct object detection. For instance, methods like LaserNet (Meyer et al. 2019), RangeRCNN (Liang et al., 2020), and RangeIoUDet (Liang et al., 2021) project point clouds onto a cylindrical coordinate system to generate dense range-view representations. These methods employ 2D convolutions, which are highly efficient yet sensitive to occlusion and scale variations, rendering them inappropriate for vehicle-infrastructure cooperative detection. Other methods, such as PIXOR (Yang et al., 2018), Birdnet (Barrera et al., 2020; Beltrán et al., 2018), and RangeDet (Fan et al., 2021), project point clouds onto 2D grids while encoding information such as height, intensity, and density to achieve object detection. These techniques exhibit stronger robustness in dealing with occluded scenes and scale changes. Voxel-based methods voxelize the point cloud and then utilize 3D networks for detection (Lang et al., 2019). However, such methods encounter several challenges, including inconsistencies between voxel size and point cloud density. Excessively large voxels may result in the loss of spatial information, while overly fine voxels require substantial memory and computational resources.

Point-based methods directly process raw point clouds, thus preserving more comprehensive structural information. Nevertheless, they also suffer from the inclusion of irrelevant points in the computation, which may affect performance.

In contrast, view-based methods feature lower latency and

higher speeds, but experience a loss of 3D spatial information due to occlusion, resulting in lower accuracy. This can be compensated for in vehicle-infrastructure cooperative detection scenarios. Recent advanced 3D perception architectures have further explored query-based paradigms to improve instance-aware capability, such as InVDriver (Zhang et al., 2025) which proposes an intra-instance aware vectorized query-based transformer for autonomous driving detection. However, such methods still suffer from pre-defined category whitelists, which motivates the BEV occupancy paradigm adopted in this work to support open-category semantic classification.

Vehicle-infrastructure cooperative object detection. In the research realm of vehicle-infrastructure cooperative object detection, numerous studies have put forward solutions to the issue of diminished detection accuracy resulting from inconsistencies in spatio-temporal information (time delay and localization error) (Xu et al., 2022). Beyond general accuracy drop, dynamic traffic scenarios also pose challenges for conventional bounding box-based detection: Liu et al. (2025) proposed an RGB-event fusion scheme to improve the robustness of object detection in dynamic scenarios, but most existing cooperative perception methods still adopt box outputs, which have cascaded errors in motion prediction and poor generalization to unseen categories.

To address the problem of temporal delay in data from both vehicles and roadside units, the real-time cooperative perception fusion framework, CooperFuse (Zheng et al., 2024), introduces a multi-agent time synchronization module. It accomplishes clock calibration among multiple agents via three rounds of communication, thereby minimizing network transmission delays. FFNet (Yu et al., 2023) proposes an algorithm framework based on feature flow, which compensates for the temporal asynchrony between vehicle and roadside data by dynamically predicting the future evolution trends of roadside sensor features. CoBEVFlow (Wei et al., 2023) presents a robust asynchronous cooperative perception system founded on BEV flow. This approach utilizes BEV flow to model scene motion vectors, relocating asynchronous perceptual features to their precise positions at the target timestamp, thus resolving the misalignment caused by delays and clock deviations in multi-agent communication. A particularly noteworthy study (Zhu et al.,

2022) explores the trade-off between accuracy and communication latency in point cloud fusion strategies (early fusion, middle fusion, late fusion) within vehicle-infrastructure cooperative perception. It proposes a simulation-based latency impact analysis framework.

This framework employs the extended Kalman filter (EKF) for trajectory prediction compensation on delayed perception results. The study further introduces Lag Compensation Error and an equivalent latency metric for comprehensive evaluation. The findings suggest that high-dynamic scenarios should prioritize low-latency late fusion, whereas static or low-speed scenarios are more appropriate for high-accuracy early fusion.

Regarding localization errors in vehicle-infrastructure cooperative perception, Zhu et al. (2022) innovatively proposed an end-to-end learning framework to tackle pose noise in multi-vehicle cooperative perception. It synergistically integrates three mechanisms: pose correction, global synchronization, and noise filtering. Specifically, it uses feature maps to predict pose deviations and compensate for localization errors, achieves pose synchronization with few nodes based on spatial overlap priors, and dynamically weights and aggregates messages (capable of fully filtering out noise). This method represents the first collaborative perception solution in vehicular networks that achieves adaptive noise correction. V2X-PC (Liu et al., 2024) is a vehicle-infrastructure cooperative perception framework, which employs point clusters as a novel collaborative message unit. It models point cluster centers as graph vertices and minimizes pose consistency errors through a parameter-free graph optimization algorithm. The results indicate its stable performance under pose noise. The framework CoAlign (Lei et al., 2024) combines a hybrid collaboration strategy (fusing intermediate features and detection results) with an innovative agent-object pose graph optimization technique. It dynamically corrects unknown pose errors without any real pose supervision and significantly enhances noise resistance through multi-scale fusion and uncertainty modeling.

However, existing methods face notable limitations that hinder their practical deployment, as illustrated in Fig. 2. Approaches relying on geometric affine transformation (Fig. 2a) are heavily dependent on ground truth (G.T.) pose information for spatial

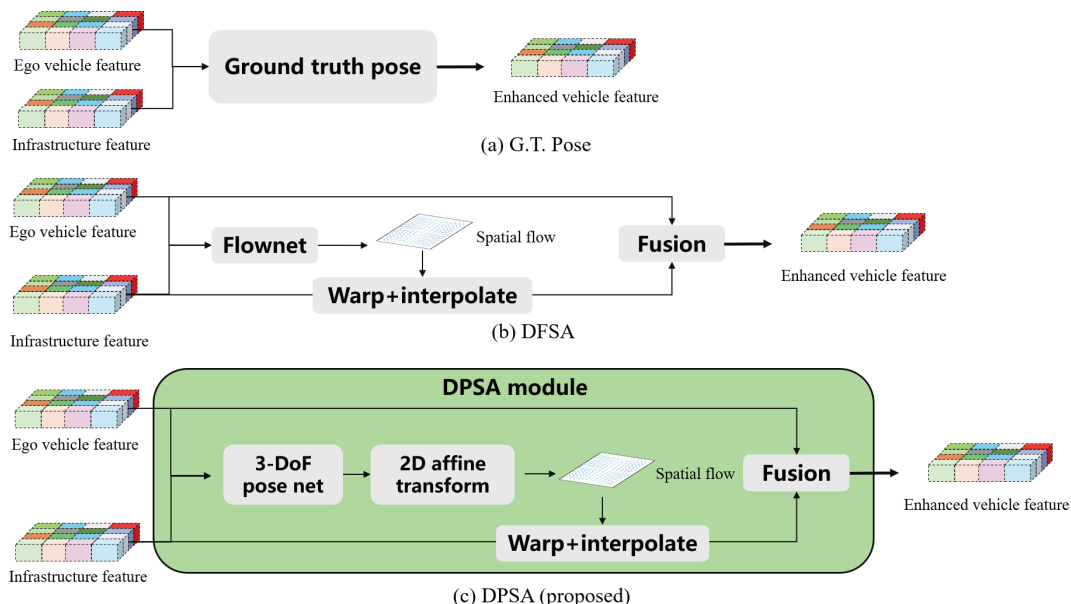


Fig. 2 Comparison of different spatial alignment methods. (a) Geometric affine transformation (G.T. Pose): relies on manual calibration; (b) DFSA: directly estimates flow without pose priors; (c) DPSA (proposed): leverages 3-DoF pose estimation for balanced accuracy and efficiency.

alignment, a requirement that entails laborious manual calibration, incurs high costs, and fails to adapt to real-time dynamics in vehicle-infrastructure cooperation. On the other hand, pure optical flow-based methods (Fig. 2b), which estimate dense flow fields without pose priors, often suffer from low computational efficiency, characterized by high parameter counts and significant overhead, as they prioritize flow estimation over geometric constraints. These shortcomings highlight a critical gap in achieving efficient and robust alignment for real-time systems. Motivated by this, we propose the DPSA module (Fig. 2c) to address the aforementioned issues.

Optical flow. Optical flow refers to the technique that matches pixels between two consecutive image frames and computes the alteration in pixel coordinates. It is widely employed for motion estimation and object tracking. In recent years, optical flow methods based on neural networks have witnessed rapid advancements and are increasingly being applied to cross-modal relative pose estimation.

RAFT (Teed and Deng, 2020) represents a novel end-to-end trainable optical flow estimation model. It extracts the similarity of full-pixel features via a global correlation pyramid, eschewing the traditional "coarse-to-fine" approach. Characterized by high accuracy, strong generalization ability, high efficiency, and a low number of parameters, it has established a new benchmark in the field of optical flow.

