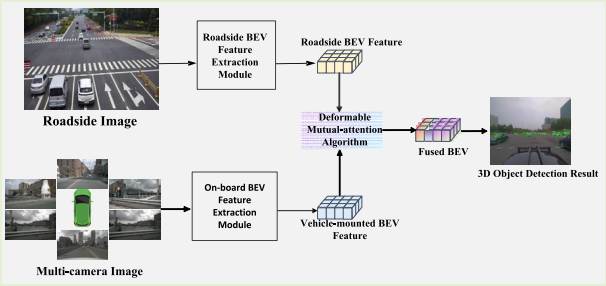
Multiview BEV Fusion From Vehicle-on-Board and Roadside Cameras for 3-D Object Detection

Tianxuan Fu, Guoqiang Mao, *Fellow, IEEE*, Xiaojiang Ren, *Member, IEEE*, Keyin Wang, *Graduate Student Member, IEEE*, Wei Xiang, *Senior Member, IEEE*, and Xiaohu Ge, *Senior Member, IEEE*

***Abstract*—3-D object detection, a sought-after feature for connected and autonomous vehicles (CAVs), offers precise localization and classification of obstacles by estimating their spatial attributes from 2-D images. This approach addresses the limitation of traditional 2-D detection, which lacks spatial information. However, the task becomes challenging in complex environments with occlusions and restricted line-of-sight views, such as intersections and winding roads. Fortunately, roadside infrastructure equipped with traffic cameras, commonly deployed at these critical locations, offers a wider field of view. It can therefore provide valuable support to CAVs and improve object detection performance. Vehicle-to-infrastructure (V2I) cooperative 3-D detection frameworks leverage vehicle-to-everything (V2X) communication to transmit roadside data to CAVs. By integrating multiview observations from both vehicle-mounted and roadside cameras, these frameworks enable a more accurate and semantically rich understanding of the driving environment. In this article, we propose V2IFormer, a multiview feature fusion framework based on bird's eye view (BEV) representations, designed to enhance 3-D object detection. Specifically, V2IFormer extracts semantic features from 2-D images captured by both vehicle-mounted and roadside cameras and transforms them into BEV features. At the CAV side, a deformable mutual-attention algorithm is employed to dynamically select spatial features and fuse the most informative BEV representations from both sources. This strategy reduces interference from irrelevant or corrupted features and improves fusion efficiency. Finally, we validate the effectiveness of V2IFormer on the DAIR-V2X benchmark. Our framework achieves state-of-the-art 3-D object detection performance, particularly in scenarios with occlusion and limited line-of-sight.**



***Index Terms*—3-D object detection, bird's eye view (BEV) perception, connected and autonomous vehicles (CAVs), vehicle-infrastructure cooperative.**

NOMENCLATURE

I_{road}	Input images captured by roadside static cameras.
I_{vehicle}	Input multiview images captured by on-board cameras.
H	Image height.
W	Image width.
N_v	Number of camera views (for vehicle cameras).
F^{2d}	Multiscale features extracted by the 2-D backbone.
F^{2d}_{vehicle}	Multiscale features extracted by the shared 2-D encoder.
C_F	Number of general feature channels.
H_{pre}	Pixel-wise height distribution predicted by HeightNet.
N_H	Number of discrete height bins.
F_{context}	Context features capturing local semantic information.
C_c	Number of context feature channels.
$F_{\text{3d height}}$	Volumetric representation.
$F_{\text{bev road}}$	Final compressed roadside BEV 2-D feature.
$F_{\text{bev vehicle}}$	Vehicle-mounted BEV 2-D feature representation.
X	Width dimension of the BEV grid.
Y	Length dimension of the BEV grid.
$F_{\text{bev fused}}$	Final fused BEV feature representation.
M	Number of attention heads.
K	Number of sampled keys/points per query.

Received 5 December 2025; revised 28 January 2026; accepted 2 February 2026. Date of publication 10 February 2026; date of current version 2 April 2026. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20446 and in part by Wuxi Key Laboratory of Integrated Vehicle-Road-Cloud Smart Transportation System. The associate editor coordinating the review of this article and approving it for publication was Dr. Wei Wei. *(Corresponding author: Guoqiang Mao.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/JSEN.2026.3661270

Δp Predicted sampling point offset.
A_{mqk} Attention weight. m is the attention head index, q is the query point index, and k is the sampled key index.
Q_p Query BEV feature at position $p = (x, y)$.
W_m Weight for aggregating the multihead attention outputs.
W'_m Projection weight for the input features for each attention head.

I. Introduction

MONOCULAR 3-D object detection has emerged as a core task in computer vision, aiming to infer object position, size, and orientation from 2-D images captured by a single camera. Its low cost and ease of large-scale deployment make it a widely adopted technology in autonomous driving, attracting sustained attention from both academia and industry. Despite its promise, vehicle-mounted monocular cameras suffer from inherent limitations. Their low installation height restricts the detection range and makes them highly vulnerable to occlusions, leading to perception blind spots. According to Waymo [1], occlusion accounts for nearly 25% of major accidents, underscoring the critical safety risks associated with restricted line-of-sight.

Roadside sensors such as cameras, millimeter-wave radars, and LiDARs are typically mounted at elevated positions, providing a wider field of view and enabling the detection of more and smaller objects than vehicle-mounted sensors. Vision-based roadside sensing is increasingly favored over LiDAR due to its lower cost, easier deployment, and seamless integration with existing urban infrastructure, including traffic poles and streetlights [2], [3]. However, roadside cameras usually operate in a monocular configuration. Their extended observation distance reduces depth estimation accuracy, making it difficult to distinguish distant vehicles from nearby road features. To alleviate this, recent works have proposed ground-aware modeling strategies. For example, BEVHeight [4] estimates ground-relative heights to enhance depth perception, while MonoGAE [4] incorporates ground-aware embeddings to improve spatial interpretation. Despite these advances, monocular roadside methods still struggle with distant object detection and occlusions. These challenges hinder their ability to fully exploit complementary information in vehicle-to-infrastructure (V2I) cooperative scenarios and highlight the need for more effective fusion frameworks.

V2I cooperative sensing utilizes 5G/vehicle-to-everything (V2X) technologies [5] to transmit roadside high-viewpoint sensor data to connected and autonomous vehicles (CAVs) in real time, thereby integrating multiview sensing data to enhance environmental perception accuracy and robustness. Existing V2I fusion strategies can be categorized into early fusion, intermediate fusion, and late fusion based on the types of data being shared, aiming to enhance 3-D object detection for CAVs through perception data exchange with roadside infrastructure [6], [7], [8]. Early fusion, which transmits raw sensor data, preserves all sensor information but demands high bandwidth due to the large volume of data being transmitted. In contrast, late fusion shares high-level processed outputs, such as bounding boxes, thereby reducing bandwidth demands at the cost of accuracy due to loss of detailed information. Therefore, a key challenge in V2I cooperative systems is determining what data to transmit under restricted communication resources. Intermediate fusion uses processed intermediate features, such as bird's eye view (BEV) representations derived from roadside sensors, to achieve a balance between efficiency and performance [4].

The intermediate fusion strategy integrates vehicle-mounted and roadside sensor data into BEV representations. BEV representations are widely adopted in camera-based autonomous driving systems because they capture comprehensive spatial context and provide a unified global feature space. The main challenge lies in effectively converting perspective-view features into BEV. Recent BEV-based models, such as BEVDepth [9] and BEVFormer [10], have significantly advanced autonomous driving by offering unified 3-D spatial representations for tasks including object detection [11], segmentation [12], and mapping [13], [14]. These representations also facilitate V2I cooperative systems [6], [8]. Despite these advances, current V2I perception methods face critical limitations. Frameworks such as V2X-ViT [7] and VIMI [8] apply attention or concatenation for V2I collaboration. However, static fusion strategies amplify spatial interference, as global attention uniformly processes all regions while concatenation fails to adapt to cross-view feature fusion. Consequently, efficiently extracting and fusing complementary features from diverse sources remains a key challenge in V2I collaborative perception.

Motivated by the aforementioned observations, this article aims to address two key challenges: 1) determining what data to transmit under limited communication resources and 2) designing an effective data fusion method for cooperative detection. Aiming at the problems of V2I cooperative 3-D object detection, we propose V2IFormer, a framework that leverages dual-side BEV fusion to enhance perception in complex traffic scenarios. The first challenge, determining what data to transmit under constrained bandwidth, is addressed by adopting BEV features as an efficient intermediate representation. These features balance essential information retention with communication efficiency. This is distinctly different from transmitting raw sensor data, which demands high bandwidth, or transmitting processed detection results, which may lose critical details [15], [16]. The second challenge, effective fusion of multisource data, is addressed through dual-side BEV representations combined with a deformable mutual-attention-based fusion mechanism. Specifically, for roadside BEV representations, V2IFormer's HeightNet adopts a linearly increasing discretization (LID) strategy to predict adaptive height distributions, effectively mitigating long-range depth attenuation and enhancing ground-relative height modeling without relying on dense depth cues. For vehicle-side BEV representations, our uncertainty-aware height modeling (UAHM) uses Gaussian-based distributions to reduce depth ambiguity.

The main contributions of this article are summarized as follows.

1) We propose V2IFormer, a BEV-based V2I collaborative perception framework that enables efficient feature

fusion between vehicles and infrastructure. Dual-side BEV representations are introduced to balance communication efficiency with perception accuracy. On the infrastructure side, a height distribution prediction module enhances long-range perception by modeling object height distributions. On the vehicle side, a Gaussian-based height modeling module improves geometric reconstruction and alleviates depth ambiguity.

2) A deformable mutual-attention module is designed to effectively mitigate spatial interference arising from irrelevant data during feature fusion. The module dynamically queries complementary features from both vehicle-side and infrastructure-side BEV representations. It adaptively selects relevant information through learned attention weights.
3) The proposed V2IFormer is comprehensively evaluated on the DAIR-V2X dataset, a large-scale benchmark for real-world V2I perception, and further validated on the V2XSet and nuScenes datasets. Experimental results demonstrate that it achieves state-of-the-art 3-D object detection performance across all these datasets.

The rest of the article is organized as follows. Section II reviews related work on 3-D object detection, collaborative perception, and BEV scene understanding, highlighting key advances and limitations. Section III presents the proposed framework, including the dual-side BEV representations and the deformable mutual-attention module. Section IV reports experimental results on the DAIR-V2X and V2XSet dataset, validating the accuracy and robustness of the proposed method. Finally, Section V concludes the article and outlines directions for future research.

II. Related Work

A. Camera-Based 3-D Object Detection

Current 3-D object detection approaches are generally categorized into image-based and LiDAR-based methods. LiDAR achieves high detection accuracy but suffers from high cost, bulky hardware, and inherent limitations, such as poor performance on small and distant objects. In contrast, camera-based detection estimates 3-D bounding boxes from monocular images and offers distinct advantages. Cameras are cost-effective, widely available, and easily deployed in both vehicles and infrastructure systems [17], [18]. For example, Tesla's full self-driving (FSD) system relies solely on cameras to deliver advanced driver assistance at a lower cost than LiDAR-equipped systems [19]. Moreover, cameras capture rich semantic information—such as color, texture, and context—crucial for robust classification and scene understanding [10]. Advances in deep learning, particularly in monocular depth estimation, have further improved their accuracy [20]. Therefore, camera-based detection offers a favorable balance between accuracy, cost, and scalability, making it promising for autonomous driving and infrastructure-assisted perception.

