

HeCoFuse: Cross-Modal Complementary V2X Cooperative Perception with Heterogeneous Sensors

Chuheng Wei*, *Member, IEEE*, Ziye Qin, *Member, IEEE*, Walter Zimmer, *Member, IEEE*,
Guoyuan Wu, *Senior Member, IEEE* and Matthew J. Barth, *Fellow, IEEE*,

Abstract—Real-world Vehicle-to-Everything (V2X) cooperative perception systems often operate under heterogeneous sensor configurations due to cost constraints and deployment variability across vehicles and infrastructure. This heterogeneity poses significant challenges for feature fusion and perception reliability. To address these issues, we propose HeCoFuse, a unified framework designed to enable effective cooperative perception across diverse sensor combinations, a unified framework designed for cooperative perception across mixed sensor setups—where nodes may carry Cameras (C), LiDARs (L), or both. By introducing a hierarchical fusion mechanism that adaptively weights features through a combination of channel-wise and spatial attention, HeCoFuse can tackle critical challenges such as cross-modality feature misalignment and imbalanced representation quality. In addition, an adaptive spatial resolution adjustment module is employed to balance computational cost and fusion effectiveness. To enhance robustness across different configurations, we further implement a cooperative learning strategy that dynamically adjusts fusion type based on available modalities. Experiments on the real-world TUMTraf-V2X dataset demonstrate that HeCoFuse achieves 43.22% 3D mAP under the full sensor configuration (LC+LC), outperforming the CoopDet3D baseline by 1.17%, and reaches an even higher 43.38% 3D mAP in the L+LC scenario, while maintaining 3D mAP in the range of 21.74%–43.38% across nine heterogeneous sensor configurations. These results, validated by our first-place finish in the CVPR 2025 DriveX challenge, establish HeCoFuse as the current state-of-the-art on TUM-Traf V2X dataset while demonstrating robust performance across diverse sensor deployments. Code is available at <https://github.com/ChuhengWei/HeCoFuse>.

I. INTRODUCTION

The Vehicle-to-Everything (V2X) enabled cooperative perception leverages information sharing between vehicles and roadside infrastructure to extend the perceptual horizon beyond line of sight and mitigate occlusions inherent to individual sensors [1]–[3]. Recent advances in V2X systems have demonstrated notable gains in detection range, accuracy, and robustness, especially under challenging conditions such as dense traffic, adverse weather, and complex urban scenes [4]. However, these studies predominantly assume that all nodes possess identical or similar sensor configurations, an assumption that barely holds in real-world deployments

due to cost constraints, incremental hardware upgrades, and heterogeneous features in vehicles and infrastructure [5].

The resulting heterogeneity creates scenarios where some nodes may carry only LiDAR, others only cameras, and yet others both sensor types [6]. This variation poses three major challenges: First, asymmetric sensor modalities complicate feature alignment and fusion, as LiDAR and camera data differ in spatial resolution, field of view, and information density. Second, feature inconsistency arises when a modality is absent at certain nodes, rendering naive fusion strategies suboptimal. Third, system robustness must be guaranteed under partial sensor failures or when interacting with nodes of varying capabilities, requiring adaptable fusion mechanisms that degrade gracefully.

To address these challenges, we propose **HeCoFuse**, a unified cooperative perception framework tailored for heterogeneous vehicle–infrastructure (V/I) settings. HeCoFuse extracts bird’s-eye-view (BEV) features from each node via modality-specific encoders, and then performs inter-node fusion through a novel hierarchical attention-based mechanism that dynamically weighs both channel-wise and spatial features according to sensor quality. In parallel, an adaptive spatial resolution module adjusts feature-map scales based on sensor configuration to balance computational cost and information fidelity. During training, we randomly sample nine representative heterogeneous configurations, enabling the model to learn robust fusion strategies that can be generalized across all sensor combinations.

Our main contributions are threefold:

- We introduce **HeCoFuse**, the first unified framework explicitly designed for cooperative perception with heterogeneous sensor configurations in V2X networks.
- We develop two novel mechanisms, Hierarchical Attention Fusion (HAF) and Adaptive Spatial Resolution (ASR), to dynamically integrate cross-modal features and harmonize feature scales under varying sensor availability.
- We conduct extensive experiments on the real-world TUMTraf-V2X dataset [7], showing that HeCoFuse achieves state-of-the-art performance when trained from scratch, and maintains robust accuracy across nine heterogeneous sensor configurations.

II. RELATED WORK

A. Cooperative Perception Strategies

Cooperative perception (CP) methods in V2X systems are typically categorized into three main fusion strategies—early

Chuheng Wei, Ziye Qin, Guoyuan Wu, and Matthew J. Barth are with the College of Engineering, Center for Environmental Research and Technology, University of California at Riverside, Riverside, CA, 92507 USA.

Ziye Qin is also with the School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China.

Walter Zimmer is with the Chair of Robotics, Artificial Intelligence and Real-time Systems, TUM School of Computation, Information and Technology, Technical University of Munich, 85748 Munich, Germany.

*Corresponding author. e-mail: chuheng.wei@email.ucr.edu

fusion, intermediate fusion, and late fusion—which correspond to data-level fusion, feature-level fusion, and fusion of individual agents’ independent outputs, respectively [8], [9]. These strategies reflect different trade-offs between communication load, perception quality, and system flexibility. Early research in CP primarily explored data-level and output-level fusion approaches, while more recent work has gravitated toward intermediate fusion methods due to their balance between performance and communication efficiency.