The flow-based global and metric radar localization in LiDAR maps (RaLF) (Nayak et al., 2024) end-to-end deep learning framework jointly learns location recognition and metric localization, converting metric localization into a pixel-level optical flow prediction problem. It utilizes the RAFT network to predict the correspondence between LiDAR data and maps, facilitating robust cross-modal localization of LiDAR scans within maps. The BEVLOC (Chen et al., 2024) end-to-end network framework fuses monocular camera images with pre-built LiDAR maps in a Bird's-Eye View (BEV) space. It leverages optical flow to calculate cross-modal correlations for predicting vehicle pose offsets, achieving high-precision pose estimation from camera images to LiDAR maps.

The aforementioned studies illustrate the potential of optical flow tracking in mitigating localization errors. Nevertheless, these methods are primarily designed for homogeneous visual data and face challenges in V2X cooperative perception, such as handling the sparsity of LiDAR point clouds and adapting to heterogeneous sensor configurations. For instance, RAFT-based approaches may lack mechanisms for incorporating vehicle-infrastructure spatial priors, while cross-modal methods like RaLF and BEVLOC depend on pre-built maps, limiting their generality in dynamic environments. These limitations underscore the need for specialized flow estimation that integrates domain-specific constraints, as addressed by the DPSA module through differentiable pose modeling. To date, no existing vehicle-infrastructure cooperative perception method has explored occupancy flow as a fusion representation, making direct quantitative comparisons with prior work infeasible. This study therefore establishes the first benchmark in this direction, comparing against vehicle-only occupancy prediction and conventional cooperative detection baselines.

3 Method

3.1 Overview

The proposed DPSANet framework presented in this study tackles the challenge of spatial misalignment in vehicle-

infrastructure cooperative perception via a novel feature alignment and fusion methodology. The framework is composed of three principal components: (1) a dual-branch feature extraction network that separately processes the point clouds from the vehicle end and the infrastructure end; (2) an innovative spatial synchronization module that rectifies the positional discrepancies between asynchronous features through optical flow estimation; and (3) a feature fusion module that integrates the aligned features for precise 3D object detection and occupancy flow analysis. As illustrated in Fig. 3, infrastructure features are first temporally aligned from their original timestamp to the vehicle frame timestamp via a feature flow predictor, then spatially aligned with vehicle features using the DPSA module.

3.2 Feature extraction and compression

Dual-branch architecture: The network adopts a dual-branch VoxelNet architecture, in which the point clouds from the vehicle end and the infrastructure end are processed via separate yet identical feature extraction pipelines. Each branch comprises a voxelization layer with configurable parameters (maximum number of points = 100, maximum number of voxels = 40,000), a PillarFeatureNet encoder with 64 output channels, a PointPillarsScatter middle encoder, and a SECOND backbone network with layer configurations [3, 5, 5] and output channels [64, 128, 256].

Feature compression strategy: To mitigate communication overhead while retaining essential information, a channel reduction mechanism is implemented for the infrastructure-end features. The infrastructure features are processed through a ReduceInfTC module, which gradually reduces the channel dimension from 768 to 384 through a series of convolutional layers with stride-2 downsampling, followed by corresponding transposed convolutions for upsampling. This compression strategy maintains the spatial resolution while substantially reducing the transmission bandwidth requirements. The feature compression can be formulated as

$$F_{\text{inf_compressed}} = \text{ReduceInfTC}(F_{\text{inf_raw}}) \quad (1)$$

where $F_{\text{inf_raw}} \in \mathbb{R}^{B \times 768 \times H \times W}$ represents the raw infrastructure features and $F_{\text{inf_compressed}} \in \mathbb{R}^{B \times 384 \times H \times W}$ denotes the compressed features.

Voxelization process: The point cloud voxelization is performed with specific parameters: `point_cloud_range` = [0, -46.08, -3, 92.16, 46.08, 1], `voxel_size` = [0.16, 0.16, 4], resulting in `output_shape` = [576, 576] for the BEV feature maps.

3.3 Spatial synchronization

3.3.1 Geometric affine transformation spatial alignment (G.T. Pose) module

As a fundamental approach, we employ geometric affine transformation based on calibration parameters to rectify spatial misalignment. This methodology calculates the transformation matrix by utilizing the rotation and translation parameters between the infrastructure and vehicle coordinate systems.

Coordinate transformation: Given the calibration parameters R_{inf2veh} (rotation matrix) and T_{inf2veh} (translation vector), we construct the affine transformation matrix:

$$\theta_{\text{rot}} = \begin{pmatrix} R_{\text{inf2veh}}[0, 0] & -R_{\text{inf2veh}}[0, 1] & 0 \\ -R_{\text{inf2veh}}[1, 0] & R_{\text{inf2veh}}[1, 1] & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (2)$$

The translation offsets in pixel coordinates are calculated as

$$\begin{cases} x_{\text{trans}} = \frac{-2 \times t_0}{P_3 - P_0} \\ y_{\text{trans}} = \frac{-2 \times t_1}{P_4 - P_1} \end{cases} \quad (3)$$

Feature warping: The infrastructure features are warped using the computed transformation:

$$F_{\text{inf_warped}} = \text{F.affine_grid}(\theta_{\text{combined}}, F_{\text{inf}}.\text{shape}, \text{False}) \quad (4)$$

Geometric flow computation: The geometric transformation is used to compute ground truth optical flow for supervision.

3.3.2 Differentiable pose-based spatial alignment (DPSA) module

Network architecture: We propose a lightweight flow estimation module specifically designed for vehicle-infrastructure spatial synchronization (Fig. 4). The network module takes infrastructure and vehicle features as input and predicts pixel-level optical flow to correct spatial misalignment.

Given the vehicle BEV feature $F_v \in \mathbb{R}^{C \times H \times W}$ and the infrastructure BEV feature $F_i \in \mathbb{R}^{C \times H \times W}$, we first utilize a shared convolutional block to preprocess these features from different viewpoints while reducing their dimension and resolution. Compressed features F_v^c, F_i^c are only 1/4 of the original resolution and their channel dimensions are also reduced, which ensures the computational efficiency of pixel-wise similarity calculation in the proposed DPSA module. Then, we calculate the pixel-wise correlation between the two features through the similarity block after normalization to obtain a cost volume, $S \in \mathbb{R}^{(H \times W) \times H \times W}$:

$$S(u_v \cdot W + v_v, u_i, v_i) = \frac{(F_v^c[u_v, v_v])^T \cdot F_i^c[u_i, v_i]}{\|F_v^c[u_v, v_v]\| \cdot \|F_i^c[u_i, v_i]\|} \quad (5)$$

where u_v and v_v are the indices of compressed vehicle feature, and u_i and v_i are the indices of compressed infrastructure feature. The cost volume is projected to low dimension by some convolutional layers and concatenated with the original vehicle and infrastructure BEV features. After processing by a fuse block built by convolutional layers, the feature is averaged along its spatial dimension and decoded to 3 DoF relative pose $[\Delta x, \Delta y, \Delta \alpha] \in \text{SE}(2)$ between vehicle coordinate and infrastructure coordinate in BEV space.

$$[\Delta x, \Delta y, \Delta \alpha] = \text{MLP}(\text{SpatialAvg}(\text{Conv}([\text{MLP}(S), F_i]))) \quad (6)$$

We generate the BEV flow based on the estimated pose and warp the infrastructure BEV features via the generated BEV flow.

$$\begin{bmatrix} x_{\text{temp}} \\ y_{\text{temp}} \end{bmatrix} = R(\Delta \alpha) \cdot \begin{bmatrix} (H - v) \cdot s_h \\ (W/2 - u) \cdot s_w \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (7)$$

$$u' = W - \left(\frac{y_{\text{temp}}}{s_w} + \frac{W}{2} \right) \quad (8)$$

$$v' = H - \left(\frac{x_{\text{temp}}}{s_h} + \frac{H}{2} \right) \quad (9)$$

where $R(\Delta \alpha)$ is the rotation matrix of $\Delta \alpha$: $R(\Delta \alpha) = \begin{bmatrix} \cos \Delta \alpha & -\sin \Delta \alpha \\ \sin \Delta \alpha & \cos \Delta \alpha \end{bmatrix} \in \mathbb{R}^{2 \times 2}$; s_h, s_w are the scale factor between BEV features and real-world perception range; (u, v) is the original pixel while (u', v') is the warped pixel.

By learning relative pose between vehicle and infrastructure in BEV space in an end-to-end framework, the spatial misalignment between vehicle and infrastructure can be well solved, significantly reducing human efforts in calibration and localization.

3.3.3 Differentiable flow-based spatial alignment (DFSA) module

We also design a variant model that directly estimates BEV flow instead of relative pose, called DFSA module. The only difference between DPSA module and DFSA module is Eq. (10), that is, the BEV flow is directly decoded from the concatenated features:

$$F_{\text{flow}} = \text{Conv}([\text{MLP}(S), F_i]) \quad (10)$$

Feature alignment: The infrastructure features are aligned using the predicted optical flow through bilinear interpolation:

$$F_{\text{inf_aligned}} = \text{warp}(F_{\text{inf}}, F_{\text{flow}}) \quad (11)$$

Supervision strategy: We utilize the geometric affine transformation results as ground truth to supervise optical flow learning:

$$L_{\text{flow}} = \|F_{\text{flow}} - F_{\text{flow_geo}}\|_1 \quad (12)$$

where $F_{\text{flow_geo}}$ is computed from the geometric transformation matrix.