Depending on the deployment scenario, camera-based 3-D object detection methods can be broadly categorized into two types: vehicle-mounted approaches, which employ onboard cameras, and infrastructure-based approaches, which use cameras installed on fixed roadside infrastructure.

1) Vehicle-Mounted Methods: Vehicle-mounted 3-D object detection methods can be broadly divided into two categories. The first extends 2-D detectors with minor modifications to jointly predict 2-D image-based and 3-D geometric attributes. Representative works include FCOS3D [21], which augments a 2-D framework for joint 2-D–3-D estimation; DETR3D [22], which employs 3-D-aware queries for direct regression; and PETR-series models [23], [24], [25], which progressively incorporate geometric cues and temporal information for improved accuracy. The second category dispenses with 2-D intermediate outputs and directly constructs structured 3-D representations. OFT [26] maps 2-D features into 3-D voxels; LSS [27] lifts features into BEV space via depth estimation; and later methods such as ImVoxelNet [28], BEVDepth [9], and CrossDTR [29] enhance volumetric or BEV reasoning with depth supervision and transformer-based embeddings.

2) Infrastructure-Based Methods: Roadside sensors, compared with vehicle-mounted sensors, offer wider environmental coverage and continuous monitoring from elevated viewpoints. However, infrastructure-based 3-D object detection faces two critical challenges: 1) distant object detection, where detection of smaller objects (e.g., pedestrians or vehicles at long range) suffers from low resolution and sparse features, and 2) occlusion, where dynamic obstructions compromise visibility. Recent advancements, such as the DAIR-V2X [6] and Rope3D [30] datasets, provide real-world benchmarks to advance roadside perception. Despite these efforts, the accuracy of the state-of-the-art detection models remains limited when applied to roadside scenarios. To address this limitation, BEVHeight [4] introduces an approach that predicts height distributions instead of relying on conventional depth estimation, significantly improving performance. In contrast, CBR [31] transforms image features into BEV space using multilayer perceptrons (MLPs), avoiding the need for extrinsic calibration but sacrificing geometric precision.

B. Cooperative Perception

V2I cooperative perception enhances situational awareness in connected environments by enabling communication among agents. On this basis, it mitigates occlusions and extends perceptual range [32]. Early fusion collaboration shares raw sensor data, such as uncompressed camera images. This approach preserves all information but requires high bandwidth, making it less suitable for communication-constrained V2X scenarios. In contrast, late fusion collaboration reduces bandwidth by transmitting only final detection results, such as 3-D bounding boxes. While efficient, this method often performs worse than single-agent approaches (where vehicles rely solely on onboard sensors) in environments with high sensor noise. For instance, Cooper [33] applies early fusion by broadcasting raw LiDAR data, which incurs significant communication costs. Conversely, late fusion strategies [6], [34] aggregate predictions from CAVs to reduce transmission overhead. However, they discard contextual information

and heavily depend on each CAV's detection accuracy and postprocessing [34].

Intermediate fusion offers a balance between accuracy and efficiency by transmitting compressed feature representations, such as BEV features. Roadside units encode raw sensor inputs into compact features and broadcast them to CAVs, which then aggregate multiview features for collaborative detection. Existing methods range from simple aggregation (e.g., summation or max pooling [35], [36]) to advanced attention-based fusion [3], [7], [37] and adaptive schemes [2].

Several recent BEV-based V2X cooperative perception methods illustrate the strengths of the BEV paradigm but also reveal important limitations that remain unresolved. For example, V2X-ViT [7] focuses on LiDAR-based BEV voxel encoding and uses a transformer to represent features and perform multiagent fusion. This approach works well for LiDAR point cloud data, but it is limited by the homogeneous treatment of features and struggles to effectively represent features for small objects and long-range occlusions. VIMI [8] employs an intermediate fusion approach by integrating vehicle and infrastructure features prior to BEV projection. This framework explicitly incorporates camera parameters as priors within its multiscale cross attention (MCA) module and designs a camera-aware channel masking (CCM) module that uses these parameters to augment the fused features and correct spatial misalignment. However, this parameter-dependent approach may lack robustness when dealing with significant calibration errors, which are common in real-world deployments. Similarly, BEV fusion [38] operates within the BEV-based fusion paradigm to mitigate information loss by integrating features in a unified BEV space. This approach effectively addresses the limitations of geometrically or semantically lossy projections seen in prior fusion paradigms, achieving superior performance through an optimized view transformation. However, its efficacy remains dependent on the accuracy of image-based depth estimation. This inherent dependency presents a fundamental limitation, leading to performance degradation in challenging scenarios. Within a BEV-based intermediate fusion framework, the practical collaborative perception framework in [15] focuses on two key challenges in V2X collaboration: which information should be exchanged over the V2X network and how the exchanged information should be fused. The method addresses these challenges by propagating compact object-level information using multiframe scene flow and then fusing the propagated objects with ego-agent features, thereby reducing communication load while maintaining detection performance. Our proposed V2IFormer directly addresses the key limitations of existing methods. It introduces three major innovations: an asymmetric dual-side BEV representation, tailored fusion strategies for roadside and vehicle-mounted features, and an adaptive deformable mutual-attention mechanism. These innovations overcome critical challenges such as geometric misalignment, depth cue issues, and feature redundancy. As a result, V2IFormer significantly improves the performance of BEV-based V2X methods, offering a more robust and efficient solution.

C. BEV Scene Understanding

BEV perception has gained significant research interest for integrating multisensor data into a unified spatial framework, enabling effective handling of complex traffic scenarios. A key challenge is transforming perspective-view camera images into accurate BEV representations. Existing solutions can be divided into implicit and explicit approaches.

1) Implicit Lifting Methods: Implicit approaches use neural networks, such as transformers and MLPs, to generate BEV features without relying on geometric priors. MLP-based methods, such as the VPN framework [39], convert multicamera images into top-down BEV views using fully connected layers. These methods have proven effective for tasks like segmentation and road layout prediction [40]. Transformer-based approaches instead employ attention mechanisms between BEV queries and image features to establish spatial correspondences [10], [23], [41]. Notably, BEVFormer [10] integrates spatial cross-attention with temporal fusion to improve representation stability. However, these methods often overlook explicit camera intrinsic parameters (e.g., focal length and principal point) and extrinsic parameters (e.g., position and orientation). As a result, performance may degrade under varying camera configurations, since accurate geometric constraints are essential for precise feature alignment.

2) Explicit Lifting Method: Explicit methods for generating BEV features leverage geometric principles and depth information to produce accurate top-down representations. Inverse perspective mapping (IPM) [42] projects images into BEV using intrinsic parameters (e.g., focal length and principal point) and extrinsic parameters (e.g., camera pose in 3-D space). However, IPM is susceptible to errors introduced by sampling and quantization artifacts. To mitigate the ill-posed nature of monocular 3-D reconstruction, later approaches incorporate depth estimation. Pseudo-LiDAR [43] converts image pixels into point clouds using predicted depth, enabling stronger 3-D reasoning in BEV space. More recent methods, such as LSS [27], predict depth distributions and exploit camera parameters to project image features into BEV [4], [9], [44]. These geometry-based methods excel at generating accurate BEV features by explicitly modeling scene structure. Building on these principles, our proposed V2IFormer introduces height discretization and prediction to achieve precise and efficient perspective-to-BEV mapping.

3) Multimodal BEV Fusion: Recent advances in BEV-based perception have demonstrated the potential of radar-camera and LiDAR-camera fusion for improving detection performance, particularly under challenging conditions. Works such as RCBEVDet [45] and RCBEVDet++ [46] explore radar-camera fusion in the BEV space, enhancing robustness in low-visibility and adverse weather conditions. Similarly, methods like PointFusion [47] focus on fusing LiDAR and camera data, but face challenges in spatial misalignment and sensor perspective differences, which impact fusion accuracy. Despite the successes of these early fusion approaches, they often rely on rigid feature alignment, which can struggle with latency-induced misalignment and temporal discrepancies in multiagent systems. In contrast, our V2IFormer model introduces a deformable mutual-attention mechanism

that dynamically aligns features across modalities, addressing the misalignment issues of previous methods. By leveraging bilateral BEV representations that fuse vehicle-side and infrastructure-side sensor data, V2IFormer improves performance in multiagent cooperative settings, where sensor data is asynchronous and subject to varying environmental conditions, such as weather and network delays.

III. Methodology

A. Overview of the Proposed Framework

The proposed V2IFormer framework, depicted in Fig. 1, is designed to enable collaborative 3-D object detection in V2I scenarios. It consists of three tightly integrated components: an infrastructure-side BEV representation branch, a vehicle-mounted BEV representation branch, and a BEV feature fusion module for vehicle-infrastructure integration. These modules are jointly optimized to exploit complementary spatial and contextual information from both infrastructure and vehicle perspectives, therefore enabling more robust and accurate 3-D object detection. Sections III-B–III-D describe each component in detail and highlight their interactions.

1) Infrastructure-Side BEV Representation Branch: The infrastructure-side branch processes images captured by roadside static cameras, denoted as $I_{\text{road}} \in \mathbb{R}^{3\times H\times W}$, where H and W are the image height and width, respectively, and the three channels correspond to RGB values. A 2-D backbone combined with a feature pyramid network (FPN) extracts multiscale image features $F^{2d} \in \mathbb{R}^{C_F\times H/16\times W/16}$, where C_F is the number of feature channels, and the spatial dimensions are downsampled by a factor of 16. To incorporate vertical spatial information, the HeightNet module (see Section III-B1) predicts a pixel-wise height distribution $H^{\text{pre}} \in \mathbb{R}^{N_H\times H/16\times W/16}$, where N_H indicates the number of discrete height bins. The vertical space is divided into N_H intervals between predefined minimum and maximum heights. Each bin represents a specific height range, and the corresponding value in H^{pre} indicates the probability of that height range for each pixel. Simultaneously, context features $F^{\text{context}} \in \mathbb{R}^{C_c\times H/16\times W/16}$ are generated to capture local semantics information, where C_C is the number of context feature channels. These two outputs—height distributions H^{pre} and context features F^{context}—are fused using a learned function [see (3)] to form a volumetric representation $F^{3d}_{\text{height}} \in \mathbb{R}^{C_c\times N_H\times H/16\times W/16}$, effectively lifting the 2-D features into 3-D space by combining semantic information with height distributions. Subsequently, the height-based projection module transforms these volumetric features into wedge-shaped volumes aligned with the ego-coordinate system. Finally, a voxel-wise aggregation is performed along the height dimension to compress the 3-D feature volume into a 2-D BEV representation $F^{\text{bev}}_{\text{road}} \in \mathbb{R}^{X\times Y\times C_F}$. Here, X and Y represent the width and length of the BEV grid, determined by the scene's spatial dimensions and grid resolution.