Early fusion involves collaboration at the input data level, where raw sensor data from multiple agents is shared and aggregated before feature extraction. One representative method, Cooper [10], integrates data from ego vehicles and other connected vehicles in a vehicular network to enhance perception capabilities. By leveraging LiDAR point clouds, Cooper fuses raw sensor data from multiple nodes at the data level. It proposes a 3D object detection framework to process aligned point cloud data, demonstrating the feasibility of performing cooperative perception through point cloud data transmission using existing vehicular network technologies. This method significantly improves detection accuracy and driving safety. To address the high communication overhead of raw data sharing, V2X-PC [11] introduced a novel approach by defining point clusters as message units instead of transmitting raw point cloud data, significantly reducing communication requirements while maintaining perception accuracy.

Intermediate fusion has emerged as the mainstream approach in CP research due to its ability to reduce communication demands while preserving essential perceptual information. Based on fusion principles, intermediate fusion techniques can be broadly classified into three categories. Traditional methods like F-Cooper [12] implement basic feature aggregation techniques such as summation and pooling operations, capitalizing on the compact nature of features to enable real-time edge computing. Attention-based methods, represented by AttentionFusion [13] and V2VFormer [14], employ various attention mechanisms to dynamically integrate multi-sensor information through weight assignment. Topology-based approaches like V2VNet [15] model multiple agents as a graph structure, using graph neural networks to facilitate information exchange. Other intermediate fusion frameworks include PillarGrid [16], CoPercept [17], VIMI [18], HM-ViT [5], CollabGAT [19], and V2X-ViT [3], each proposing unique strategies for effective feature integration.

Late fusion methods integrate detection results after independent processing at each agent, typically using traditional data fusion techniques. These approaches were common during the early stages of CP development, assigning confidence levels to different modules and combining their outputs for final detection. Zhao et al. [20] proposed a method using Dempster-Shafer theory [21], [22] to update detection confidence and improve lane detection accuracy, while VIPS [23] introduced a system encoding simplified object representations for efficient post-detection matching. Other late fusion approaches include Khalifa’s multi-view

cooperative framework [24], Pereira’s Poster platform [25], and CooperFuse [26], which calculates motion and scale consistency across frames for robust bounding box fusion.

Beyond these three primary categories, hybrid fusion methods combine features from different fusion stages, leveraging their respective strengths for enhanced cooperative perception. DiscoNet [27] integrates knowledge distillation to align early and intermediate fusion strategies, while ML-Cooper [28] employs reinforcement learning to dynamically adjust information utilization under varying bandwidth constraints. Additional hybrid approaches include Arnold’s dual early-late fusion method [29] and Liu’s region-based fusion framework [30], which adapt their strategies based on sensor coverage and environmental conditions.

B. Heterogeneous Sensor Fusion

Recent studies have begun to explore heterogeneous cooperative perception in V2X scenarios, where vehicles and infrastructure may be equipped with different sensor modalities and detection networks. HEAD [31] proposes a bandwidth-efficient cooperative perception framework that fuses the classification and regression heads of diverse 3D object detection models, enabling collaboration across vehicles equipped with different sensor types such as LiDAR and camera-based systems. By leveraging self-attention mechanisms and compact feature representations, HEAD achieves comparable perception performance to intermediate fusion while significantly reducing communication overhead. In the method proposed by Luo et al. [32], a vehicle-to-infrastructure (V2I) cooperative perception framework was proposed to address beyond visual range scenarios by combining heterogeneous sensor data. This method enhances YOLOv4 [33] and CenterPoint [34] backbones through novel convolutional modules and performs multi-object post-fusion, demonstrating improvements in urban intersection settings. HM-ViT [5] introduces a unified hetero-modal vehicle-to-vehicle (V2V) cooperative perception framework using a heterogeneous 3D graph transformer to jointly model inter- and intra-agent interactions. Unlike prior approaches limited to homogeneous sensor configurations, HM-ViT effectively integrates LiDAR and image features across multiple dynamic agents, achieving state-of-the-art performance on OPV2V [13].

Despite the progress made by these methods, they still exhibit notable limitations. First, most of them focus on either V2V or V2I scenarios independently, rather than addressing the broader V2X cooperative setting. More importantly, their treatment of heterogeneous modalities is often limited to single-sensor configurations, such as fusing features from one camera and one LiDAR, without supporting more diverse or flexible combinations of multi-camera and multi-LiDAR setups. This restricts their scalability and effectiveness in real-world deployments where sensor configurations vary significantly across agents. To address this gap, our method aims to develop a generalized fusion framework that can accommodate arbitrary combinations of camera and LiDAR sensors across vehicles and infrastructure. By enabling more