3.4 Feature fusion strategy

Multi-scale feature processing: Before fusion, both vehicle-end and infrastructure-end features are downsampled by a factor of 0.25 to reduce computational complexity while maintaining sufficient spatial resolution for accurate alignment. This downsampling is implemented using interpolate with bilinear mode.

Pixel-weighted fusion: The aligned features are concatenated along the channel dimension and processed through a PixelWeightedFusion module, which consists of a 3×3 convolutional layer followed by batch normalization and ReLU activation. This module learns optimal fusion weights for different spatial locations and feature channels.

The final fused features are computed as

$$F_{\text{fused}} = \text{PixelWeightedFusion}(\text{concat}([F_{\text{veh}}, F_{\text{inf_aligned}}])) \quad (13)$$

where $F_{\text{veh}} \in \mathbb{R}^{B \times 384 \times H \times W}$ represents vehicle-end features and $F_{\text{inf_compressed}} \in \mathbb{R}^{B \times 384 \times H \times W}$ denotes the spatially aligned infrastructure features.

Mathematical representation: The fusion process can be expressed as

$$F_{\text{fused}} = \sigma \left(W_{\text{fusion}} \cdot \begin{bmatrix} F_{\text{veh}} \\ F_{\text{inf_aligned}} \end{bmatrix} + b_{\text{fusion}} \right) \quad (14)$$

where $W_{\text{fusion}} \in \mathbb{R}^{384 \times 768}$ and $b_{\text{fusion}} \in \mathbb{R}^{384}$ are learnable parameters, and σ represents the ReLU activation function.

Feature concatenation: The concatenation operation combines vehicle features (384 channels) with aligned infrastructure features (384 channels) along the channel dimension, resulting in 768-channel features that are then processed by the fusion module.

3.5 3D object detection output

Detection head: The fused features are fed into an Anchor3DHead that generates 3D bounding box predictions. The head utilizes predefined 3D anchors with specific sizes and orientations for different object classes (Car, Truck, Van).

Anchor generation: For each class, anchors are generated with predefined dimensions and rotations:

Pedestrian: [0.6, 0.8, 1.73] meters with rotations [0, 1.57]

Cyclist: [0.6, 1.76, 1.73] meters with rotations [0, 1.57]

Car: [1.6, 3.9, 1.56] meters with rotations [0, 1.57]

Multi-task prediction: The detection head simultaneously

predicts classification scores, 3D bounding box regression parameters, and direction classification, enabling comprehensive 3D object detection. The head configuration includes Focal Loss for classification ($\gamma = 2.0$, $\alpha = 0.25$), Smooth L1 Loss for regression ($\beta = 0.111$), and Cross-entropy loss for direction prediction.

Output formulation: The detection outputs can be formulated as

$$P_{\text{cls}} = \sigma(W_{\text{cls}} \cdot F_{\text{fused}} + b_{\text{cls}}) \quad (15)$$

$$P_{\text{reg}} = W_{\text{reg}} \cdot F_{\text{fused}} + b_{\text{reg}} \quad (16)$$

$$P_{\text{dir}} = \sigma(W_{\text{dir}} \cdot F_{\text{fused}} + b_{\text{dir}}) \quad (17)$$

where P_{cls} , P_{reg} , and P_{dir} represent classification probabilities, regression parameters, and direction predictions, respectively.

Bbox assignment: The training employs MaxIoUAssigner with different IoU thresholds for each class: Pedestrian (0.5), Cyclist (0.5), and Car (0.6), ensuring appropriate positive and negative sample assignment.

Training strategy: The training objective combines multiple loss components.

- Classification loss: Focal Loss for handling class imbalance with $\gamma = 2.0$ and $\alpha = 0.25$.
- Regression loss: Smooth L1 Loss for 3D bounding box regression with $\beta = 0.111$.
- Direction loss: Cross-entropy loss for orientation prediction.
- Flow supervision loss: L1 loss between predicted optical flow and geometric flow ground truth (weight: 0.01).

Total loss: The overall training loss is computed as

$$L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{reg}} \cdot L_{\text{reg}} + \lambda_{\text{dir}} \cdot L_{\text{dir}} + \lambda_{\text{flow}} \cdot L_{\text{flow}} \quad (18)$$

where $\lambda_{\text{reg}} = 2.0$, $\lambda_{\text{dir}} = 0.2$, and $\lambda_{\text{flow}} = 0.01$ are the loss weights.

Training configuration: The network is trained using the Adam optimizer with a learning rate of 0.003, weight decay of 0.01, and gradient clipping at 35. The training follows a two-stage strategy: first training the base model without temporal latency, then fine-tuning with latency simulation.

Data augmentation: During training, we apply various augmentations including random rotation, scaling, and point cloud dropout to improve model robustness and generalization capability.

Evaluation metrics: The model performance is evaluated using Car 3D IoU AP and Car BEV IoU AP metrics on the DAIR-V2X dataset, providing comprehensive assessment of both 3D and Bird's-Eye-View detection accuracy.

This training strategy ensures that the model learns to effectively handle spatial misalignment through optical flow estimation while maintaining high detection accuracy in vehicle-infrastructure cooperative scenarios.

3.6 Occupancy flow output

We propose a novel OccupancyHead architecture that extends the FFNet-VIC3D framework for comprehensive spatio-temporal occupancy prediction in V2X cooperative perception scenarios. Existing motion/trajectory prediction works mostly output agent-level trajectories, such as the kinematics-aware multigraph attention residual network for heterogeneous trajectory prediction, but such outputs are difficult to directly integrate with feature-level fusion of perception ends. The proposed head operates as a multi-task learning module that simultaneously predicts current occupancy, motion flow, and future occupancy states, enabling joint spatiotemporal modeling of dynamic traffic scenes.

Multi-task architecture design. The OccupancyHead comprises three specialized prediction branches processing high-dimensional features from the backbone network (384 channels) to produce dense occupancy predictions: (1) an occupancy classification branch outputting 4-class semantic occupancy maps at 256×256 resolution, (2) a motion flow estimation branch predicting 2D optical flow vectors, and (3) a future occupancy prediction branch generating 4-frame future occupancy sequences. The network architecture can be formulated as

$$O_t = \text{Softmax}(\text{Conv}_{3 \times 3}(\text{Upsample}(F))) \in \mathbb{R}^{B \times 11 \times 256 \times 256} \quad (19)$$

$$F_t = \text{Conv}_{3 \times 3}(\text{Upsample}(F)) \in \mathbb{R}^{B \times 2 \times 256 \times 256} \quad (20)$$

$$O_{t+1:t+4} = \text{Softmax}(\text{Conv}_{3 \times 3}(\text{Upsample}(F))) \in \mathbb{R}^{B \times 44 \times 256 \times 256} \quad (21)$$

where $F \in \mathbb{R}^{B \times 384 \times H \times W}$ represents the fused backbone features, and the upsample module (a component that increases the spatial resolution of feature maps) progressively upscales spatial resolution through three transposed convolution layers.

Spatial resolution and coordinate system. The head operates on a standardized coordinate system defined by the point cloud range (m-meter):

$x \in [0, 92.16]$ m, $y \in [-46.08, 46.08]$ m, $z \in [-3.92, 1.00]$ m, corresponding to a $92.16 \text{ m} \times 92.16 \text{ m}$ coverage area with 0.16 m voxel resolution. This yields an intermediate feature resolution of 576×576 from the backbone, which is subsequently downsampled to 256×256 for final occupancy prediction, providing a spatial resolution of 0.36 m per occupancy grid cell.

Class-imbalanced loss function design. To address the severe class imbalance inherent in occupancy prediction (where background pixels dominate the scene), we implement a hybrid loss strategy combining weighted CrossEntropy Loss with Dice Loss. The class weights are empirically set as $w = [0.1, 2.0, 1.5, 1.5]$ for the 4 classes (background + 3 object categories), emphasizing vehicle-related classes while suppressing rare categories. The total occupancy loss is formulated as

$$L_{\text{occ}} = L_{\text{CE}}(O_t, Y_t) + \lambda_d L_{\text{Dice}}(O_t, Y_t) \quad (22)$$

where $\lambda_d = 0.5$ controls the Dice Loss contribution, and the weighted CrossEntropy Loss is computed as

$$L_{\text{CE}} = \frac{1}{N} \sum_{i=1}^N w_{y_i} \log(p_{y_i}) \quad (23)$$

Motion flow prediction. The motion flow branch employs SmoothL1Loss with $\beta = 0.01$ to handle the sparse nature of optical flow predictions

$$L_{\text{flow}} = \frac{1}{N} \sum_{i=1}^N \text{SmoothL1}(F_{\text{pred},i} - F_{\text{gt},i}, \beta) \quad (24)$$

This design is particularly effective for small displacement values commonly encountered in vehicle motion prediction.