2) Vehicle-Mounted BEV Representation Branch: Parallel to the infrastructure-side branch, the vehicle-mounted branch processes multiview images captured by on-board vehicle cameras, represented as $I_{\text{vehicle}} \in \mathbb{R}^{N_v\times 3\times H\times W}$, where N_v is the number of camera views, and H and W represent the image height and width, respectively, with three RGB channels. A shared 2-D encoder extracts multiscale features $F^{2d}_{\text{vehicle}} \in \mathbb{R}^{N_v\times C_F\times H/16\times W/16}$, where C_F is the number of feature channels, and the spatial dimensions are downsampled by a factor of 16. These features are then lifted from 2-D image space to 3-D space using a cross-view transformer, which aligns multiview image features with the spatial layout of the BEV grid by leveraging camera intrinsic (e.g., focal length and principal point) and extrinsic (e.g., camera position and orientation) parameters. This process accounts for the geometric relationships across multiple camera perspectives to ensure accurate 3-D projection. Subsequently, a BEV pooling operation aggregates the transformed features within each grid cell of the BEV plane and compresses them along the height dimension to produce the vehicle-mounted BEV representation $F^{\text{bev}}_{\text{vehicle}} \in \mathbb{R}^{X\times Y\times C_F}$, where X and Y denote the width and length of the BEV grid, determined by the scene's spatial dimensions and grid resolution. Fig. 1 illustrates the configuration of on-board vehicle cameras and roadside infrastructure cameras, along with the feature fusion process, to clarify the spatial relationships between the two perspectives and their integration into a unified BEV representation.

3) On-Board Fusion of Vehicle and Infrastructure BEV Features: The BEV fusion module serves as the core of V2I cooperation by integrating $F^{\text{bev}}_{\text{road}}$ and $F^{\text{bev}}_{\text{vehicle}}$ via a deformable mutual-attention mechanism. This module adaptively attends to spatially overlapped sensing regions from both sources, generating a unified and enriched BEV representation $F^{\text{bev}}_{\text{fused}}$. The fused features are then passed through the 3-D object detection head that predicts bounding boxes parameterized by location (x, y, z), dimension (l, w, h), and orientation θ.

B. Constructing Roadside BEV Features

Accurate transformation of 2-D image features into 3-D representations is a critical step for BEV perception in V2I systems [4], [9], [27]. Common monocular methods such as BEVDepth [9] rely on depth prediction, producing pixel-level depth distributions to infer 3-D structure. However, their performance significantly degrades for small or distant objects, which occupy only a few pixels in the image. This limitation leads to unstable or inaccurate depth estimation due to insufficient spatial evidence. To address this issue, recent works such as BEVHeight [4] propose height-based prediction, which estimates object height as an intermediate geometric feature. Compared to direct depth prediction, height estimation offers greater robustness to distance, as height cues remain relatively stable even when the object occupies a small number of pixels only. This strategy is particularly effective for traffic environments, where object categories like vehicles typically exhibit consistent height ranges. By using height prediction as an intermediate representation, it not only maintains geometric consistency in BEV space but also significantly enhances the stability and accuracy of distant object detection. Therefore, inspired by height-based modeling strategies such as BEVHeight [4], we adopt an alternative approach that estimates object height as an intermediate geometric representation. Unlike depth, object height is relatively invariant to distance and less sensitive to projection distortion.

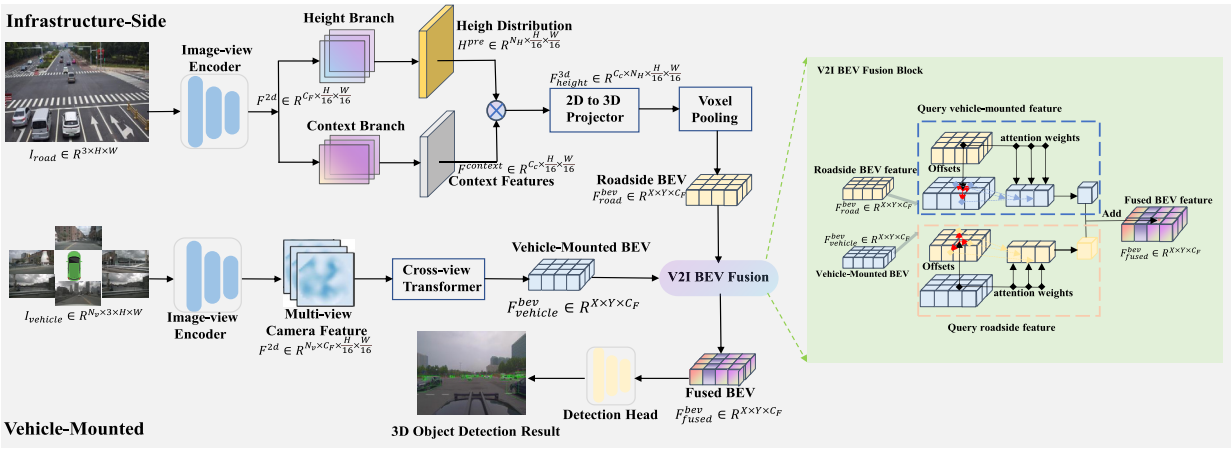


Fig. 1. V2IFormer framework consists of three main components: the infrastructure-side branch, the vehicle-mounted branch, and the BEV feature fusion module. The infrastructure branch generates height-based BEV features from ground-height predictions, while the vehicle-mounted branch uses a spatial encoder to create BEV features from vehicle-mounted images and a temporal encoder to integrate historical features. The fusion module employs deformable mutual-attention to combine both BEV features, producing fused features for the detection head.

1) HeightNet: Based on the insights from the above analysis, we design a dedicated height estimation network, termed HeightNet, to predict pixel-wise height distributions. Pixel-wise height refers to the estimated height of objects in a 3-D scene, assigned to each pixel in a 2-D image, forming a discretized height distribution that captures the spatial variation of heights across the image. To estimate per-pixel object heights from 2-D image features, HeightNet is designed to predict a discrete height distribution for each pixel by integrating contextual and geometric cues. HeightNet consists of two main branches: a context branch for extracting semantic information, and a height estimation branch for predicting discretized height values. Starting from the 2-D image features F^{2d}, we extract context features F^{context} using residual blocks [48] and a channel attention mechanism, followed by a deformable convolution (DCN) [49] to predict per-pixel heights. Simultaneously, we encode the camera's extrinsic parameters E and intrinsic parameters I through two MLPs to generate camera-aware features F_{cam}, enhancing the network's adaptation to varying camera viewpoints.

We formulate height prediction as a classification task by discretizing the height range $N_H = 90$ bins (determined through empirical analysis of the height distribution) using one-hot encoding. Previous studies [50], [51] typically employed uniform discretization (UD) with equal intervals, which are less effective for roadside height estimation due to their inability to adapt to the nonuniform distribution of object heights. To address this limitation, we compare three discretization strategies, each with distinct characteristics: 1) UD, which uses equal bin widths but struggles with sparse sampling at lower heights, leading to poor accuracy for distant and small objects; 2) spacing-increasing discretization (SID), which employs logarithmically increasing bin sizes, where smaller bins are concentrated at lower heights and larger bins at higher heights, making it suitable for long-distance detection but overly sparse for short-range objects; and (3) LID [52], which uses bin widths that increase linearly with

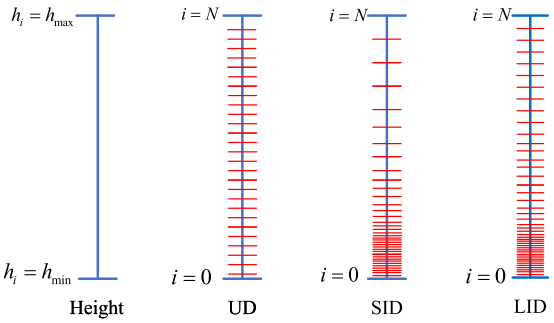


Fig. 2. Height discretization method. The continuous height values h_i are quantized into N discrete intervals within the bounded range $[h_{\min}, h_{\max}]$.

height, ensuring balanced sampling density and consistent relative error across all height ranges for improved accuracy on short-range objects and stability for distant ones. These techniques are illustrated in Fig. 2. Based on our analysis, LID provides a more balanced accuracy for nearby objects and stability for distant ones by linearly increasing bin width sizes, ensuring a consistent relative error across all height ranges. Thus, we adopt LID in this work. Its mathematical formulation is given by

$$h_i = h_{\min} + (h_{\max} - h_{\min}) \cdot \frac{i\,(i+1)}{N_H\,(N_H+1)} \tag{1}$$

where the continuous target height values are divided into N_H bins, and the discretized height can be represented as h_i, covering a predefined interval $[h_{\min}, h_{\max}]$. The height prediction module, illustrated in Fig. 3, is designed to estimate pixel-wise heights. The architecture begins with residual blocks, as proposed in [48], which process the 2-D features F^{2d} conditioned on camera parameters F_{cam} to enhance feature quality. These refined features are then passed through a DCN layer [49] that adaptively samples spatial information

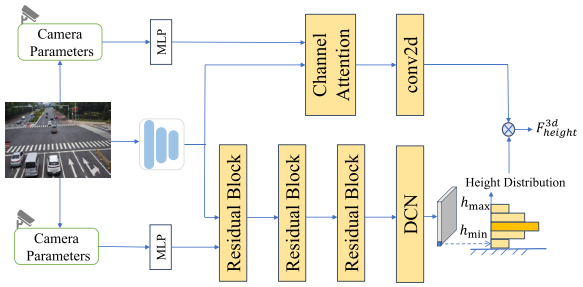


Fig. 3. Framework of the HeightNet. The per-pixel height distribution is directly predicted from the image features. This process employs multiple stacked residual blocks followed by a DCN layer. In addition, the contextual features F^{context} are extracted from the 2-D image features F^{2d} using a channel attention mechanism and a 2-D convolutional block.

to compute precise height values for each pixel. The complete process is formally expressed as

$$H^{\text{pre}} = \psi_h\left(\text{DCN}\left(\text{Res}\left(F^{2d}|F_{\text{cam}}\right)\right)\right) \tag{2}$$

where $\psi_h(\cdot)$ represents the prediction network, DCN$(\cdot)$ denotes the DCN operation, and Res$(\cdot)$ refers to the residual blocks. Here, $H^{\text{pre}} \in \mathbb{R}^{N_H \times H/16 \times W/16}$ denotes the predicted height distribution.