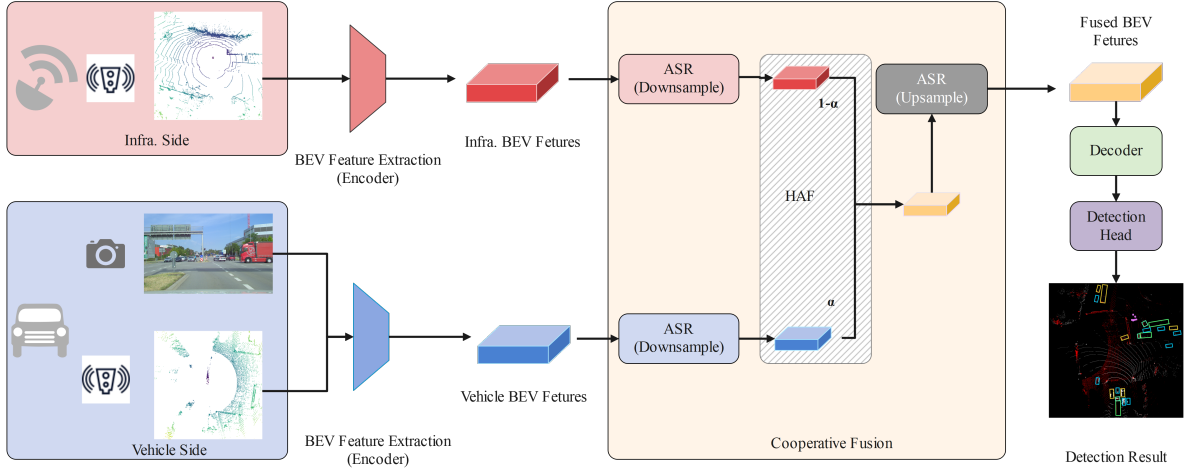


Fig. 1: HeCoFuse framework overview. Here, the LC+L case is used as an example (vehicle: LiDAR+Camera; infrastructure: LiDAR); for other vehicle–infrastructure sensor combinations, the per-node setups vary accordingly. The framework integrates features via Hierarchical Attention Fusion (HAF) and balances resolution through Adaptive Spatial Resolution (ASR) across different scenarios.

flexible hetero-modal fusion, we seek to improve perception robustness under varied deployment conditions in V2X systems.

III. METHODOLOGY

A. Heterogeneous V2X Cooperative Perception Framework - HeCoFuse

Real-world V2X deployments rarely feature identical sensor configurations across nodes due to cost constraints, infrastructure limitations, and hardware variations. This heterogeneity creates significant challenges for cooperative perception systems that typically assume sensor consistency. We introduce **HeCoFuse**, a unified framework designed to maintain robust perception performance across diverse sensor configurations, which framework is illustrated in Fig. 1. Our approach addresses nine distinct sensor arrangements:

- **Full sensor configuration:** LC+LC
- **Homogeneous single-type:** L+L, C+C
- **Heterogeneous single-type:** L+C, C+L
- **Mixed configurations:** LC+C, LC+L, C+LC, L+LC

where L denotes LiDAR and C denotes camera, and in the notation ‘X + Y’ the symbol ‘+’ separates the vehicle node (X) from the infrastructure node (Y). The same ‘X+Y’ convention is used throughout this paper.

The core of HeCoFuse is a cooperative learning strategy that enables seamless operation across these heterogeneous configurations. Rather than training separate models for each sensor arrangement, we employ a unified learning approach where the network learns to extract maximum information from available sensors while maintaining consistent feature representations. This is achieved through a modular architecture with feature adaptation components that ensure dimensional compatibility between different sensor inputs. During training, we randomly sample from all nine configurations, enabling the model to learn robust feature extraction

and fusion strategies that generalize across sensor combinations. This approach ensures graceful degradation when high-quality sensors are unavailable. For instance, when a vehicle with only cameras interacts with infrastructure equipped with only LiDAR sensors, the system maintains effective perception by adaptively fusing the complementary information.

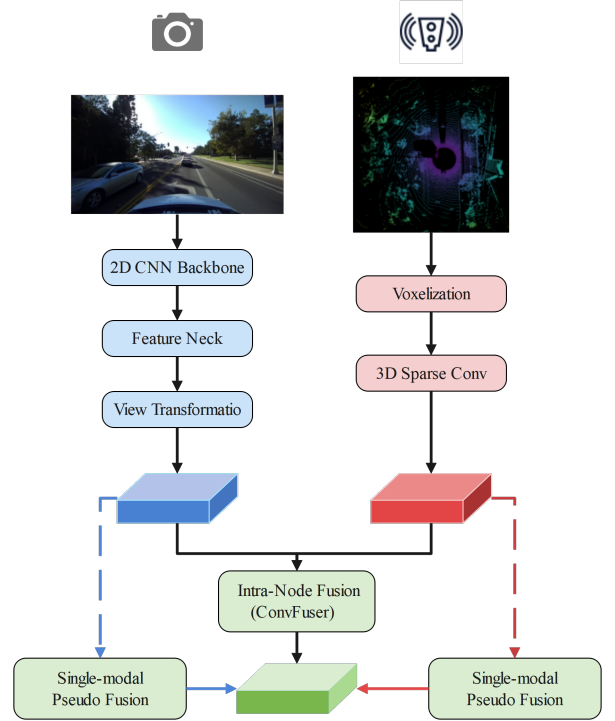


Fig. 2: BEV feature extraction process for heterogeneous sensors. Our framework handles various sensor configurations through specialized encoding paths for LiDAR and camera, along with intra-node fusion for multi-modal configurations and PseudoFusion for single-sensor scenarios.