Future occupancy prediction. The future prediction branch leverages motion flow information to model temporal dynamics, predicting occupancy states for 4 consecutive future frames.

Each future frame prediction utilizes the same loss formulation as current occupancy, with equal loss weights (1.0) for all temporal predictions:

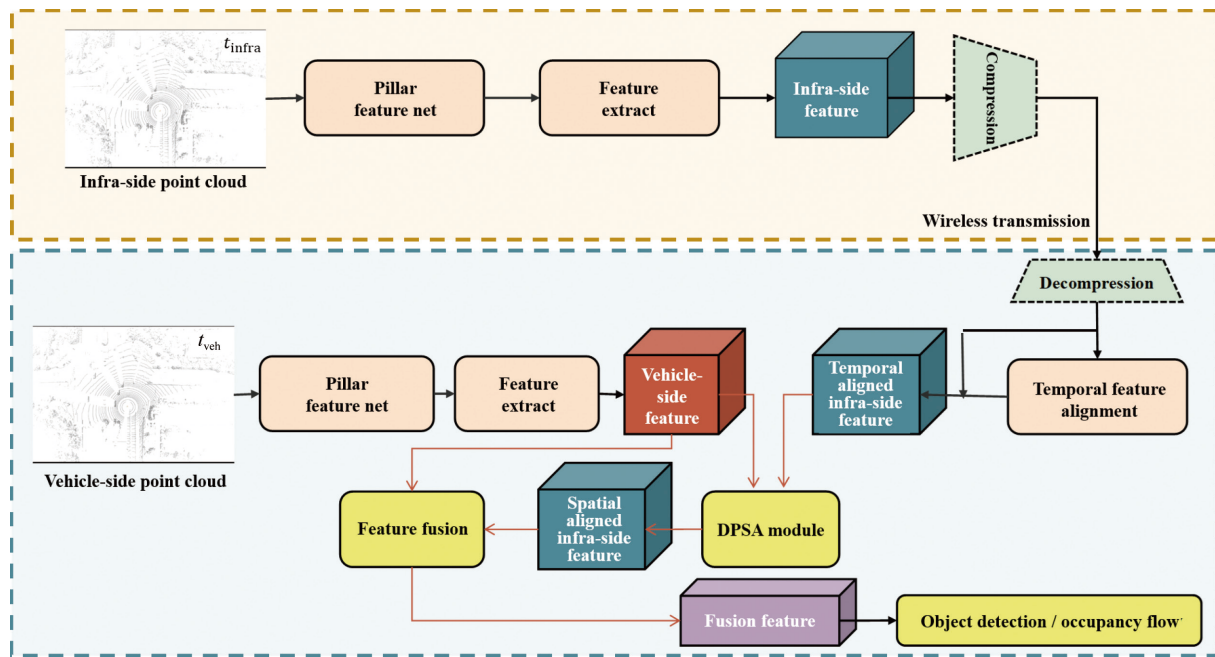


Fig. 3 Overall pipeline of the proposed DPSANet framework.

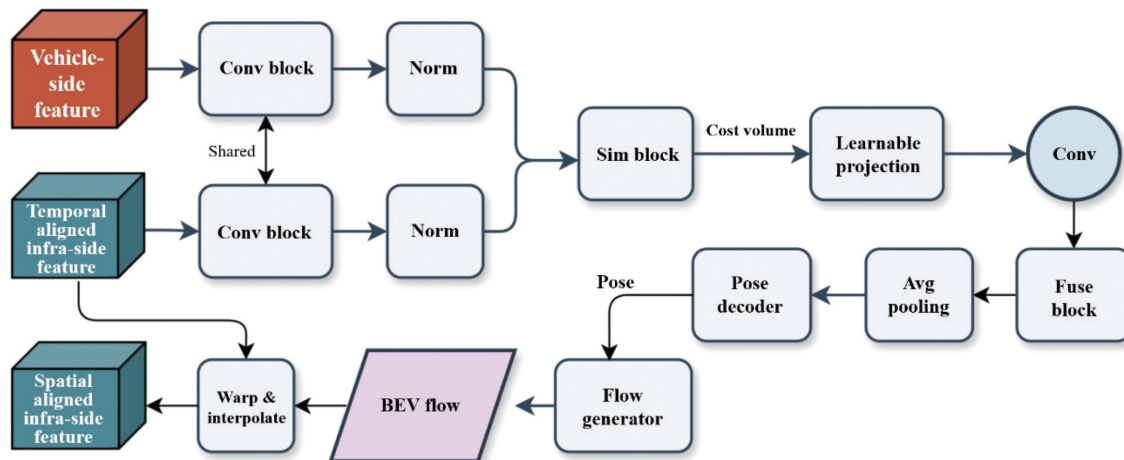


Fig. 4 Detailed architecture of the DPSA module.

$$L_{\text{future}} = \frac{1}{4} \sum_{t=1}^4 L_{\text{occ}}(O_{t+1}, Y_{t+1}) \quad (25)$$

Network integration and training. The OccupancyHead integrates seamlessly with the FFNet-VIC3D backbone, processing 384-channel fused features from vehicle-infrastructure cooperative perception. The complete loss function combines all three tasks:

$$L_{\text{total}} = L_{\text{occ}} + L_{\text{flow}} + L_{\text{future}} \quad (26)$$

This multi-task learning approach enables the network to jointly optimize spatial occupancy classification, temporal motion modeling, and future state prediction, providing comprehensive spatiotemporal understanding essential for autonomous driving in cooperative V2X scenarios.

4 Experiments

4.1 Experiment settings

4.1.1 Dataset

DAIR-V2X (Yu et al., 2022) is a large-scale vehicle-infrastructure

cooperative perception dataset that is composed of real-world images and point clouds. Encompassing over 100 scenes and 18,000 data pairs, it concurrently captures data from infrastructure and vehicle sensors at an equipped intersection as an autonomous driving vehicle traverses through. All frames are annotated with 10 three-dimensional (3D) object classes. Moreover, 9311 pairs possess cooperative 3D annotations (covering cars, buses, trucks, and vans) for both vehicle and infrastructure views. The dataset is partitioned into training, validation, and test sets at a ratio of 5:2:3, and model evaluation is carried out on the validation set.

V2X-Seq (Yu et al., 2023) stands as the world's inaugural real-scenario temporal vehicle-to-everything (V2X) dataset, which consists of two sub-datasets. The first is a Temporal Perception Dataset utilized for 3D detection and tracking. The second is a Trajectory Prediction Dataset (covering 28 intersections, with over 200,000 ten-second clips, 8 types of obstacle trajectories at a frequency of 10 Hz, along with traffic light data). It offers high-quality cooperative trajectory ground truth and supports evaluations for V2X temporal perception and trajectory prediction tasks.

4.1.2 Implementation details

Framework and codebase. We adopted MMDetection3D as our codebase and developed our occupancy prediction model based on the FFNet-VIC3D architecture. This model expands the V2XVoxelNet detector to enable multi-task learning, encompassing occupancy prediction, motion flow estimation, and future frame prediction.

Training configuration. Our occupancy prediction model was trained on the V2X-Seq-SPD dataset over 32 epochs. We utilized the AdamW optimizer with an initial learning rate of 1×10^{-5} and a weight decay of 1×10^{-2} . To address memory limitations, we employed a batch size of 1 per GPU, along with 4 data loading workers per GPU. The learning rate scheduling adhered to a cyclic policy with a target ratio of $(10, 1 \times 10^{-4})$, a cyclic count of 1, and a step ratio up of 0.4. Gradient clipping was applied using the L2 norm with a maximum norm of 1.0.

Model architecture. For the LiDAR-based occupancy prediction task, we selected SECOND as our 3D backbone in conjunction with PointPillars voxelization. The features were represented as a Bird's-Eye-View (BEV) map with dimensions of (576, 576) and a resolution of 0.16 m/pixel in both length and width. The point cloud range was set to $x \in [0, 92.16]$, $y \in [-46.08, 46.08]$, $z \in [-3.92, 1]$ meters, covering an area of (92.16 m $\times$ 92.16 m) around the vehicle. The voxel size was configured as [0.16 m, 0.16 m, 4.0 m], with a maximum of 100 points per voxel and 40,000 voxels per frame.

Multi-task learning. Our model concurrently predicts: (1) the occupancy of the current frame with 4 semantic classes (background + 3 object categories), (2) the motion flow between consecutive frames, (3) the future occupancy for the next 4 frames, and (4) the spatial flow for infrastructure-to-vehicle feature alignment. The occupancy head employed a combination of CrossEntropyLoss and DiceLoss with class weights to mitigate class imbalance, while motion flow prediction utilized SmoothL1Loss.

Feature fusion strategy. We implemented two fusion modes: geometric flow-based alignment using calibration data and learned spatial flow estimation. The geometric flow mode leveraged ground truth pose transformations for precise spatial alignment, while the learned mode employed a PoseFlow_SVD network to estimate spatial transformations. Feature fusion supported three modes: 'fusion' (both vehicle and infrastructure), 'veh_only' (vehicle only), and 'inf_only' (infrastructure only).