2) Height-Based 2-D-to-3-D Feature Projection: The predicted height distribution H^{pre} [obtained from (2)] enables us to lift the 2-D features into 3-D space. To construct a 3-D volumetric representation from 2-D image features, we project the pixel-wise height distribution H^{pre} into a frustum-shaped 3-D space. First, spatial context features F^{context} are derived from the 2-D image features F^{2d} using channel attention and convolutional operations. These features are then combined with the predicted height H^{pre} via an outer product operation

$$F^{3d}_{\text{height}} = F^{\text{context}} \otimes H^{\text{pre}} \tag{3}$$

where $\otimes$ denotes the outer product, effectively lifting 2-D image features into a 3-D volumetric space by incorporating height information. The resulting feature volume $F^{3d}_{\text{height}} \in \mathbb{R}^{C_c \times N_H \times H/16 \times W/16}$ encodes both spatial and height-aware cues, where C_c is the channel dimension and N_H represents the number of height bins.

To align the 3-D features with the ego vehicle's perspective, we introduce a three-step projection module that transforms the frustum-shaped volume into a wedge-shaped volumetric representation in the ego coordinate system (Fig. 4). This transformation ensures compatibility with BEV feature fusion for autonomous driving applications.

Step 1: Image Pixel to Camera Coordinate

For each image pixel $p_{\text{imag}} = (u, v)$, we first choose the associated point P_{ref} in the reference plane (where depth is set to 1). The projection from image coordinates to the camera coordinate system is computed using the camera intrinsic matrix I, as follows:

$$P^{\text{cam}}_{\text{ref}} = I^{-1}\,[u, v, 1]^T \tag{4}$$

where u and v denoting pixel indices. Here, I is the intrinsic camera matrix that maps 3-D points in the camera coordinate system to 2-D image coordinates.

Step 2: Camera Coordinate to Virtual Coordinate

The reference point $P^{\text{cam}}_{\text{ref}}$ is then transformed into the virtual coordinate using a transformation matrix $T^{\text{virt}}_{\text{cam}}$, which incorporates the camera's extrinsic parameters (rotation and translation relative to the ground)

$$P^{\text{virt}}_{\text{ref}} = T^{\text{virt}}_{\text{cam}} P^{\text{cam}}_{\text{ref}}. \tag{5}$$

Here, $T^{\text{virt}}_{\text{cam}}$ is a rotation matrix that aligns the virtual coordinate system such that the origin is at the camera's optical center, the Y-axis points downward toward the ground, the Z-axis points forward, and the X-axis points to the right. A virtual coordinate system is used to compute the ego point P_{ego}. This step normalizes observations from different camera mounting angles and positions into a canonical virtual space, effectively mitigating the impact of varied roadside camera placements.

Step 3: Virtual to Ego Coordinate

For each height bin h_i, we compute the corresponding 3-D point in the ego vehicle coordinate system using the principle of similar triangles. Specifically, we consider the geometric relationship between the reference point and the target height point as

$$\Delta O_{\text{virt}} M P_{\text{ref}} \sim \Delta O_{\text{virt}} N P_{\text{ego}} \tag{6}$$

where O_{virt} is the origin of the virtual system (camera center), P_{ref} is the reference point at unit depth, and P_{ego} is the projected 3-D point corresponding to height bin h_i. We define H as the known configurable distance from O_{virt} to the ground. Using the property of triangle similarity ($O_{\text{virt}}N / O_{\text{virt}}M = O_{\text{virt}}P_{\text{ref}} / O_{\text{virt}}P_{\text{ego}}$), the 3-D point in ego coordinates is computed as

$$P^{\text{height}}_{\text{ego}} = T^{\text{ego}}_{\text{virt}} \frac{H - h_i}{y_{\text{virt}}} P^{\text{virt}}_{\text{ref}} \tag{7}$$

where $T^{\text{ego}}_{\text{virt}}$ denotes the transformation matrix from the virtual coordinate system to the ego coordinate system. y_{virt} is the Y-coordinate of $P^{\text{virt}}_{\text{ref}}$, and $H - h_i / y_{\text{virt}}$ is the scaling factor derived from the similar triangle principle. This operation assigns each pixel (u, v) and height bin h_i to a unique 3-D position in the ego coordinate system.

This geometric formulation ensures high applicability to diverse infrastructure layouts. The scaling factor $H - h_i / y_{\text{virt}}$ automatically adapts to the actual mounting height H and pitch angle encoded in y_{virt}. Consequently, the framework can accurately map features to a unified BEV space regardless of whether the roadside camera is positioned at different heights or horizontal offsets, provided the extrinsic parameters are known. This intrinsic robustness allows V2IFormer to generalize across various intersection geometries without the need for scene-specific recalibration or retraining.

After this transformation, all height-bin features are projected onto the BEV grid. To reduce the 3-D volume to a compact 2-D BEV feature map, we perform sum-pooling along the height dimension N_H. This preserves the channel information while aggregating height direction features, resulting in a concise 2-D representation that is well-suited for downstream tasks such as object detection and multisensor fusion in autonomous driving.

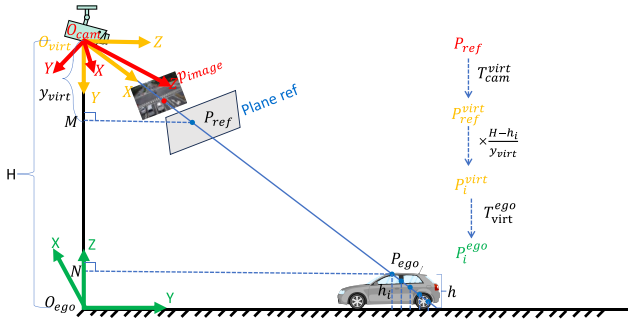


Fig. 4. Height-based 2-D–3-D projector. The $O_{ego}XYZ$ implies the coordinate system, O_{virt} has the same origin as O_{cam} but with Y-axis perpendicular to the ground. A pixel in the image plane, with an estimated height h_i from the predicted distribution, is projected through this 2-D-to-3-D process into a 3-D point P_{ego} in the ego coordinate system.

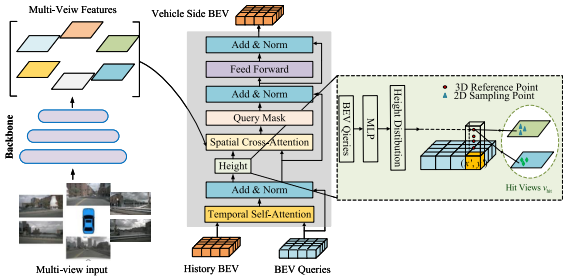


Fig. 5. Main architecture of extracting multiview vehicle-mounted BEV feature.

C. Constructing Vehicle-Side BEV Features

The architecture of the vehicle-side BEV representation is illustrated in Fig. 5. We focus on two key components: 1) BEV space construction with explicit height uncertainty modeling, which ensures accurate projection of sensor data into the BEV space by accounting for vertical dimension uncertainties, and 2) segmentation-based query masks (SQMs), which enable focused feature extraction by dynamically isolating irrelevant regions. The remaining components will be briefly introduced for completeness.

1) BEV Space Construction Through Explicit Height Uncertainty Modeling: The transformation of multicamera image features into BEV features provides a unified environmental representation that significantly enhances perception tasks for autonomous driving systems. This conversion process requires an advanced feature sampling technology. As illustrated in Fig. 5, each grid position in the BEV space is associated with multiple 3-D reference points derived from its spatial coordinates and height information. These 3-D points undergo projection onto the image plane to establish the corresponding 2-D reference points, enabling precise feature extraction. Our approach builds upon BEVFormer's architecture [10], which effectively aggregates multiview image features into the BEV space through its innovative spatial cross-attention mechanism. This framework serves as the foundation for our enhanced BEV representation system.

As shown in Fig. 5, the BEVFormer architecture [10] employs six encoder layers that maintain the standard transformer structure while incorporating three key modifications: 1) BEV queries; 2) spatial cross-attention; and 3) temporal self-attention. The BEV queries serve as grid-shaped learnable embeddings that systematically extract relevant features from multiple camera views via attention mechanisms. The spatial cross-attention module operates on these queries to integrate features across different camera perspectives, while the temporal self-attention mechanism enables the fusion of current observations with historical BEV representations, thereby capturing dynamic scene evolution.

The conventional spatial cross-attention mechanism employs predefined anchor heights (a discrete set of elevation values assigned to each BEV voxel), implicitly encoding height information through attention weights. For each sampled BEV voxel, the attention weights compute a weighted average of the anchor heights. However, the discretization of predefined anchor heights leads to discontinuities, which in turn produce sparse BEV projections and incomplete spatial coverage.

To address the challenge of height ambiguity in real-world scenarios, we propose a UAHM approach. By introducing a continuous height representation and explicitly modeling height uncertainty, our method enables smoother and more reliable BEV feature aggregation. As illustrated in Fig. 5, our approach employs a lightweight MLP to predict a discrete probability distribution P_i over predefined height bins for each BEV grid position. The mean μ and variance σ^2 are then derived analytically from P_i through expectation computation. This design offers two key advantages: First, predicting a full distribution (rather than regressing continuous values) improves stability by avoiding numerical instabilities and ensuring non-negative variance. Second, the distribution provides an interpretable uncertainty model, explicitly capturing the probability spread across heights. By explicitly modeling height uncertainty as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, our approach supports continuous and flexible sampling of 3-D reference points, improving BEV feature aggregation smoothness and reliability. The mean μ and variance σ^2 of the height distribution are computed as follows:

$$\mu = \sum_{i=0}^{N_h-1} P_i h_i \tag{8}$$

$$\sigma^2 = \sum_{i=0}^{N_h-1} P_i \left(h_i - \mu\right)^2. \tag{9}$$

Here, h_i and $P_i(\cdot)$ denote the height value of the ith bin and its corresponding probability, respectively. To sample heights from the predicted distribution, we select N_h values (empirically set to $N_h = 10$ to match urban scene height variations) from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. This sampling strategy ensures robust coverage of the height uncertainty range, enabling continuous and flexible height projections without introducing discontinuities.

The sampled heights are then utilized in our multiview projection pipeline. For each view, a 3-D point (x_r, y_r, h_i) is defined in the ego-vehicle coordinate system, where (x_r, y_r) corresponds to a BEV grid position. These 3-D points are then

projected onto the 2-D image planes of multiple camera views. Specifically, the 2-D coordinates (x_{ij}, y_{ij}) are computed as

$$h_{ij} \cdot \left[x_{ij}, y_{ij}, 1\right]^T = T_j \left[x_r, y_r, h_i, 1\right]^T. \tag{10}$$

Here, (x_r, y_r) is the 2-D point on jth camera view projected from ith 3-D point (x_r, y_r, h_i), $T_j \in \mathbb{R}^{3\times4}$ is the known projection matrix of the jth camera, and h_{ij} represents the height of the ith 3-D point as obtained from the perspective of the jth camera, determined by the preceding projection process. This projection mechanism ensures accurate feature sampling across multicamera views.