HeCoFuse adopts a modular approach to feature extraction that adapts dynamically to available sensors at each node, as illustrated in Fig. 2. For a given node $n \in \{vehicle, infrastructure\}$ with sensor set $S_n \subseteq \{camera, lidar\}$, features are extracted through sensor-specific pathways before fusion:

$$F_n = \begin{cases} \text{Fuse}(F_n^{camera}, F_n^{lidar}), & \text{if } |S_n| = 2 \\ F_n^s, & \text{if } |S_n| = 1, s \in S_n \end{cases} \quad (1)$$

1) LiDAR Feature Encoding: For LiDAR processing, we transform unstructured point cloud data into Bird’s Eye View (BEV) features through a multi-stage pipeline. First, our dynamic voxelization module converts raw point clouds into a structured 3D grid, adaptively handling varying point densities across different scenes. The voxelization process assigns points to voxel cells and aggregates information within each cell, significantly reducing computational complexity while preserving essential spatial information.

These voxelized features are then processed by a 3D sparse convolutional backbone that efficiently extracts spatial patterns. By operating only on non-empty voxels, the sparse convolution maintains computational efficiency while capturing detailed geometric structure. Finally, the 3D features are projected onto the ground plane to create a dense bird’s-eye-view representation that serves as input to subsequent fusion stages. This approach effectively leverages the precise spatial information provided by LiDAR while remaining computationally tractable through sparse operations.

2) Camera Feature Encoding: The camera pathway transforms perspective images into BEV space through a sequential process designed to maintain rich semantic information while establishing geometric consistency. Starting with multi-view images, we apply a 2D convolutional backbone to extract features that capture appearance, texture, and object information. These initial features are enhanced through a feature neck network (e.g., FPN) that creates multi-scale representations, providing both fine-grained details and high-level semantic understanding.

The critical step in camera BEV encoding is the view transformation module, which converts perspective view features to BEV space. This involves first estimating depth distributions for each image pixel and then projecting features along these distributions into 3D space using camera-to-LiDAR transformation matrices and camera intrinsic parameters. Finally, the height dimension is collapsed to form the BEV representation. This approach addresses the inherent challenges in transforming 2D perspective views to a 3D coordinate system, enabling effective fusion with LiDAR features despite their fundamentally different sensing modalities.

3) BEV Fusion: When a node contains both LiDAR and camera sensors (e.g., in LC+LC, LC+L, LC+C configurations), we apply a BEV Fusion module to combine complementary information from both modalities:

$$F_n = \text{BEVFusion}(F_n^{camera}, F_n^{lidar}) \quad (2)$$

Our BEVFusion concatenates the features along the channel dimension and applies a convolutional layer with batch normalization and activation:

$$\text{BEVFusion}(F_1, F_2) = \phi_{act}(\phi_{norm}(\phi_{conv}([F_1, F_2]))) \quad (3)$$

This intra-node fusion leverages the complementary strengths of each modality—LiDAR provides accurate spatial information while cameras offer rich semantic features. The fused representation enhances detection performance by combining the strengths of both sensors.

4) Partial Sensor Configuration (LiDAR-only / Camera-only): A key challenge in heterogeneous V2X systems is handling nodes with single sensor types. For these configurations, we employ PseudoFusion to ensure feature compatibility across different nodes.

For L-only or C-only configurations, PseudoFusion employs sensor-specific feature adapters to transform features to a consistent dimensionality:

$$F_n^{adapted} = \phi_{adapter}^s(F_n^s), \quad s \in \{lidar, camera\} \quad (4)$$

where $\phi_{adapter}^s$ is a sensor-specific 1×1 convolutional layer followed by normalization and activation. This ensures that regardless of the input sensor, the output features maintain consistent channel dimensions:

$$F_n^{adapted} \in \mathbb{R}^{B \times C_{out} \times H \times W} \quad (5)$$

To handle the absence of a specific sensor, PseudoFusion checks sensor availability and creates appropriate feature representations:

$$F_n = \begin{cases} F_n^{adapted}, & \text{sensor } s \text{ available} \\ \phi_{fallback}(B, C_{out}, H, W), & \text{no sensors available} \end{cases} \quad (6)$$

where $\phi_{fallback}$ creates a baseline feature representation for extreme cases where sensor data is temporarily unavailable. This approach ensures that HeCoFuse maintains robust performance across all nine possible sensor configurations by guaranteeing feature compatibility regardless of sensor availability.

B. Joint Attention-based Fusion with Adaptive Resolution

The core challenge in heterogeneous V2X fusion lies in effectively combining features from nodes with different sensor configurations while maintaining computational efficiency. As illustrated in Fig. 3, we address this through a two-pronged approach: hierarchical attention fusion and adaptive spatial resolution adjustment.

1) Hierarchical Attention Fusion (HAF): Given BEV features $F_{vehicle}$ and F_{infra} from potentially asymmetric sensor sources, our channel attention mechanism first assesses which feature channels provide the most reliable information:

$$\alpha = \sigma(W_{channel}) \in \mathbb{R}^{1 \times C \times 1 \times 1} \quad (7)$$

For heterogeneous configurations, this enables the network to adaptively weight channels from different sensor modalities. For instance, in an L+C configuration (vehicle with LiDAR, infrastructure with camera), α will preferentially weight LiDAR-dominated channels for distance estimation while favoring camera-dominated channels for appearance features:

$$F_{channel} = \alpha \odot F_{vehicle} + (1 - \alpha) \odot F_{infra} \quad (8)$$