Hardware configuration. All training and evaluation processes were conducted on NVIDIA A800-SXM4-80GB GPUs. The model supported both single-sequence training mode for debugging purposes and multi-sequence training mode for full-dataset training using sequence-level data splits.

Evaluation metrics. The performance of occupancy prediction was evaluated using multiple metrics, including occupancy mean

Intersection over Union (mIoU), overall accuracy, motion flow endpoint error (EPE), motion flow accuracy at 1-pixel and 3-pixel thresholds, and future prediction temporal consistency. The model achieved real-time inference while maintaining high accuracy across all prediction tasks.

Data processing. The dataset supported both single-sequence and multi-sequence training modes, with automatic sequence detection and filtering based on data availability. Future frame prediction necessitated temporal consistency checking to ensure an adequate number of future frames were available for training. The model handled missing data gracefully by providing default values for unavailable occupancy or flow data.

4.2 Comparison with SOTA

4.2.1 Object detection results with SOTA

To assess the performance of the proposed DPSANet (ours) in vehicle-infrastructure collaborative 3D object detection, comparisons are conducted with state-of-the-art (SOTA) methods on the DAIR-V2X dataset, as presented in Table 1. The table presents key metrics, including fusion type, Required Position Precision (a crucial precondition for spatial-aligned feature fusion in most existing methods), average precision (AP) at an Intersection over Union (IoU) threshold of 0.5, average precision at an IoU threshold of 0.7, and communication volume (Comm, measured on a $\log_2$ scale).

A significant difference is observed in location precision. Most SOTA middle fusion methods (e.g., Where2comm, DiscoNet, V2VNet (Wang et al. 2020), V2X-ViT, TransIFF, FFNet) depend on high location precision to guarantee accurate feature alignment. This is a demanding and technically arduous requirement for real-world implementation. In stark contrast, our DPSANet operates under low location precision, simulating practical scenarios where precise vehicle-infrastructure pose estimation is either unobtainable or inherently prone to errors.

Despite this limitation, DPSANet attains competitive Average Precision (AP) performance, specifically an AP at an Intersection over Union (IoU) threshold of 0.5 reaching 59.51 and an AP at an IoU threshold of 0.7 reaching 50.86. These outcomes approximate the performance of methods that rely on high location precision (e.g., TransIFF: an AP (IoU = 0.5) of 59.62 and an AP (IoU = 0.7) of 46.03; FFNet: an AP (IoU = 0.5) of 63.87 and an AP (IoU = 0.7) of 54.58) and significantly outperform late-fusion baselines (PointPillars: an AP (IoU = 0.5) of 54.32 and an AP (IoU = 0.7) of 44.58).

Simultaneously, DPSANet maintains favorable communication efficiency, with a communication (Comm) value of 14.96. This value is lower than that of FFNet (Comm = 16.87) and Where2comm (Comm = 18.19), and it is substantially reduced in comparison to methods with fixed high communication overhead

Table 1 Object detection results with SOTA on DAIR-V2X dataset

Method	Reference	Fusion type	Required position precision	AP@IoU = 0.5	AP@IoU = 0.7	Comm ($\log_2$)
PointPillars	CVPR 2019	Veh.-only	—	50.03	43.57	0
PointPillars	CVPR 2019	Late	High (≤ 0.2 m)	54.32	44.58	8.39
Where2comm	NIPS 2022	Middle	High (≤ 0.2 m)	58.46	44.46	18.19
DiscoNet	NIPS 2021	Middle		54.29	44.88	24.62
V2VNet	ECCV 2020	Middle		56.01	42.25	24.62
V2X-ViT	ECCV 2022	Middle		54.26	43.35	24.62
TransIFF	ICCV 2023	Middle		59.62	46.03	12.39
FFNet	NIPS 2023	Middle		63.87	54.58	16.87
DPSANet (ours)		Middle	Low (≤ 2 m)	59.51	50.86	14.96

(DiscoNet, V2VNet, V2X - ViT: all with Comm = 24.62).

This indicates the robustness of DPSANet: It effectively alleviates the dependence on high-precision spatial calibration (a crucial practical bottleneck for cooperative perception systems) while still achieving competitive detection accuracy and communication efficiency. Such advantages make DPSANet highly adaptable to real-world scenarios where perfect location precision is seldom attainable.

4.3 Ablation studies

4.3.1 Effectiveness of the DPSA module in Object Detection Task

To verify the effectiveness of the proposed DPSA module in vehicle-infrastructure cooperative perception tasks, comparative experiments were devised to explore two core aspects: the impact of spatial synchronization and the performance of different spatial synchronization strategies. The experimental results are presented in Table 2.

4.3.1.1 Influence of spatial synchronization

Initially, a baseline comparison was carried out between the Spatial Misaligned group and the Spatial Aligned group, encompassing G.T. Pose, DFSANet, and DPSANet (ours). The Spatial Misaligned approach directly fuses uncalibrated infrastructure-side and vehicle-side features, bypassing spatial synchronization. Conversely, all Spatial Aligned approaches execute feature fusion only subsequent to attaining spatial alignment through their respective strategies.

As presented in Table 2, all Spatial Aligned approaches significantly outperform the Spatial Misaligned approach across all evaluation metrics, including the Average Precision for Cars (AP@Car) under both 3D IoU and BEV IoU thresholds. For instance, at a 3D IoU threshold of 0.5, the AP@Car of our Spatial Aligned approach is 6.05 percentage points higher than that of the Spatial Misaligned approach. At a BEV IoU threshold of 0.7, this disparity broadens to 5.07 percentage points. This validates that spatial synchronization is a crucial precondition for enhancing the performance of vehicle-infrastructure cooperative perception. Aligned features eliminate spatial offsets among multi-source sensing data, facilitating more robust feature fusion and consequently improving perception accuracy.

4.3.1.2 Performance of distinct spatial synchronization strategies

To further validate the superiority of the proposed DPSA module, a comparative analysis was conducted on the three sub - methods within the Spatial Aligned group, as elaborated below:

Spatial aligned—G.T. Pose: This approach utilizes ground-truth (G.T.) calibrated vehicle-road relative poses for spatial synchronization. Theoretically, it attains the optimal alignment effect, which is manifested in its leading Average Precision (AP) performance (e.g., 3D Intersection over Union (IoU) AP@0.5: 54.81; Bird's-Eye View (BEV) IoU AP@0.5: 63.87). Nevertheless, it depends entirely on manual pose annotation, which is not only labor-intensive and costly but also lacks generalizability to dynamic real-world scenarios (e.g., diverse road conditions, transient vehicle motion). Consequently, this method is not viable for large-scale engineering deployment.

Spatial aligned—DFSANet: This method realizes spatial synchronization by appending a convolutional layer after the Fuse Block to generate an $H \times W \times 2$ flow field. Although its AP metrics are slightly higher than those of the Spatial Misaligned method (e.g., 3D IoU AP@0.7: 28.44 vs. 28.17), it demonstrates substantial performance deficiencies compared to the proposed method. Additionally, its DFSA module necessitates 40.9M parameters and an average inference time of 59.18 ms, far exceeding the efficiency constraints of real-time vehicle-road cooperative systems.

Spatial aligned—DPSANet (ours): The proposed DPSA module adopts a two-stage spatial synchronization mechanism. Firstly, spatial average pooling and fully connected layers are deployed after the Fuse Block to estimate the 3-degree-of-freedom (3-DoF) vehicle-road relative pose.

Secondly, geometric transformation is applied to map the estimated pose into an $H \times W \times 2$ flow field for precise spatial alignment. By integrating the geometric prior of 3-DoF pose deviation (a fundamental characteristic of vehicle-road relative positioning), this module effectively alleviates alignment errors.

Quantitatively, the proposed method outperforms BEV Flow across all key metrics. At a 3D IoU threshold of 0.7, it achieves a 2.92-percentage-point improvement (31.36 vs. 28.44). At a BEV IoU threshold of 0.7, this improvement reaches 4.71 percentage points (50.86 vs. 46.15). In terms of efficiency, the proposed module has 40.7M parameters (slightly fewer than BEV Flow's 40.9M) and reduces the average inference time by 95.8% (from 59.18 to 2.50 ms), which is a crucial advantage for meeting the real-time requirements of engineering applications.