2) Segmentation-Based Query Mask: To improve the reliability of feature aggregation in the BEV space by tackling unconstrained height predictions for unannotated queries, we introduce an SQM. We observe that many queries do not correspond to any annotated object. For these queries, simply assigning low height values can lead to inconsistent representations across ground, buildings, trees, and other unannotated structures. Consequently, the predicted heights for such queries remain unconstrained, potentially introducing noise into the BEV feature space. To address this issue and prevent the aggregation of irrelevant background features, we introduce an SQM. Specifically, the SQM generates a binary segmentation mask that identifies whether each query corresponds to a relevant object. Features from queries classified as irrelevant are completely masked out (set to zero), thereby significantly improving the reliability of feature aggregation in the BEV space.

3) Other Modules: Temporal fusion modules [10] in Fig. 5 are designed to integrate information from preceding frames into the current BEV representation through a self-attention mechanism. In this framework, the query corresponds to the BEV features of the current frame, while the values are constructed by concatenating BEV features from both current and previous frames. This enables the network to capture temporal continuity and mitigate motion-induced inconsistencies.

Here, the BEV features are further enhanced by integrating multisource data from both vehicle-mounted and roadside sensing systems. Specifically, the cooperative vehicle-infrastructure BEV feature fusion module (detailed in Section III-D) combines spatial representations from both perspectives. By leveraging the complementary viewpoints of vehicle onboard sensors and roadside infrastructure, this fusion strategy generates a more comprehensive and robust BEV representation, improving perception accuracy in complex traffic environments.

D. Vehicle-Infrastructure BEV Feature Fusion

The BEV features $F^{\text{bev}}_{\text{road}}$ and $F^{\text{bev}}_{\text{vehicle}}$ exhibit fundamentally different error characteristics due to their observation perspectives. Simple feature stacking fails to align features from the same spatial region accurately. As depicted in Fig. 1, the V2I BEV Fusion Block leverages deformable mutual attention to suppress spatial interference from irrelevant regions and effectively fuse roadside and vehicle-mounted BEV features in relevant areas. We employ BEV queries to extract features from both roadside and vehicle-side BEV data. Instead of processing all spatial positions, the model focuses on key locations near each BEV query for feature aggregation, adaptively assigning attention weights to these points.

Given the BEV features $F^{\text{bev}}_{\text{road}} \in \mathbb{R}^{X\times Y\times C_F}$ and $F^{\text{bev}}_{\text{vehicle}} \in \mathbb{R}^{X\times Y\times C_F}$, we first apply a 1×1 convolution to align their feature channels. Subsequently, we generate query BEV features $Q_p \in \mathbb{R}^{X\times Y\times C_F}$ as follows:

$$Q_p = \text{conv}_{1\times1}\left(F^{\text{bev}}_{\text{road}} \oplus F^{\text{bev}}_{\text{vehicle}}\right) \tag{11}$$

where Q_p denotes the query at position $p = (x, y)$ from the concatenated features. These queries merge roadside and vehicle features, capturing spatial context effectively. For precise fusion, we apply deformable mutual-attention to the roadside feature map, computed as

$$\begin{aligned}&\text{DeformAttn}\left(Q_p, p, F^{\text{bev}}_{\text{road}}\right)\\ &= \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{\text{mqk}} \cdot W'_m F^{\text{bev}}_{\text{road}}\left(p + \Delta p_{\text{mqk}}\right)\right]\end{aligned} \tag{12}$$

where M is the number of attention heads, m indexes the attention head, K is the total number of sampled keys, and k indexes the sampled keys. $A_{\text{mqk}} \in [0, 1]$ is the attention weight, representing the correlation of the kth key to the qth query in the mth attention head, with $\sum_{k=1}^{K} A_{\text{mqk}} = 1$. Δp_{mqk} represents the predicted offset based on the query feature Q_p for the reference point p. $F^{\text{bev}}_{\text{road}}(p + \Delta p_{\text{mqk}})$ denotes the input feature at the $p + \Delta p_{\text{mqk}}$. W_m aggregates the multihead attention outputs, and W'_m projects the input features for each head.

The same process applies to vehicle-side feature $F^{\text{bev}}_{\text{vehicle}}$. The fused BEV features $F^{\text{bev}}_{\text{fused}}$ at spatial position p are then computed as

$$\begin{aligned}F^{\text{bev}}_{\text{fused}} &= \text{BevFuse}\left(F^{\text{bev}}_{\text{road}}, F^{\text{bev}}_{\text{vehicle}}\right)\\ &= \sum_{p=0}^{HW}\left[Q'_p + \sum_{V\in\left\{F^{\text{bev}}_{\text{road}}, F^{\text{bev}}_{\text{vehicle}}\right\}} \text{DeformAttn}\left(Q_p, p, V\right)\right]\end{aligned} \tag{13}$$

where H and W are the spatial shape of BEV features, Q'_p is the sampled feature at the reference point p from the $F^{\text{bev}}_{\text{road}}$ in roadside scenarios or the $F^{\text{bev}}_{\text{vehicle}}$ in vehicle-mounted scenes. The symbol V represents the input feature set used for fusion at each spatial position p.

This approach offers two key benefits. First, it selectively aggregates features from reliable locations, avoiding corrupted regions, such as those affected by occlusions or sensor noise. Second, it assigns lower attention weights to less reliable areas, prioritizing high-quality features.

IV. Experiments and Results

We present the results of our V2I collaborative perception approach evaluated on the DAIR-V2X dataset. The experimental setup is briefly introduced, along with a benchmark dataset for V2I perception. We then compare the proposed V2IFormer with state-of-the-art methods.

TABLE I

COMPARISON OF 3-D OBJECT DETECTION PERFORMANCE ACROSS DIFFERENT FUSION METHODS. AB REPRESENTS THE AVERAGE TRANSMISSION SIZE IN BYTES. "-" DENOTES NO INFORMATION PROVIDED. V2IFORMER OUTPERFORMS ALL FUSION STRATEGIES, PARTICULARLY IN EARLY FUSION, ACHIEVING SIGNIFICANT IMPROVEMENTS IN $mAP_{3\text{-}D}$ AND mAP_{BEV}, WHILE MAINTAINING 1/10 THE TRANSMISSION COST OF EARLY FUSION METHODS. QUANTITATIVE EVALUATION RESULTS ON THE DAIR-V2X-C DATASET ARE PRESENTED IN THE TABLE BELOW, WITH THE BEST VALUES HIGHLIGHTED IN BOLD. ALL RESULTS ARE REPORTED AS PERCENTAGES (%)

Fusion	Model	$mAP_{3D(IOU=0.5)}$	$mAP_{3D(IOU=0.7)}$	$mAP_{BEV(IOU=0.5)}$	$mAP_{BEV(IOU=0.7)}$	AB (Byte)
Single Vehicle	ImvoxelNet [28]	12.03	-	13.62	-	0
Early Fusion	PointPillars [53]	57.35	-	64.06	-	1382.28K
Late Fusion	TCLF [6]	40.79	-	46.80	-	539.60
Intermediate Fusion	DiscoNet [54]	52.83	29.19	61.25	50.11	120K
Intermediate Fusion	V2VNet [55]	52.02	28.54	60.78	50.02	120K
Intermediate Fusion	FFNet [56]	55.81	30.23	63.54	54.16	120K
Intermediate Fusion	V2IFormer (Ours)	68.93	48.31	75.52	63.60	146.24K

A. V2I Collaborative Perception-Experiments

1) DAIR-V2X Dataset: Many V2I studies use simulated datasets with idealized V2X communication, which do not reflect real-world conditions. We evaluate multiview 3-D object detection using the DAIR-V2X dataset [6], which contains real-world frames. Released in 2022 by Baidu Apollo and Tsinghua University's Institute for AI Industry Research (AIR), DAIR-V2X is the first public benchmark for vehicle-infrastructure cooperative autonomous driving. It covers 10 km of urban roads, 10 km of highways, and 28 intersections in Beijing's High-Level Autonomous Driving Demonstration Zone, capturing diverse traffic scenarios across various weather (sunny, rainy, and foggy) and lighting conditions (day and night) in urban and highway environments. The DAIR-V2X dataset offers rich multimodal data, including raw images, LiDAR point clouds, timestamped annotations, and calibration files, supporting research in cooperative 3-D object detection and related perception tasks. Its DAIR-V2X-C subset includes 38 845 frames of images and point clouds for cooperative detection. We use the VIC-Sync portion of DAIR-V2X-C, comprising 9311 synchronized vehicle-infrastructure frame pairs, for training and evaluation. Annotations, originally in world coordinates, are converted to vehicle coordinates for the experiments, with translated labels serving as the benchmark. The dataset is split into training, validation, and test sets with a 5:2:3 ratio. All models are evaluated on the validation set using the KITTI-defined average precision (AP) for 3-D bounding boxes [17].

2) NuScenes Dataset: We evaluate the proposed method on the nuScenes dataset [57], a widely used benchmark for autonomous driving. It contains 1000 multimodal video sequences of about 20 s each, with keyframes annotated at 2 Hz. Each keyframe includes synchronized data from six cameras covering 360°, one LiDAR, five radars, GPS, and IMU. In total, the dataset provides 28 130 training samples and 6019 validation samples.

The V2XSet dataset: V2XSet is a large-scale V2X perception dataset generated in simulation using CARLA [58] and OpenCDA [59]. It is specifically designed to address the lack of fully public datasets for evaluating real-world challenges such as localization errors and transmission delays. To this end, V2XSet uniquely integrates realistic noise into V2X collaborative perception. The dataset contains 11 447 frames, corresponding to 33 081 samples when accounting for frames per agent, and is split into 6694/1920/2833 samples for training, validation, and testing, respectively. V2XSet covers 55 distinct scenes across eight towns in CARLA. These scenes span five roadway types: straight segments, curvy roads, mid-blocks, entrance ramps, and intersections. Intersection scenes are deliberately oversampled due to their inherent complexity, high traffic volume, and severe occlusions. Each scene lasts up to 25 s and involves between two and seven interacting agents. All agents, including both vehicles and infrastructure units, are equipped with 32-channel LiDAR sensors operating at 10 Hz with a 120-m range. The sensor placement is carefully designed to highlight geometric disparities. Infrastructure sensors are mounted on traffic or street light poles at a height of 4.3 m and are deployed at key locations such as intersections, mid-blocks, and ramps. This high-altitude placement contrasts with the lower mounting height of vehicle sensors. As a result, the inherent variation in installation height leads to distinct measurement patterns that faithfully reflect the differing sensor perspectives of infrastructure and vehicle agents.

3) Evaluation Metrics: The DAIR-V2X dataset [6] defines evaluation metrics for performance assessment: AP [18] for 3-D object detection accuracy and average byte (AB) for transmission cost. AP compares predicted 3-D object detections with ground-truth annotations in the DAIR-V2X-C dataset. For VIC3D evaluation, focusing on the vehicle's egocentric perception, we filter objects outside a cubic region in the vehicle coordinate system, defined as $x \in [x_{\min}, x_{\max}]$ m, $y \in [y_{\min}, y_{\max}]$ m, $z \in [z_{\min}, z_{\max}]$ m. Metrics are evaluated across detection ranges: overall 0–100 m, using an IoU threshold of 0.5 for two components: $AP_{3\text{-}D}$ and AP_{BEV}. AB measures the average size of transmitted data.