As shown in the central portion of Fig. 3, our spatial attention further refines this fusion by addressing modality-specific strengths in different spatial regions. For nodes with varying sensor configurations, spatial attention maps A_n highlight regions where each node's particular sensors provide the most reliable information:

$$A_n = \sigma(\phi_{conv2}(\phi_{conv1}(F_n))) \in \mathbb{R}^{B \times 1 \times H \times W} \quad (9)$$

In heterogeneous settings, these attention maps become particularly important—when a vehicle with only LiDAR interacts with infrastructure with only cameras, spatial attention emphasizes LiDAR's strengths in accurate distance measurements for distant objects while prioritizing camera-based detections in visually complex regions:

$$F_{fused} = (F_{vehicle} \odot \alpha) \odot A_{vehicle} + (F_{infra} \odot (1 - \alpha)) \odot A_{infra} \quad (10)$$

where $\odot$ denotes the Hadamard (element-wise) product, allowing for fine-grained weighting of each feature element independently.

2) **Adaptive Spatial Resolution (ASR)**: Different sensor configurations produce BEV features with varying characteristics, requiring different computational resources. As depicted in the upper and lower portions of Fig. 3, our approach incorporates adaptive resolution adjustment that works in concert with attention fusion.

For heterogeneous configurations, we first apply configuration-aware downsampling:

$$\begin{aligned} F_{n,down} &= \phi_{downsample}(F_n, s_n) \\ &= \text{AdaptiveAvgPool2d}(F_n, H/s_n, W/s_n), \end{aligned} \quad (11)$$

where s_n is determined based on the specific sensor configuration at node n :

$$s_n = \begin{cases} s_{high}, & \text{if node } n \text{ has camera-only} \\ s_{low}, & \text{if node } n \text{ has LiDAR-only} \\ s_{medium}, & \text{if node } n \text{ has both sensors} \end{cases} \quad (12)$$

Hierarchical attention fusion is then applied to these downsampled features. After fusion, we apply configuration-aware upsampling that prioritizes restoration of detailed information from the most precise sensor configuration:

$$F_{final} = \phi_{upsample}(F_{fused}, \min(s_{vehicle}, s_{infra})) \quad (13)$$

This integrated approach, combining hierarchical attention fusion with adaptive resolution adjustment, enables HeCo-Fuse to effectively bridge the gap between different sensor configurations while reducing computational overhead by 30-45% in heterogeneous configurations.

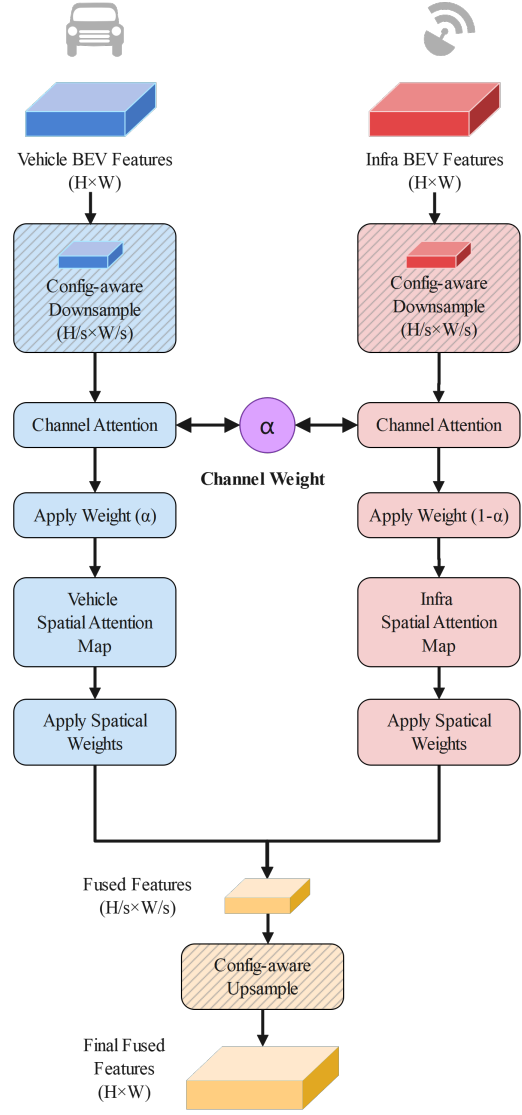


Fig. 3: Architecture of inter-node feature fusion combining Hierarchical Attention Fusion (HAF) and Adaptive Spatial Resolution (ASR) mechanisms. The shaded regions denote the ASR modules.

IV. EXPERIMENTAL EVALUATIONS

A. Dataset Selection

Both simulated and real-world datasets offer distinct advantages for cooperative perception research. Simulated datasets like V2XSet [3] and OPV2V [13] provide controlled environments, while real-world datasets such as DAIR-V2X [36], V2X-Seq [37], and V2XReal [38] capture authentic traffic dynamics and sensor noise. For our heterogeneous cooperative perception framework, we require a dataset meeting several criteria: real-world data collection, multi-modal sensing with both LiDAR and camera streams at each node, consistent sensor types across nodes, and precise calibration between vehicle and infrastructure. The dataset should also feature diverse traffic scenarios with sufficient annotations to evaluate detection performance.

TABLE I: Quantitative comparison of HeCoFuse with baseline and challenge methods across different sensor configurations.