To simulate localization inaccuracies, Gaussian noise is added to the ground-truth vehicle pose. Translation perturbations $\varepsilon_t = (\varepsilon_x, \varepsilon_y, \varepsilon_z)^T$ with $\varepsilon_x, \varepsilon_y, \varepsilon_z \sim \mathcal{N}(0, \delta_t^2)$ (unit: m) and rotation perturbations $\varepsilon_r = (\varepsilon_\phi, \varepsilon_\theta, \varepsilon_\psi)^T$ with $\varepsilon_\phi, \varepsilon_\theta, \varepsilon_\psi \sim \mathcal{N}(0, \delta_r^2)$ (unit: $^\circ$) are applied independently. The horizontal axis in Fig. 5 reports the paired standard deviation (δ_t, δ_r); e.g., "0.1" denotes $\delta_t = 0.1$ m, $\delta_r = 0.1^\circ$. As shown in Fig. 5, this ablation experiment assesses the influence of localization errors on object detection performance by employing Average Precision (AP). In the context of Bird's-Eye View (BEV) detection (left panel), the G.T. Pose group demonstrates a notable decline in AP as errors increase (e.g., from approximately 63.5 to 48 at an Intersection over Union (IoU) of 0.5), and higher IoU thresholds (0.7) display greater sensitivity. Conversely, the proposed DPSANet sustains a stable BEV AP (approximately 59.5 at IoU = 0.5, approximately 50.5 at IoU = 0.7) across the range of errors and outperforms the spatial misalignment baseline, indicating robust tolerance to localization uncertainty. Similar patterns are observed in 3D detection. Overall, these results confirm the outstanding robustness of DPSANet to localization errors in both modalities, offering reliable methodological support for autonomous driving detection under conditions of localization uncertainty.

4.3.2 Effectiveness of the DPSA module in occupancy flow task

To evaluate the DPSA module's efficacy for occupancy tasks, comparative experiments included four methods: Veh_only

Table 2 Performance comparison of methods in object detection task

Method	AP@Car 3D IoU threshold			AP@Car BEV IoU threshold			Params. of DPSA module	Average time (ms)
	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7		
Spatial misaligned	55.35	47.42	28.17	56.56	52.78	45.79	—	—
G.T. Pose	63.92	54.81	28.94	66.37	63.87	54.58	—	—
Spatial aligned DFSANet	56.05	48.09	28.44	57.49	53.57	46.15	40.9M	59.18
DPSANet (ours)	62.68	53.47	31.36	64.45	59.51	50.86	40.7M	2.50

(vehicle-side baseline), Veh & Infra Fusion with ground truth pose (G.T. Pose), unsynchronized fusion (Spatial misaligned), and our DPSA-integrated fusion. Table 3 summarizes the quantitative comparison, where G.T. Pose and Veh_only serve as the upper- and lower-bound references, respectively; our DPSA-integrated fusion consistently outperforms the Spatial misaligned setting in both occupancy accuracy (mIoU) and motion flow error (EPE), and substantially narrows the gap toward the G.T. Pose upper bound. Since no existing cooperative perception method employs occupancy flow for feature fusion, direct comparison with a “cooperative occupancy flow SOTA” is not possible. Instead, our evaluation focuses on two aspects: (1) the improvement of collaborative fusion over single-vehicle occupancy prediction, and (2) the benefits of spatial synchronization (ours vs. misaligned) under realistic conditions.

4.3.2.1 Comparative analysis of baseline and fusion approaches

Fusion methods were compared against the Veh_only baseline to highlight multi-source data advantages. Fusion methods outperform Veh_only by leveraging infrastructure data to enhance detection of occluded/infrequent objects (e.g., trucks/vans), improving mIoU, reducing motion flow EPE, and stabilizing future occupancy prediction. Multi-source fusion

inherently boosts cooperative perception performance, laying the groundwork for further gains via spatial synchronization.

4.3.2.2 Critical role of spatial synchronization

Spatial alignment impact was analyzed by comparing G.T. Pose (ground-truth synchronization) and Spatial misaligned (unsynchronized fusion). Spatial misaligned directly fuses uncalibrated features, while G.T. Pose achieves optimal alignment via annotated vehicle-road poses. Table 3 shows G.T. Pose outperforms Spatial misaligned across metrics: mIoU (0.5526 vs. 0.4539), motion flow EPE (0.01044 vs. 0.0113). This confirms spatial synchronization enhances fusion quality; unsynchronized fusion (Spatial misaligned) only marginally improves mIoU over Veh_only (0.4539 vs. 0.4260), emphasizing synchronization necessity.

4.3.2.3 Performance of the proposed DPSA module

The proposed DPSANet achieves end-to-end spatial synchronization without ground-truth annotations by estimating 3-degree-of-freedom vehicle-road pose via spatial average pooling and fully connected layers, then computing flow maps using geometric priors for alignment.

Performance-wise, DPSA (ours) outperforms Spatial misaligned in mIoU (0.4726 vs. 0.4539) and motion flow EPE (0.01104 vs. 0.0113), indicating geometric priors reduce alignment errors. While slightly lower than G.T. Pose (mIoU 0.4726 vs. 0.5526), it avoids manual annotation costs for real-world viability. Future occupancy accuracy (0.9855) is comparable to G.T. Pose (0.9801) and Spatial misaligned (0.9866), validating long-term prediction stability.

4.3.2.4 Visualization of the proposed method

To conduct an intuitive assessment of the performance of the proposed DPSANet in vehicle-road cooperative perception, visualization results for Occupancy-BEV and motion flow are presented. Three methods are compared: DPSA (ours) (the proposed pose-estimation-based spatial synchronization), G.T. Pose (ground-truth pose-guided spatial synchronization), and Veh_only (vehicle-side perception without roadside collaboration).

In Occupancy-BEV representations (Fig. 6), the proposed DPSANet generates detection results that exhibit a high degree of consistency with the ground truth (G.T.) pose. Both methods can robustly capture the distribution of objects, including occluded targets that the Veh_only method fails to detect. In comparison to the Veh_only method, the DPSANet (ours) leverages roadside perception to improve the detection of occluded objects, which underscores the advantage of cross-node feature fusion in cooperative systems.

Regarding motion flow prediction (Fig. 7), DPSA (ours) retrieves more comprehensive motion state information of vehicles and provides more complete predictions for dynamic objects. This performance is comparable to that of G.T. Pose (which benefits from geometrically optimal ground-truth poses) and significantly outperforms Veh_only, where motion-related details are scarce.

These visual results confirm that DPSANet (ours) can approximate the performance of ground-truth-based spatial synchronization while eliminating the dependence on manual pose annotation. It effectively improves the perception of occluded targets and the accuracy of dynamic motion prediction, outperforming single-vehicle perception in terms of both semantic completeness and motion fidelity.

5 Conclusions

This work addresses key challenges in vehicle-infrastructure cooperative perception by introducing a differentiable pose-based

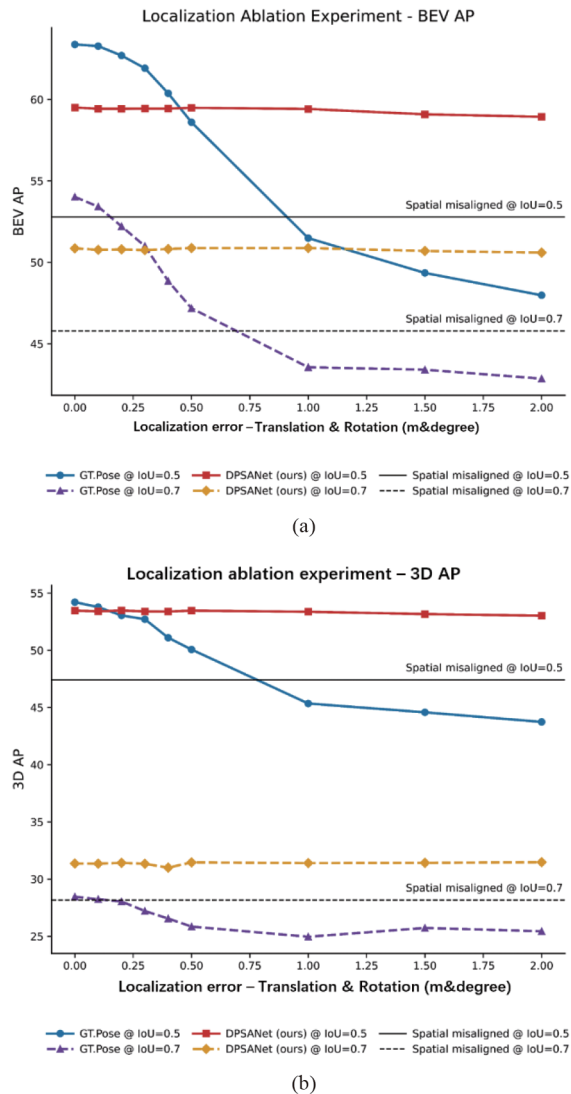


Fig. 5 Influence of localization errors on spatial alignment and cooperative perception performance. (a) Average precision metrics of BEV perception results; (b) average precision metrics of 3D perception results.