4) Implementation Details: To evaluate performance, we compare the proposed V2IFormer with several state-of-the-art methods on the DAIR-V2X-C benchmark. We compared our proposed V2IFormer with four different fusion methods: nonfusion (e.g., ImvoxelNet [28]), early fusion (e.g., PointPillars [53]), late fusion (e.g., TCLF [6]), and middle fusion (e.g., DiscoNet [54], V2VNet [55], and FFNet [56]). The performance of these methods was evaluated on the DAIR-V2X-C dataset using the KITTI evaluation metrics [6]: BEV mAP and 3-D mAP, with IoU thresholds of 0.5. Models are trained for 150 epochs using the AdamW

optimizer [60] with an initial learning rate of 2×10^{-4} on eight NVIDIA RTX 3090 GPUs.

B. Benchmark Results

1) Results on the DAIR-V2X-C Dataset and Performance Comparison: Table I presents the 3-D object detection performance of different methods on the DAIR-V2X-C dataset. It is clear that all cooperative-based methods consistently achieve significantly improved 3-D object detection performance compared to the single-vehicle approach. To ensure the practical relevance of this performance, our analysis assumes a high-precision V2X communication environment where the synchronization offset between the vehicle-side and infrastructure-side sensors is maintained within ±10 ms. This strict temporal alignment is a prerequisite for the feature fusion process, ensuring that the aggregated data represents a consistent physical state of the target vehicle. V2IFormer outperforms all other fusion strategies, particularly with its intermediate fusion approach. First, the proposed V2IFormer model substantially outperforms the nonfusion image-based method ImvoxelNet, with an improvement of 61.90% in mAP_{BEV} (IoU = 0.5) and 56.90% in $\text{mAP}_{\text{3-D}}$ (IoU = 0.5). These results confirm that incorporating infrastructure data significantly enhances 3-D detection performance. Second, although the late fusion method TCLF requires low transmission bandwidth (539.60 B), it achieves a mAP_{BEV} (IoU = 0.5) of only 46.80%, which is 28.72% lower than that of V2IFormer. This indicates that, while late fusion is efficient in terms of communication, it considerably compromises detection accuracy. Third, early fusion methods such as PointPillars yield a mAP_{BEV} (IoU = 0.5) of 64.06%, but demand high data transmission (1382.28 kB). In contrast, intermediate fusion methods achieve markedly better performance with greatly reduced bandwidth. Among intermediate fusion models, V2IFormer achieves the highest performance across all evaluation metrics. It exceeds DiscoNet by 14.10% in $\text{mAP}_{\text{3-D}}$ (IoU = 0.5) and 14.27% in mAP_{BEV} (IoU = 0.5), and outperforms FFNet by 13.12% and 11.98% in the same metrics, respectively, while operating at a similar communication cost (146.24 kB). In summary, V2IFormer establishes a new state-of-the-art on the DAIR-V2X-C dataset. The results demonstrate that our intermediate fusion scheme not only achieves superior accuracy but also optimizes the bandwidth-latency tradeoff. By requiring only 1/10 of the bandwidth of early fusion, the proposed method ensures that transmission delays remain well within the synchronization margin, thereby enhancing the robustness and practical feasibility.

Further analysis on the DAIR-V2X-C dataset, which defines short-range as 0–30 m, mid-range as 30–50 m, and long-range as 50–100 m. The experimental results presented in Table II highlight the performance of various fusion strategies on the DAIR-V2X-C dataset, focusing on both 3-D object detection $\text{AP}_{\text{3-D}}$ and BEV detection AP_{BEV} across different distance ranges. Notably, the single-vehicle method based on ImVoxelNet exhibits limited performance, especially in long-range detection (50–100 m), where $\text{AP}_{\text{3-D}}$ reaches only 2.28% and AP_{BEV} is 2.82%. This is primarily due to its reliance on monocular camera data, which lacks precise depth estimation, thus failing to capture distant objects accurately. Early fusion methods, such as PointPillars, significantly outperform the single-vehicle approach, particularly at mid-range distances (30–50 m), achieving $\text{AP}_{\text{3-D}}$ of 60.38% and AP_{BEV} of 64.08%. However, these methods suffer from high communication overhead due to the large volume of data transmitted between vehicles and infrastructure, which hinders the efficiency and low-latency response required for practical deployment. In contrast, the proposed intermediate fusion strategy (V2IFormer) excels across all metrics, achieving $\text{AP}_{\text{3-D}}$ of 62.16% in the 0–30-m range and 56.89% in the 50–100-m range, with AP_{BEV} reaching 69.85% at 0–30 m and 64.40% at 50–100 m. It achieves a robust performance in long-range detection while minimizing data transmission overhead, making it a highly scalable solution for real-world cooperative vehicle systems. Overall, the results clearly demonstrate that intermediate feature fusion, particularly with V2IFormer, provides superior accuracy, reliability, and bandwidth efficiency, making it a promising solution to the challenges of V2X perception systems in autonomous driving.

2) Handling Dynamic Weathers and Lighting Conditions: V2IFormer is evaluated on the real-world datasets that naturally encompass diverse and dynamically changing environmental conditions, including variations in challenging weather (rainy and foggy). While recent advances in 3-D object detection have achieved remarkable performance on standard vehicle-centric benchmarks such as KITTI [17], nuScenes [57], and Waymo [61], prior studies have shown that their performance under adverse conditions remains limited [62], [63]. In contrast, the DAIR-V2X cooperative perception dataset used in our study is collected in the wild and covers a wide variety of real-world environmental conditions. However, DAIR-V2X does not provide frame-level or scene-level annotations specifying the exact weather type for each sample. This lack of fine-grained environmental metadata makes it impossible to reliably partition the dataset into distinct categories such as "rain-only," "fog-only," or "clear-daylight," which are necessary for quantitative robustness benchmarking. We adopt qualitative robustness assessments to evaluate V2IFormer's behavior in challenging conditions. As illustrated in Fig. 6, we present comparisons under foggy, rainy, and occluded scenarios using ImvoxelNet [28], PointPillars [53], and our V2IFormer. Across all these scenarios, V2IFormer consistently produces more complete and accurate detections, handling partially occluded vehicles, small pedestrians, and distant or low-visibility targets more effectively. This robustness stems from V2IFormer's bilateral BEV representation and deformable BEV feature fusion, which enable it to exploit complementary cues from vehicle-side and infrastructure-side sensors, even when one modality is degraded by adverse weather or lighting.

While V2IFormer demonstrates compelling performance on standard benchmarks, its robustness under adverse weather and occlusion is critical for real-world V2I deployment. Although the DAIR-V2X dataset does not provide explicit

TABLE II
QUANTITATIVE EVALUATION RESULTS ON THE DAIR-V2X-C DATASET ARE PRESENTED IN THE TABLE BELOW, WITH THE BEST VALUES HIGHLIGHTED IN BOLD. ALL SCORES ARE REPORTED IN PERCENTAGES (%)

Fusion	Model	AB (Byte)	$AP_{3D}(IoU = 0.5)$			$AP_{BEV}(IoU = 0.5)$		
			0-30m	30-50m	50-100m	0-30m	30-50m	50-100m
Single Vehicle	ImvoxelNet [28]	0	16.25	7.25	2.28	17.66	8.58	2.82
Early Fusion	PointPillars [53]	1382.28K	53.07	60.38	33.05	55.80	64.08	36.17
Late Fusion	TCLF [6]	539.60K	34.67	29.69	15.76	38.24	34.27	19.40
Intermediate Fusion	V2IFormer (Ours)	146.24K	62.16	57.75	56.89	69.85	62.31	64.40

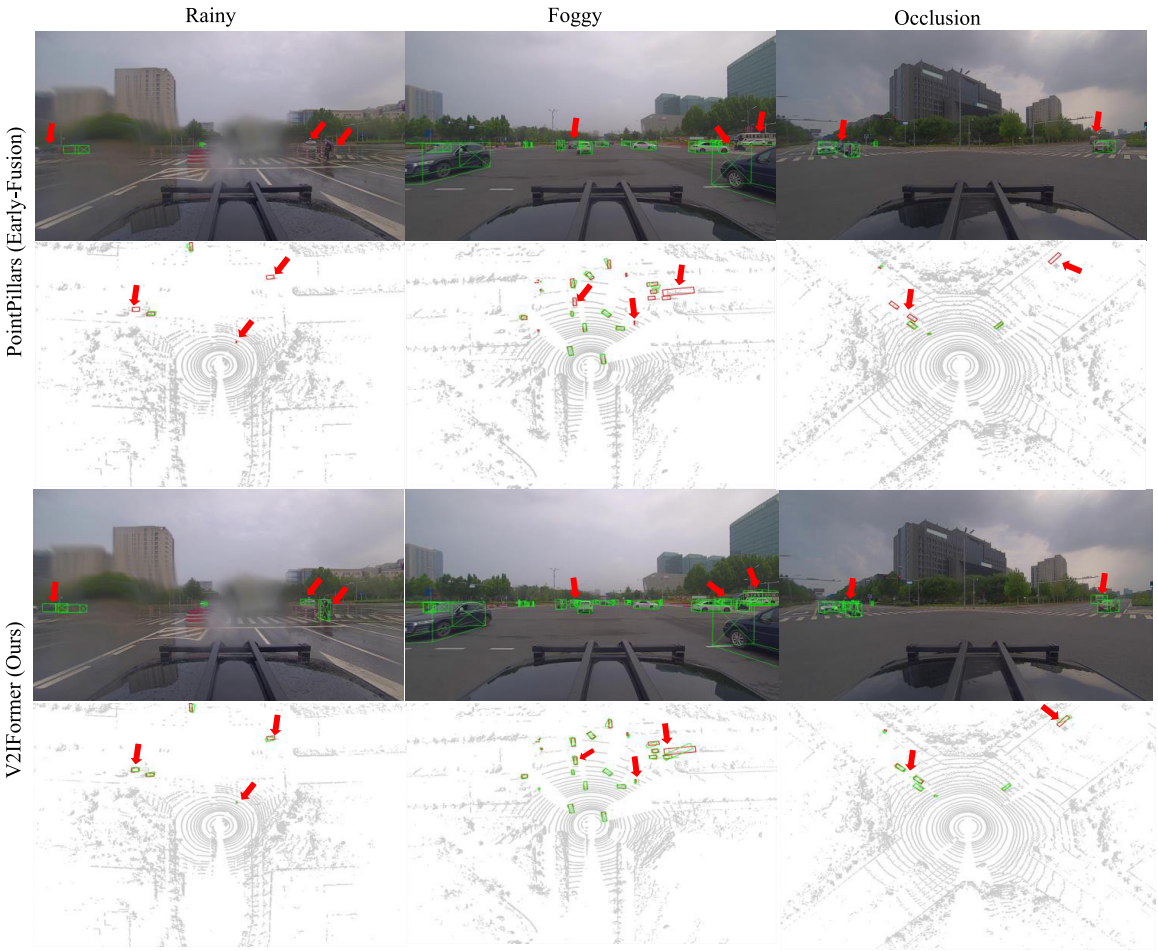


Fig. 6. Qualitative comparisons in occlusion conditions and various weather scenarios. V2IFormer correctly detects partially occluded vehicles as well as distant vehicles and pedestrians in rainy and foggy scenes.