Method	Type	Config		Precision ($\uparrow$)	Recall ($\uparrow$)	3D IoU ($\uparrow$)	Pos. RMSE ($\downarrow$)	Rot. RMSE ($\downarrow$)	3D mAP ($\uparrow$)
		Vehicle	Infra.						
CoopDet3D *		LC	LC	42.34	42.86	0.36	0.48	0.03	42.05
BUPTMM †		LC	LC	43.63	44.55	0.38	0.45	0.04	43.35
CV123 †		LC	LC	43.53	44.32	0.38	0.45	0.03	43.25
TEAM12138 †		LC	LC	43.52	44.32	0.38	0.45	0.03	43.25
KaAI †		LC	LC	43.51	44.35	0.38	0.45	0.03	43.22
HeCoFuse	Full sensor configuration	LC	LC	43.52	44.34	0.38	0.45	0.03	43.22
HeCoFuse	Homogeneous single-type	L	L	42.50	43.52	0.33	0.48	0.06	42.10
HeCoFuse		C	C	22.30	17.33	0.18	0.65	0.15	21.74
HeCoFuse	Heterogeneous single-type	L	C	30.52	26.64	0.21	0.65	0.13	30.04
HeCoFuse		C	L	33.22	24.72	0.21	0.51	0.11	32.68
HeCoFuse	Mixed configurations:	LC	C	31.45	27.37	0.20	0.72	0.10	30.85
HeCoFuse		LC	L	39.64	38.53	0.34	0.52	0.05	39.17
HeCoFuse		C	LC	35.27	33.08	0.26	0.57	0.11	34.76
HeCoFuse		L	LC	43.73	42.91	0.33	0.50	0.04	43.38

The **bolded values** indicate the best-performing model for each metric.

* The results reported in [7] are obtained using pre-trained Camera and LiDAR models, while in our reproduction, the training is performed from scratch.

† These methods represent the other top-5 performing algorithms from the CVPR DriveX Workshop TUMTraF-V2X Challenge [35], excluding HeCoFuse.

We selected the TUMTraF-V2X dataset [7] for our experiments due to its advantages for heterogeneous cooperative perception research. Collected at an urban intersection in Munich, it includes synchronized data from vehicle and infrastructure perspectives with 2,000 camera images, 5,000 LiDAR point clouds, and 29,000 annotated 3D bounding boxes across 8 object categories. Both nodes feature mechanical LiDAR sensors and cameras with similar specifications, enabling simulation of various heterogeneous configurations while maintaining a consistent baseline for evaluation.

B. Quantitative Analysis

We trained our HeCoFuse framework from scratch on the TUMTraF-V2X mini dataset (half of the complete dataset) for 20 epochs using a single GeForce RTX 4090 GPU. For both vehicle and infrastructure nodes, we employed YOLOv8 [39] as the backbone encoder for camera data processing, while PointPillars [40] was utilized as the encoder for LiDAR point cloud processing. To comprehensively evaluate our method, we adopted multiple metrics including Precision, Recall, 3D IoU, Position RMSE, Rotation RMSE, and 3D mAP across various object categories. Table I presents the performance comparison against the official TUMTraF-V2X baseline method, CoopDet3D [7]. These metrics evaluate different aspects of detection performance: Precision and Recall measure detection accuracy, 3D IoU quantifies the spatial overlap between predicted and ground truth bounding boxes, Position and Rotation RMSE assess localization accuracy, while 3D mAP provides a comprehensive measure of detection performance across varying confidence thresholds.

Our HeCoFuse method achieved first place in the CVPR 2025 DriveX Workshop TUMTraF-V2X Challenge [35] based on the 3D mAP metric with the L+LC configuration (vehicle with LiDAR only, infrastructure with both LiDAR and camera), reaching 43.38% 3D mAP. Across all six evaluation metrics, HeCoFuse achieves the best or tied-for-best performance on five metrics, demonstrating the superiority of our approach. Under the full sensor configuration (LC+LC),

HeCoFuse achieves 43.22% 3D mAP, representing a 1.17% improvement over the baseline's 42.05%. Notably, the L+LC configuration delivers the highest overall performance with 43.38% 3D mAP and superior precision (43.73%), highlighting the effectiveness of our hierarchical attention fusion mechanism in heterogeneous settings.

The results reveal several important insights about heterogeneous sensor fusion. First, LiDAR sensors contribute significantly more to detection performance than cameras in our framework. For instance, comparing L+L (42.10% mAP) with C+C (21.74% mAP) configurations shows a substantial 20.36% performance gap, indicating LiDAR's superior spatial accuracy for 3D object detection. Second, infrastructure sensors play a pivotal role in overall system performance due to their wider field of view and elevated mounting position. This is evident when comparing L+LC (43.37% mAP) with LC+L (39.17% mAP), where the former achieves 4.20% higher mAP despite using fewer total sensors. Interestingly, some metrics for the L+LC configuration even surpass those of the fully-equipped LC+LC case; this may be because the roadside camera already provides sufficiently rich perspectives, and the vehicle's additional camera, limited by its mounting viewpoint, adds little extra benefit.

These findings validate our adaptive fusion approach's capability to dynamically leverage the most reliable information sources across heterogeneous configurations, maintaining robust performance even with sensor asymmetry. This represents a critical capability for real-world V2X deployments where sensor configurations often vary widely.