Table 3 Performance comparison of methods in occupancy task

Method	Occupancy BEV IoU					Motion flow	Future occupancy
	Background	Car	Truck	Van	mIoU	EPE	Accuracy
Veh_only	0.9896	0.4664	0.1028	0.1454	0.4260	0.016 18	0.9671
G.T. Pose	0.9939	0.579	0.310	0.309	0.552	0.0104	0.9801
Veh & infra fusion	Spatial misaligned	0.9928	0.501	0.177	0.144	0.453	0.0113
	DPSANet (ours)	0.9925	0.521	0.088	0.287	0.473	0.0110
							0.9855

spatial alignment (DPSA) module, which eliminates the dependency on ground-truth poses without compromising performance. Comprehensive evaluations on DAIR-V2X and V2X-Seq datasets demonstrate that our approach surpasses misaligned and GT-dependent fusion methods in critical metrics such as 3D IoU AP@0.7 and mean Intersection over Union (mIoU). The practical utility of the DPSA module is highlighted by its low inference time (2.50 ms) and sustained robustness under low location precision. Furthermore, we validate the superiority of occupancy flow over traditional object detection. By enabling integrated spatiotemporal modeling of occupancy changes and motion flows, our method supports high-accuracy future prediction and resists occlusions through multi-source temporal context. The consistent outperformance of multi-source fusion over vehicle-only baselines underscores the complementary value of cross-node sensing.

Despite these promising results, it is important to acknowledge the limitations of this study. The current model shows constrained capability in modeling non-rigid objects (e.g., pedestrians) and exhibits declining accuracy in long-term future predictions beyond four frames (0.4 s), which is insufficient for full autonomous driving planning tasks requiring 1–2 s horizons.

Additionally, performance may be sensitive to extreme environmental conditions not fully represented in the current datasets. Future work will focus on several directions: integrating temporal attention mechanisms and longer training sequences to extend prediction horizons beyond one second, incorporating multi-modal data (e.g., camera images) to improve alignment robustness, and adapting the framework to diverse V2X communication protocols.

Author contributions

Yanding Yang and **Rujun Yan** contributed equally to this work. **Yanding Yang** is responsible for conceptualization, methodology, software, investigation, and writing – original draft. **Rujun Yan** is responsible for conceptualization, methodology, software, investigation, and validation. **Jinyu Miao** is responsible for methodology, software, and investigation. **Zhiwei Zhao** is responsible for methodology, software, and investigation. **Kewei Wang** is responsible for investigation, validation, and writing – original draft. **Kun Jiang** is responsible for resources, supervision, project administration, and writing – review & editing. **Diange Yang** is responsible for resources. All the authors have approved the final manuscript.

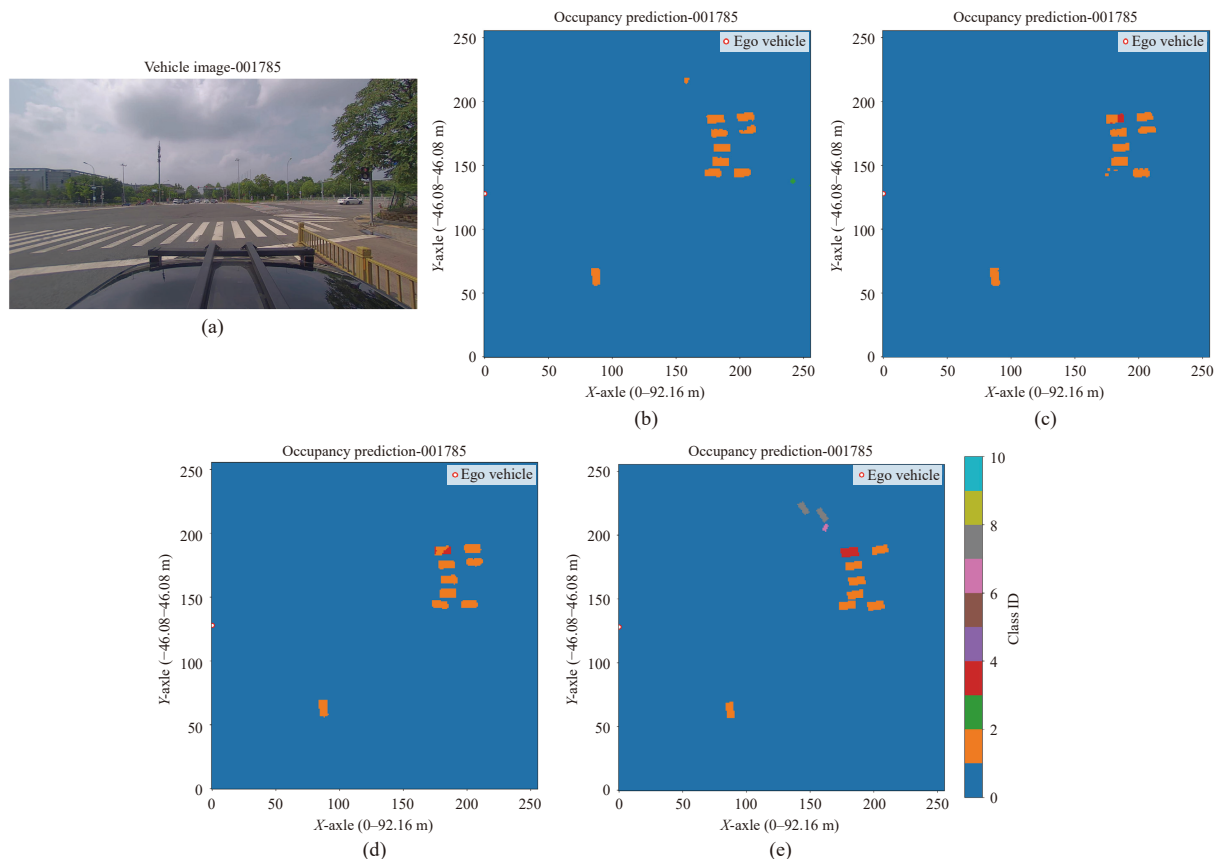


Fig. 6 Visualization results for Occupancy-BEV on DAIR-Seq datasets. (a) Vehicle image; (b) G.T. Pose result; (c) vehicle-only result; (d) DPSANet result; and (e) ground truth.

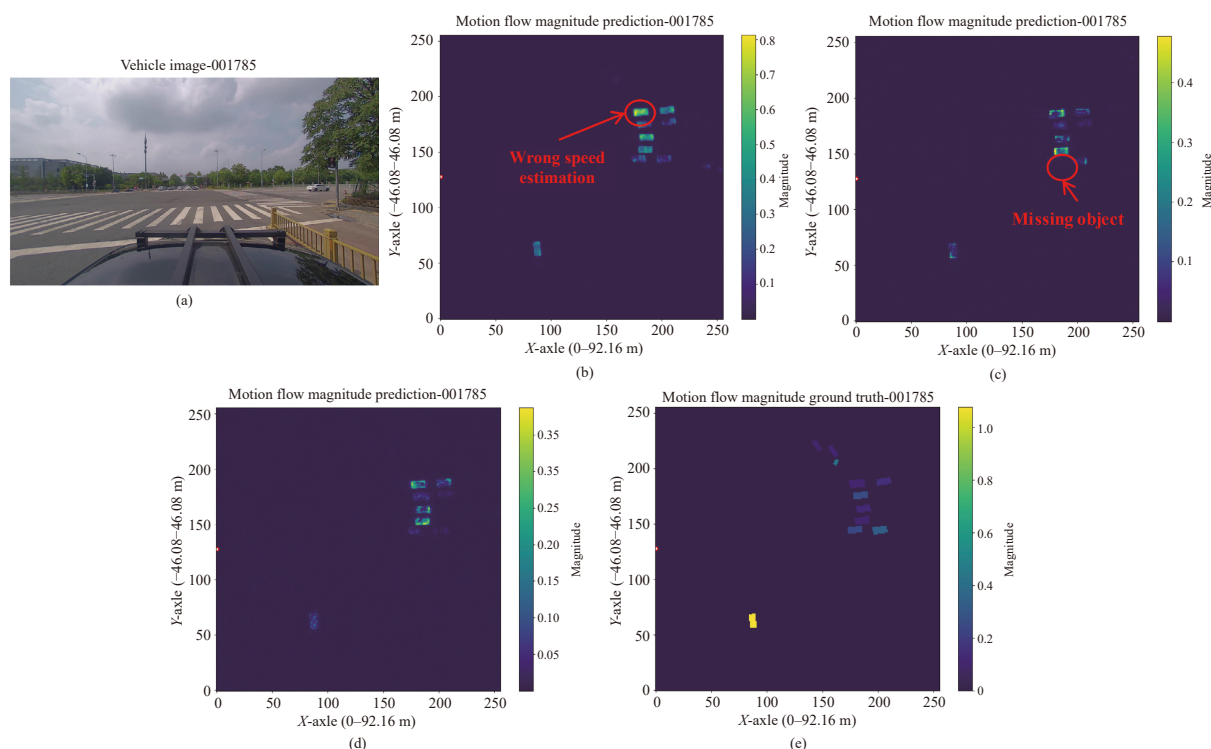


Fig. 7 Visualization results for motion flow on DAIR-Seq datasets. (a) Vehicle image; (b) G.T. Pose result; (c) vehicle-only result; (d) DPSANet result; (e) ground truth.