per-frame weather metadata, we employ quantitative proxies to evaluate the model's stability. Specifically, we manually curated representative subsets from the test set to conduct controlled evaluations under three environmental categories: normal (clear daylight), low-light (dusk/night), and adverse weather (rainy/foggy). Furthermore, we quantified robustness against occlusion by specifically evaluating targets in occluded scenarios. As summarized in Table III, the quantitative results across these proxies further substantiate our robustness claims. In terms of environmental variations, while the single-vehicle baseline (no fusion) suffers a performance drop of approximately 24.39 percentage points (from 48.52% to 24.13%) under adverse weather, V2IFormer maintains a significantly more stable performance floor. Notably, under adverse weather and occlusion, V2IFormer achieves mAP scores of 48.54% and 30.25%, respectively. More importantly, the advantage of our cooperative scheme becomes more pronounced as environmental conditions degrade. In occluded scenarios where the vehicle-side perspective loses critical visual cues, V2IFormer achieves a significant improvement over the single-vehicle approach, increasing the mAP from 12.42% to 30.25%. This quantitative analysis confirms that the bilateral

TABLE III
QUANTITATIVE ROBUSTNESS EVALUATION OF V2IFORMER UNDER CHALLENGING CONDITIONS ON THE DAIR-V2X DATASET

Environment	Single-Vehicle	V2IFormer
Normal (clear)	48.52%	69.21%
Low-light (dusk/night)	32.64%	58.42%
Adverse (rain/fog)	24.13%	48.54%
Occlusion	12.42%	36.25%

BEV feature fusion effectively leverages infrastructure-side perspectives to compensate for local modality degradation, ensuring reliable robustness in challenging real-world scenarios.

3) Quantitative Evaluation and Robustness Analysis on V2XSet: Table IV presents the quantitative results on V2XSet under two evaluation settings. The perfect setting assumes ideal V2X communication with accurate pose information and zero transmission delay, representing the upper bound of cooperative perception performance. In contrast, the noisy setting introduces localization disturbances and communication delays, providing a more realistic and stringent benchmark for assessing practical robustness. Under the perfect setting, all cooperative perception approaches outperform the single-agent baseline. Our V2IFormer achieves the best overall performance, reaching 0.925 AP@0.5 and 0.864 AP@0.7, outperforming all existing fusion methods. Other cooperative models, such as V2X-ViT (0.882 AP@0.5/0.712 AP@0.7) and DiscoNet (0.844 AP@0.5/0.695 AP@0.7), also exhibit significant improvements over No Fusion, confirming the benefits of multiagent feature fusion. The performance gap widens notably under the noisy setting. Traditional fusion strategies degrade severely when exposed to pose errors and delay. Late fusion and early fusion drop to 0.307 AP@0.7 and 0.384 AP@0.7, respectively—both falling below the no fusion baseline (0.402 AP@0.7). These results highlight the vulnerability of basic fusion mechanisms to spatial misalignment and temporal inconsistencies. In contrast, V2IFormer demonstrates remarkable robustness. It achieves 0.910 AP@0.5 and 0.780 AP@0.7 under the noisy setting, outperforming all other models by a considerable margin. Compared with V2X-ViT (0.836 AP@0.5/0.614 AP@0.7), V2IFormer delivers a 0.166 AP@0.7 improvement, indicating a significantly stronger ability to handle V2X noise. Moreover, V2IFormer exhibits the smallest performance drop between perfect and noisy settings (only 0.084 AP@0.7), whereas competing methods suffer much larger declines (e.g., V2X-ViT: 0.098, OPV2V: 0.177, and V2VNet: 0.184). Overall, these results demonstrate that the proposed bilateral BEV representation and deformable BEV feature fusion are highly effective in mitigating V2X noise, enabling V2IFormer to maintain stable and reliable perception performance in realistic deployment scenarios.

4) Analysis on V2XSet Time Delay Analysis: As shown in Table V, V2IFormer maintains strong detection performance even under substantial communication delays. AP@0.7 decreases gradually from 0.571 at 100 ms to 0.426 at 400 ms, and AP@0.5 decreases from 0.729 to 0.469 over the same range. This indicates that V2IFormer remains reliable and robust across a wide spectrum of delay conditions. Notably, even at 400 ms, which far exceeds typical V2X latency in real systems, V2IFormer still achieves 0.426 AP@0.7. Across all delay levels, V2IFormer consistently outperforms the single-agent No Fusion baseline. At 100 ms, V2IFormer achieves 0.571 AP@0.7, exceeding no fusion (0.402) by 17%. Even under an extreme 400 ms delay, V2IFormer maintains 0.426 AP@0.7, slightly above no fusion. These results clearly demonstrate that V2IFormer is significantly more resilient to temporal misalignment, maintaining a stable advantage even under severe communication delays.

TABLE IV
3-D DETECTION PERFORMANCE COMPARISON ON V2XSET

Models	Noisy		Perfect	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Fusion	0.606	0.402	0.606	0.402
Late Fusion	0.549	0.307	0.727	0.620
Early Fusion	0.720	0.384	0.819	0.710
DiscoNet [54]	0.798	0.541	0.844	0.695
V2VNet [55]	0.791	0.493	0.845	0.677
OPV2V [3]	0.709	0.487	0.807	0.664
V2X-ViT [7]	0.836	0.614	0.882	0.712
V2IFormer (Ours)	0.910	0.780	0.925	0.864

TABLE V
TIME DELAY ANALYSIS ON AP@0.5 AND AP@0.7

Delay	AP@0.5	AP@0.7
100 ms	0.729	0.571
200 ms	0.612	0.442
300 ms	0.547	0.438
400 ms	0.469	0.426

5) Inference Time Analysis on V2XSet: To evaluate the practical deployability of our framework, we analyze the computational efficiency and per-frame inference time of V2IFormer. Table VI presents the runtime of the proposed V2IFormer, including the asymmetric dual-side BEV encoders and the deformable mutual-attention fusion module. The overall perception pipeline achieves an average inference time of 108 ms per frame, providing the low-latency response necessary for efficient cooperative perception in autonomous driving systems. The lightweight design of HeightNet and UAHM reduces the computational cost of BEV lifting on both sides, while the sparsity-aware implementation of the mutual-attention mechanism minimizes redundant computation by focusing attention on spatially relevant BEV regions. These characteristics enable V2IFormer to maintain the high computational efficiency required for cooperative perception systems while preserving state-of-the-art accuracy. This analysis confirms that V2IFormer strikes a favorable balance between computational efficiency and detection performance, making it well-suited for deployment.

6) Hardware Efficiency and Latency Analysis: To support our claim of real-world deployment potential, we performed a careful evaluation of the model's computational requirements and speed. The model's key performance metrics, evaluated on an NVIDIA RTX 3090 GPU, are summarized in Table VII. The results in Table VII clearly show that our model is

TABLE VI
INFERENCE TIME OF EACH MODULE OF THE ALGORITHM (ms)

Delay	Inference time
Extracting roadside BEV features	61
Extracting vehicle-side BEV features	76
BEV feature fusion	32
Average total time	108

TABLE VII
SUMMARY OF COMPUTATIONAL AND RUN TIME PERFORMANCE METRICS

Metric	Model parameters	GPU memory	Latency	FPS
Value	13.54M	1.43G	108 ms	9.28

TABLE VIII
ABLATION STUDY OF COMPONENTS IN V2IFORMER

LID	UAHM	VIFM	mAP@3D
✓			40.60
✓	✓		42.10
✓	✓		52.04
✓	✓	✓	68.93

TABLE IX
ABLATION STUDY ON HEIGHT MODELING

Model	Height Modeling Approach	mAP@3D
BEVFormer	Implicit	28.80
V2IFormer (Ours)	Explicit	29.20

well-suited for use in actual vehicles. First, the architecture is very hardware-efficient: it uses only 13.54-M parameters and has an extremely low peak GPU memory usage of 1.43 GB. This small resource footprint is essential for smooth deployment onto resource-limited embedded platforms, which are the required hardware for automotive integration. Furthermore, the model reliably meets the strict speed requirements for autonomous driving. The measured average inference latency is 108 ms, leading to a processing speed of 9.28 FPS. These results validate the framework's computational efficiency and its potential for practical deployment.

C. Ablation Studies

We conduct a series of ablations to verify the core components of V2IFormer and present the findings of each experiment.

1) Key Components in V2IFormer: Our experiments evaluate three key components in V2IFormer (as shown in Table VIII). Specifically, "LID" refers to the LID applied within the infrastructure-side block, "UAHM" denotes the UAHM in the vehicle-side BEV generation module, and "VIFM" indicates the vehicle-infrastructure BEV feature fusion performed through the offset-learning deformable mutual-attention module. We progressively activated these components to assess their individual and combined contributions. When "VIFM" is disabled, we adopt a standard cross-attention mechanism as a baseline. The results clearly show that each module contributes to performance improvement. First, adding just the LID module improves 3-D detection accuracy by 1.5% (from 40.60% to 42.10%). When we combine LID with UAHM, performance improves from 42.10% to 52.04%, showing UAHM's strong impact. The complete model with all three components reaches 68.93%, with VIFM contributing an extra 16.89%. This step-by-step improvement proves each component matters: LID enhances infrastructure-side data processing, UAHM addresses uncertainty in vehicle-side data, and VIFM effectively integrates both data sources.

2) Effectiveness of Explicit Height Modeling: Table IX reports an ablation study on different height-lifting strategies for BEV feature generation using the nuScenes dataset. We compare our approach with representative methods, such as the implicit height lifting adopted in BEVFormer. The proposed Gaussian-based adaptive height lifting aggregates image features along the vertical dimension through a probabilistic representation, which captures vertical context, alleviates quantization artifacts, and preserves computational efficiency. Experimental results show that our method achieves the best performance, with an mAP of 29.20%, outperforming BEVFormer by 0.4%. These results verify the effectiveness of the proposed height modeling strategy in enhancing 3-D perception.

Note: "Implicit" refers to height information derived indirectly from 2-D features without explicit 3-D coordinate modeling. "Explicit" refers to height modeled directly using 3-D spatial coordinates, improving modeling accuracy.

3) Ablation Study on Sampling Density and Attention Heads: We evaluate the impact of sampling density K and the number of attention heads M to verify the efficiency and scalability of the deformable mutual-attention module. As shown in Table X, detection performance is significantly constrained by sparse sampling configurations. Increasing the sampling density K from 16 to 64 leads to a substantial improvement in mAP (from 74.3% to 79.7%). This indicates that a higher sampling density allows the module to capture more intricate spatial dependencies. However, performance gains diminish beyond $K = 64$, with mAP improving by only 0.1% while latency increases to a prohibitive 194.55 ms. We attribute this to the sampling points having sufficiently covered the key local features, making additional points redundant for representation learning. Table XI highlights the tradeoff between representation diversity and computational overhead. Scaling M from 2 to 8 boosts the mAP by 5.1%, maintaining a manageable latency of 107.76 ms. Beyond this point, doubling M to 16 leads to performance saturation (78.7% mAP) accompanied by a severe latency penalty. To balance high-fidelity 3-D detection with the low-latency response required for cooperative autonomous driving, we choose $K = 32$ and $M = 8$ as the default configuration.