C. Qualitative Analysis

To provide a comprehensive understanding of HeCoFuse's performance in real-world scenarios, we present qualitative visualization results using the full sensor configuration (LC+LC) on the TUMTraF-V2X test set, as shown in Fig. 4. The visualization demonstrates our system's detection capabilities from multiple perspectives: the BEV representation of fused LiDAR point clouds, the vehicle's camera view, and

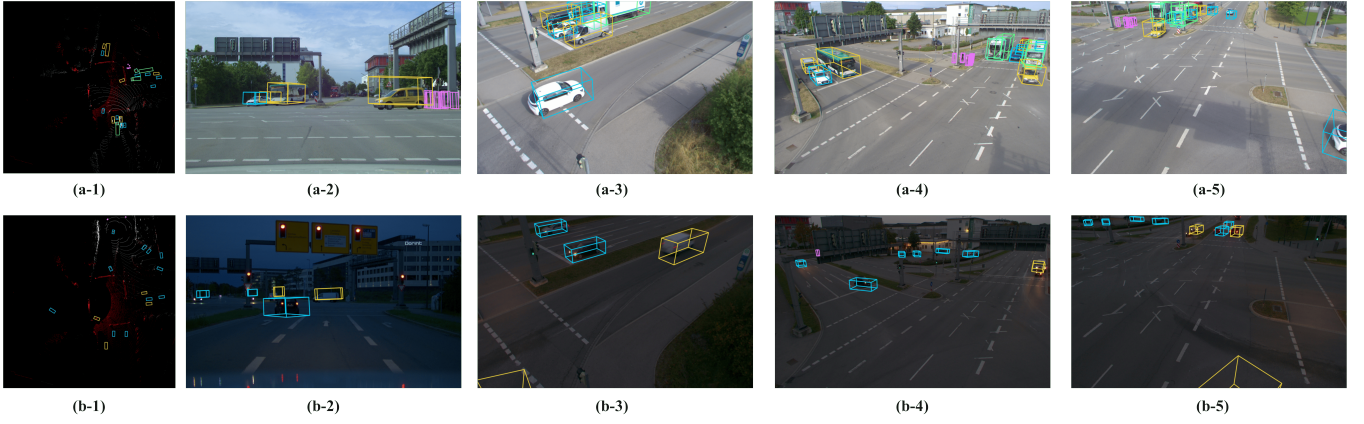


Fig. 4: Qualitative visualization of HeCoFuse detection results in the full sensor configuration (LC+LC) on the TUMTraF-V2X test set. The figure illustrates detection results in both daytime (first row) and nighttime (second row) conditions. (a-1, b-1) BEV perspective showing fused LiDAR point clouds with detection results from both vehicle and infrastructure; (a-2, b-2) vehicle-mounted camera view with corresponding 2D projections of 3D detections; (a-3, a-4, a-5, b-3, b-4, b-5) three different infrastructure camera views with projected detection results.

three different infrastructure camera viewpoints. We specifically selected examples from both daytime and nighttime conditions to illustrate the robustness of our approach across varying lighting conditions.

Our framework successfully detects multiple object categories including passenger vehicles, buses, trains, and pedestrians across different viewing angles and environmental conditions. Notably, the detection performance remains consistent in challenging nighttime scenarios (b-1 to b-5), where camera-only methods typically struggle due to poor lighting. This robustness can be attributed to our hierarchical attention fusion mechanism, which dynamically adjusts the contribution of each sensor modality based on its reliability in different environmental conditions. For instance, in nighttime scenarios, the fusion mechanism automatically places greater emphasis on LiDAR features, which are less affected by lighting variations, while still leveraging the complementary semantic information from cameras. The multi-perspective visualization further demonstrates how HeCoFuse effectively integrates information from spatially distributed sensors, addressing occlusion issues and extending the perceptual range beyond what would be possible with a single sensor node. These qualitative results reinforce our quantitative findings and highlight the practical advantages of heterogeneous sensor fusion for robust cooperative perception in complex urban environments.

V. CONCLUSION

In this paper, we introduce HeCoFuse, a unified cross-modal cooperative perception framework specifically designed to address the challenges of heterogeneous sensor configurations in vehicle-infrastructure cooperative perception systems. Our approach bridges the gap between idealized research assumptions and real-world deployments where vehicles and infrastructure possess varying sensor types and capabilities. HeCoFuse incorporates several key innovations:

a hierarchical attention-based fusion mechanism that dynamically integrates features based on their quality across different modalities and nodes, an adaptive spatial resolution adjustment module that balances computational efficiency with information preservation, and a cooperative learning strategy that enables consistent performance across nine different sensor configurations. Through extensive experiments on the real-world TUMTraF-V2X dataset, we demonstrate that HeCoFuse achieves state-of-the-art performance in fully equipped settings while maintaining robust accuracy in heterogeneous scenarios, particularly when infrastructure nodes are equipped with multiple sensor modalities. Notably, our framework shows exceptional resilience in asymmetric configurations common in practical deployments, where some nodes may only have partial sensing capabilities.

Several promising directions remain for future research. First, extending the validation to scenarios with more than two nodes would further demonstrate the scalability of our approach in complex urban environments with multiple vehicles and infrastructure units. Second, challenging weather conditions such as rain, fog and snow represent critical test cases for heterogeneous fusion approaches, as different sensor modalities exhibit complementary strengths and weaknesses under adverse conditions. Such conditions place greater demands on cooperative perception systems and would further validate the adaptive fusion capabilities of our framework. Finally, expanding the camera configurations to incorporate additional viewpoints on the vehicle side could help address the performance gap between LiDAR and camera-based detection, potentially enabling more balanced multi-modal fusion even when LiDAR sensors are unavailable. This would be particularly valuable for low-cost cooperative perception systems where camera sensors might be the primary sensing modality available on some nodes.