Replication and data sharing

The source codes and replication package are available on ETS data at <https://ets-data.science.com/ETSD.2025.9190074>.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Nos. 52394264 and 52372414).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

- Barrera, A., Guindel, C., Beltran, J., Garcia, F., 2020. BirdNet+: End-to-end 3D object detection in LiDAR Bird's Eye View. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 1–6.
- Beltrán, J., Guindel, C., Moreno, F. M., Cruzado, D., García, F., De La Escalera, A., 2018. BirdNet: A 3D object detection framework from LiDAR information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 3517–3523.
- Chen, N., Wang, J., Chen, H., Shen, Y., Wang, S., Tan, X., 2024. BEVLOC: End-to-end 6-DoF localization via cross-modality correlation under Bird's Eye View. In: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8155–8159.
- Chen, Q., Tang, S., Yang, Q., Fu, S., 2019. Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 514–524.
- Chen, Z., Shi, Y., Jia, J., 2023. TransIFF: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3D detection with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 18159–18168.
- Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z., 2021. RangeDet: In defense of range view for LiDAR-based 3D object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2898–2907.
- Gao, Z., Jia, B., Xie, D., Wang, W., Wu, J., 2025. A discussion on the complexity and transit mechanisms of urban traffic systems. *Engineering*, **44**, 24–29.
- Gao, X., Zhang, X., Lu, Y., Huang, Y., Yang, L., Xiong, Y., et al., 2024. A survey of collaborative perception in intelligent vehicles at intersections. *IEEE Trans Intell Veh*, <https://doi.org/10.1109/TIV.2024.3395783>
- Hao, H., Wang, Y., Chen, J., 2024. Empowering scenario planning with artificial intelligence: A perspective on building smart and resilient cities. *Engineering*, **43**, 272–283.
- Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S., 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In: *Advances in Neural Information Processing Systems*, **35**, 4874–4886.
- Huang, T., Liu, J., Zhou, X., Nguyen, D. C., Rahimi Azghadi, M., Xia, Y., et al., 2025. Vehicle-to-everything cooperative perception for autonomous driving. *Proc IEEE*, **113**, 443–477.
- Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. Pointpillars: Fast encoders for object detection from point clouds. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12689–12697.
- Lei, Z., Ni, Z., Han, R., Tang, S., Feng, C., Chen, S., et al., 2024. Robust collaborative perception without external localization and clock devices. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), 7280–7286.
- Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W., 2021. Learning distilled collaboration graph for multi-agent perception. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 29541–29552.
- Liang, Z., Zhang, M., Zhang, Z., Zhao, X., Pu, S., 2020. RangeRCNN: Towards Fast and Accurate 3D Object Detection with Range Image Representation. <https://doi.org/10.48550/arXiv.2009.00206>

- Liang, Z., Zhang, Z., Zhang, M., Zhao, X., Pu, S., 2021. RangeIoUDet: Range image based real-time 3D object detector optimized by intersection over union. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7136–7145.
- Liu, S., Ding, Z., Fu, J., Li, H., Chen, S., Zhang, S., Zhou, X., 2024. V2X-PC: Vehicle-to-everything Collaborative Perception via Point Cluster. <https://doi.org/10.48550/arXiv.2403.16635>
- Liu, Z., Sun, Y., Wang, Y., Yang, N., Li, S. E., Zhao, X., 2025. Beyond conventional vision: RGB-event fusion for robust object detection in dynamic traffic scenarios. *Commun Transp Res*, 5, 100202.
- Mao, J., Shi, S., Wang, X., Li, H., 2023. 3D object detection for autonomous driving: A comprehensive survey. *Int J Comput Vis*, 131, 1909–1963.
- Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C. K., 2019. LaserNet: An efficient probabilistic 3D object detector for autonomous driving. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12669–12678.
- Nayak, A., Cattaneo, D., Valada, A., 2024. RaLF: Flow-based global and metric radar localization in LiDAR maps. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), 5097–5103.
- Qiao, D., Zulkernine, F., 2023. Adaptive feature fusion for cooperative perception using LiDAR point clouds. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1186–1195.
- Song, Z., Liu, L., Jia, F., Luo, Y., Jia, C., Zhang, G., et al., 2024. Robustness-aware 3D object detection in autonomous driving: A review and outlook. *IEEE Trans Intell Transport Syst*, 25, 15407–15436.
- Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision, 402–419.
- Teufel, S., Gamerding, J., Volk, G., Bringmann, O., 2023. Collective PV-RCNN: A novel fusion technique using collective detections for enhanced local LiDAR-based perception. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), 1828–1834.
- Wang, T. H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R., 2020. V2VNet: vehicle-to-vehicle communication for joint perception and prediction. In: *Computer Vision – ECCV*, 2020, 605–621.
- Wei, S., Wei, Y., Hu, Y., Lu, Y., Zhong, Y., Chen, S., et al., 2023. Asynchrony-robust collaborative perception via bird's eye view flow. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, 28462–28477.
- Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M. H., Ma, J., 2022. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In: European Conference on Computer Vision, 107–124.
- Yang, B., Luo, W., Urtasun, R., 2018. PIXOR: Real-time 3D object detection from point clouds. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7652–7660.
- Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., et al., 2022. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21329–21338.
- Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., Nie, Z., 2023. Flow-based feature fusion for vehicle-infrastructure cooperative 3D object detection. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, 34493–34503.
- Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5486–5495
- Zhang, B., Huang, H., Liu, C., Zhang, Y., Xu, Z., 2025. InVDriver: Intra-instance aware vectorized query-based autonomous driving transformer. *J Int Con Veh*, 8, 9210060.

Zhang, X., Jiang, K., Sun, K., Liu, M., Shao, Z., Xiao, P., et al., 2021. Intrinsic parameters modeling for LiDAR and point cloud correction method. *Automot Eng*, 43, 1228–1237.

Zheng, Z., Xia, X., Gao, L., Xiang, H., Ma, J., 2024. CooperFuse: A real-time cooperative perception fusion framework. In: 2024 IEEE Intelligent Vehicles Symposium (IV), 533–538.

Zhu, Y., Zhao, C., Du, Y., 2022. Latency impact analysis of point cloud fusion modes for cooperative perception in autonomous driving. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 1024–1029.



Yanding Yang received the M.S. degree of Science from the University of Nottingham, UK, in 2005. He is currently studying for a Ph.D. degree in mechanical engineering at the School of Vehicle and Delivery, Tsinghua University, China. He is currently the Director of Dongfeng Motor Corporation Research & Development Institute, China, a high-level talent in the automotive industry, mainly responsible for the development of new energy and intelligent connected vehicle technology.

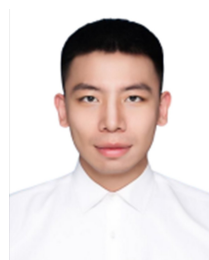


Rujun Yan received the B.E. degree in vehicle engineering (electronic information) from Tsinghua University, China, in 2024. Since 2024, he has been pursuing a M.S. degree in vehicle engineering (autonomous driving) at Tsinghua University, China. His research interests include autonomous driving, vehicle-road cooperative perception, and data-driven vehicle dynamics.



autonomous driving.

Jinyu Miao received the B.S. degree in automation and the M.S. degree in control science and engineering from the School of Automation Science and Electrical Engineering, Beihang University, China, in 2019 and 2022, respectively. He is now currently working toward a Ph.D. degree at the School of Vehicle and Mobility, Tsinghua University, China. His research interests include high-precision localization, vehicle dynamics, and planning for



Zhiwei Zhao received the M.S. degree in mechanical engineering from Beijing Institute of Technology, China, in 2025. He is currently a SLAM Algorithm Engineer at Dongfeng Motor Corporation Research & Development Institute, China. His main research interests include visual SLAM and multi-sensor fusion localizing.



Kewei Wang received the Ph.D. degree in automotive engineering from the School of Automotive Engineering, Wuhan University of Technology, China, in 2021. He is currently working at Dongfeng USharing Technology Co., Ltd., China. He has participated in a number of national major research projects. His research interests include environment perception, multi-sensor fusion, and artificial intelligence.



Kun Jiang received the B.S. degree in mechanical and automation engineering from Shanghai Jiao Tong University, China, in 2011, and the M.S. degree in mechatronics system and the Ph.D. degree in information and systems technologies from the University of Technology of Compiègne (UTC), France, in 2013 and 2016, respectively. He is currently an Associate Professor at Tsinghua University, China. His research interests include autonomous vehicles,

high precision digital map, and sensor fusion.



Diange Yang received the B.S. and Ph.D. degrees from Tsinghua University, China, in 1996 and 2001, respectively. He is currently a Professor at the School of Vehicle and Mobility, Tsinghua University. His research work mainly focuses on intelligent connected vehicles and autonomous driving. He has received numerous awards during his career, including the Distinguished Young Science Technology talent of Chinese Automobile Industry in 2011 and

the Excellent Young Scientist of Beijing in 2010. He was a recipient of the Second Prize of the State Technological Invention Awards of China in 2010 and 2013.