REFERENCES

- [1] G. Yu, H. Li, Y. Wang, P. Chen, and B. Zhou, "A review on cooperative perception and control supported infrastructure-vehicle system," *Green Energy and Intelligent Transportation*, vol. 1, no. 3, p. 100023, 2022.
- [2] S.-W. Kim, B. Qin, Z. J. Chong, X. Shen, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 663–680, 2014.
- [3] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 107–124.
- [4] M. Karvat and S. Givigi, "Adver-city: Open-source multi-modal dataset for collaborative perception under adverse weather conditions," *arXiv preprint arXiv:2410.06380*, 2024.
- [5] H. Xiang, R. Xu, and J. Ma, "Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 284–295.
- [6] C. Wei, Z. Qin, Z. Zhang, G. Wu, and M. J. Barth, "Integrating multi-modal sensors: A review of fusion techniques for intelligent vehicles," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025.
- [7] W. Zimmer, G. A. Wardana, S. Sriharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 668–22 677.
- [8] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2x cooperative perception for autonomous driving: Recent advances and challenges," *arXiv preprint arXiv:2310.03525*, 2023.
- [9] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets, and challenges," *IEEE Intelligent Transportation Systems Magazine*, 2023.
- [10] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524.
- [11] S. Liu, Z. Ding, J. Fu, H. Li, S. Chen, S. Zhang, and X. Zhou, "V2x-pc: Vehicle-to-everything collaborative perception via point cluster," *arXiv preprint arXiv:2403.16635*, 2024.
- [12] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.
- [13] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.
- [14] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "V2vformer: Vehicle-to-vehicle cooperative perception with spatial-channel transformer," *IEEE Transactions on Intelligent Vehicles*, 2024.
- [15] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.
- [16] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 1743–1749.
- [17] Y. Zhang, B. Chen, J. Qin, F. Hu, and J. Hao, "Coopercept: Cooperative perception for 3d object detection of autonomous vehicles," *Drones*, vol. 8, no. 6, p. 228, 2024.
- [18] Z. Wang, S. Fan, X. Huo, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, "Vimi: Vehicle-infrastructure multi-view intermediate fusion for camera-based 3d object detection," *arXiv preprint arXiv:2303.10975*, 2023.
- [19] A. N. Ahmed, S. Mercelis, and A. Anwar, "Collabgat: Collaborative perception using graph attention network," *IEEE Access*, 2024.
- [20] X. Zhao, K. Mu, F. Hui, and C. Prehofer, "A cooperative vehicle-infrastructure based urban driving environment perception method using a ds theory-based credibility map," *Optik*, vol. 138, pp. 407–415, 2017.
- [21] A. P. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.
- [22] A. P. Dempster, "Upper and lower probabilities induced by a multi-valued mapping," in *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008, pp. 57–72.
- [23] S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "Vips: Real-time perception fusion for infrastructure-assisted autonomous driving," in *Proceedings of the 28th annual international conference on mobile computing and networking*, 2022, pp. 133–146.
- [24] A. B. Khalifa, I. Alouani, M. A. Mahjoub, and A. Rivenq, "A novel multi-view pedestrian detection database for collaborative intelligent transportation systems," *Future generation computer systems*, vol. 113, pp. 506–527, 2020.
- [25] G. Pereira, P. M. d'Orey, and A. Aguiar, "Poster: Cooperative perception platform for intelligent transportation systems," in *2020 IEEE Vehicular Networking Conference (VNC)*. IEEE, 2020, pp. 1–2.
- [26] Z. Zheng, X. Xia, L. Gao, H. Xiang, and J. Ma, "Cooperfuse: A real-time cooperative perception fusion framework," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 533–538.
- [27] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.
- [28] Q. Xie, X. Zhou, T. Qiu, Q. Zhang, and W. Qu, "Soft actor-critic-based multilevel cooperative perception for connected autonomous vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 370–21 381, 2022.
- [29] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2020.
- [30] P. Liu, Z. Wang, G. Yu, B. Zhou, and P. Chen, "Region-based hybrid collaborative perception for connected autonomous vehicles," *IEEE Transactions on Vehicular Technology*, 2023.
- [31] D. Qu, Q. Chen, Y. Zhu, Y. Zhu, S. S. Avedisov, S. Fu, and Q. Yang, "Head: A bandwidth-efficient cooperative perception approach for heterogeneous connected and autonomous vehicles," *arXiv preprint arXiv:2408.15428*, 2024.
- [32] T. Luo, L. Chen, T. Luan, Y. Li, and Y. Li, "A vehicle-to-infrastructure beyond visual range cooperative perception method based on heterogeneous sensors," *Energies*, vol. 15, no. 21, p. 7956, 2022.
- [33] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [34] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.
- [35] "Cvpr 2025 drivex workshop: Tumtraf-v2x cooperative perception challenge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2025, cooperative Perception Challenge Track.
- [36] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.
- [37] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.
- [38] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng *et al.*, "V2x-real: A large-scale dataset for vehicle-to-everything cooperative perception," *arXiv preprint arXiv:2403.16034*, 2024.
- [39] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [40] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.