4) Robustness of Height Modeling: As discussed in the introduction, both depth and height modeling contribute to 2-D-to-3-D mapping, but height modeling offers superior robustness to variations in camera setup. Depth estimation in the image space is highly sensitive to parameters such as

TABLE X
ABLATION STUDY ON SAMPLING DENSITY K REGARDING DETECTION ACCURACY AND LATENCY

Sampling Density (K)	mAP (%)	Latency (ms)
16	74.3%	85.20
32	78.5%	107.76
64	79.7%	142.35
128	79.8%	194.55

TABLE XI
IMPACT OF THE NUMBER OF ATTENTION HEADS M ON PERFORMANCE AND COMPUTATIONAL OVERHEAD

Number of Attention Heads (M)	mAP (%)	Latency (ms)
2	73.1	92.45
4	75.8	98.12
8	78.2	107.76
16	78.7	156.30

TABLE XII
ABLATION STUDY ON THE ROBUSTNESS OF HEIGHT MODELING

Model	Back camera mAP	Overall mAP
BEVDepth (Depth modeling)	27.30	33.30
V2IFormer (Height modeling)	27.90	29.20

focal length and mounting position, whereas height modeling projects features into a unified BEV space, thereby reducing this sensitivity and yielding more stable results. To validate this advantage, we compared our method with a depth modeling approach based on explicit height lifting, represented by the LSS module in BEVDepth [9]. As shown in Table XII, although BEVDepth benefits from LiDAR supervision and dedicated network designs, our purely vision-based height modeling achieves comparable accuracy across multiple metrics. Moreover, our method outperforms BEVDepth on the rear camera with varying focal lengths and configurations, further demonstrating its robustness. These findings highlight the practicality and reliability of height modeling in real-world applications.

V. CONCLUSION

In this article, we proposed V2IFormer, a robust framework for V2I cooperative 3-D object detection that leveraged multiview monocular cameras to achieve accurate perception in complex environments. The proposed approach employed an intermediate fusion strategy, transmitting BEV features to balance communication efficiency and semantic richness, addressing bandwidth constraints and mitigating information loss in V2I perception. We introduce a dual-side BEV representation framework: infrastructure-side perception integrates height distribution prediction with semantic encoding, while vehicle-mounted BEV generation employs a Gaussian-based height modeling approach to enhance feature representation. A deformable mutual-attention mechanism further enhanced fusion performance by dynamically aligning and selecting relevant spatial features across vehicle and infrastructure viewpoints, ensuring robustness in challenging scenarios. Extensive evaluation on the DAIR-V2X benchmark validated V2IFormer's state-of-the-art detection accuracy and consistent performance under occluded scenes and diverse weather conditions. Future research will explore edge-assisted processing to ensure efficient deployment in large-scale autonomous vehicle networks.

ACKNOWLEDGMENT

Tianxuan Fu and Keyin Wang are with the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: futianxuan@stu.xidian.edu.cn; keyinwang@stu.xidian.edu.cn).

Guoqiang Mao is with the School of Transportation, Southeast University, Nanjing 211189, China (e-mail: g.mao@ieee.org).

Xiaojiang Ren is with Guangzhou Research Institute, Xidian University, Xi'an 710071, China (e-mail: xjren@xidian.edu.cn).

Wei Xiang is with the Cisco-La Trobe Centre for AI and IoT, La Trobe University, Melbourne, VIC 3086, Australia (e-mail: w.xiang@latrobe.edu.au).

Xiaohu Ge is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xhge@hust.edu.cn).

REFERENCES

[1] M. Schwall, T. Daniel, T. Victor, F. Favaro, and H. Hohnhold, "Waymo public road safety performance data," 2020, *arXiv:2011.00038*.

[2] D. Qiao and F. Zulkernine, "Adaptive feature fusion for cooperative perception using LiDAR point clouds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1186–1195.

[3] 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 *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589.

[4] L. Yang et al., "BEVHeight: A robust framework for vision-based roadside 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 21611–21620.

[5] M. Noor-A-Rahim et al., "6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proc. IEEE*, vol. 110, no. 6, pp. 712–734, Jun. 2022.

[6] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 21329–21338.

[7] R. Xu, X. Hao, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 107–124.

[8] Z. Wang et al., "VIMI: Vehicle-infrastructure multi-view intermediate fusion for camera-based 3D object detection," 2023, *arXiv:2303.10975*.

[9] Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 1477–1485.

[10] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 2020–2036, Mar. 2025.

[11] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[12] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5935–5943.

[13] J. Schramm et al., "BEVCar: Camera-radar fusion for BEV map and object segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 1435–1442.

[14] J. Wang, F. Li, Y. An, X. Zhang, and H. Sun, "Toward robust LiDAR-camera fusion in BEV space via mutual deformable attention and temporal aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5753–5764, Jul. 2024.

[15] M.-Q. Dao, J. S. Berrio, V. Frémont, M. Shan, E. Héry, and S. Worrall, "Practical collaborative perception: A framework for asynchronous and multi-agent 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 12163–12175, Sep. 2024.

[16] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2774–2781.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[19] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, "3D object detection from images for autonomous driving: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3537–3556, May 2024.

[20] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9287–9296.

[21] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 913–922.

[22] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, Nov. 2021, pp. 180–191.

[23] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 531–548.

[24] Y. Liu et al., "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3239–3249.

[25] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3598–3608.

[26] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3D object detection," 2018, *arXiv:1811.08188*.

[27] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. ECCV*, Aug. 2020, pp. 194–210.

[28] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1265–1274.

[29] C.-Y. Tseng, Y.-R. Chen, H.-Y. Lee, T.-H. Wu, W.-C. Chen, and W. H. Hsu, "CrossDTR: Cross-view and depth-guided transformers for 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4850–4857.

[30] X. Ye et al., "Rope3D: The roadside perception dataset for autonomous driving and monocular 3D object detection task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21309–21318.

[31] S. Fan, Z. Wang, X. Huo, Y. Wang, and J. Liu, "Calibration-free BEV representation for infrastructure perception," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 9008–9013.

[32] S.-W. Kim et al., "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 663–680, Apr. 2015.

[33] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 514–524.

[34] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2021, pp. 953–959.

[35] Q. Chen, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 88–100.

[36] E. E. Marvasti, A. Raftari, and Y. P. Fallah, "Cooperative LiDAR object detection via feature sharing in deep networks," U.S. Patent 17 928 473, Aug. 24, 2023.

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

[38] Z. Liu et al., "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," 2022, *arXiv:2205.13542*.

[39] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.

[40] W. Yang et al., "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15531–15540.

[41] J. Yan et al., "Cross modal transformer: Towards fast and robust 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 18222–18232.

[42] H. A. Mallot, H. H. Bülthoff, J. J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biol. Cybern.*, vol. 64, no. 3, pp. 177–185, Jan. 1991.

[43] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8445–8453.

[44] Z. Teng et al., "360bev: Panoramic semantic mapping for indoor bird's-eye view," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 373–382.

[45] Z. Lin et al., "RCBEVDet: Radar-camera fusion in Bird's eye view for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14928–14937.

[46] Z. Lin, Z. Liu, Y. Wang, L. Zhang, and C. Zhu, "Rcbevdet++: Toward high-accuracy radar-camera fusion 3D perception network," 2024, *arXiv:2409.04979*.

[47] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[49] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9308–9316.

[50] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[51] Y. Tang, S. Dörn, and C. Savani, "Center3D: Center-based monocular 3D object detection with joint depth understanding," in *Proc. DAGM German Conf. Pattern Recognit.*, Sep. 2020, pp. 289–302.

[52] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8551–8560.

[53] A. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12689–12697.

[54] Y. Li, S. Ren, P. Wu, S. Chen, F. Chen, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 29541–29552.

[55] 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 *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 605–621.

[56] H. Yu et al., "Vehicle-infrastructure cooperative 3D object detection via feature flow prediction," 2023, *arXiv:2303.10552*.

[57] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[58] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, vol. 78, Nov. 2017, pp. 1–16.

[59] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "OpenCDA: An open cooperative driving automation framework integrated with co-simulation," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Mar. 2021, pp. 1155–1162.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[61] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2443–2451.

[62] M. Hahner et al., "LiDAR snowfall simulation for robust 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16343–16353.

[63] M. Hahner, C. Sakaridis, D. Dai, and L. V. Gool, "Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 15283–15292.

Tianxuan Fu received the master's degree in transport information engineering and control from Lanzhou Jiaotong University, Lanzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Communication Engineering, Xidian University, Xi'an, China.

His research interests include object detection, target tracking, data fusion, and intelligent transportation systems.

Guoqiang Mao (Fellow, IEEE) was a leading Professor, the Founding Director of the Research Institute of Smart Transportation, and the Vice-Director of the ISN State Key Laboratory, Xidian University, Xi'an, China, from 2014 to 2019. Before that, he was with the University of Technology Sydney, Ultimo, NSW, Australia, and the University of Sydney, Sydney, NSW, Australia. He is the Chair Professor and the Director of the Center for Smart Driving and Intelligent Transportation Systems, Southeast University, Nanjing, China. He has published 300 papers in international conferences and journals that have been cited more than 15 000 times. His H-index is 57 and is in the list of the top 2% most-cited scientists worldwide by Stanford University in 2022, 2023, and 2024, both by Single Year and by Career Impact. His research interests include intelligent transport systems, the Internet of Things, wireless localization techniques, mobile communication systems, and applied graph theory and its applications in telecommunications.

Mr. Mao is a Fellow of AAIA and IET. He received the "Top Editor" Award for outstanding contributions to IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2011, 2014, and 2015. He has served as the chair, the co-chair, and a TPC member for a number of international conferences. From 2014 to 2017, he was the Co-Chair of the IEEE ITS Technical Committee on Communication Networks. He has been an Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS since 2018, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2014 to 2019, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2010 to 2020. He has been serving as the Vice-Director of the Smart Transportation Information Engineering Society, Chinese Institute of Electronics, since 2022.

Xiaojiang Ren (Member, IEEE) received the Ph.D. degree in computer science from the Australian National University, Canberra, ACT, Australia, in 2016.

He is an Associate Professor with Xidian University, Xi'an, China. His research interests include intelligent transport systems, the Internet of Things, wireless sensor networks, routing protocol design for wireless networks, and optimization problems.

Keyin Wang (Graduate Student Member, IEEE) received the master's degree in control engineering from Hubei University of Automotive Technology, Shiyan, China, in 2022. He is currently pursuing the Ph.D. degree with the College of Communication Engineering, Xidian University, Xi'an, China.

His research interests include vehicle localization based on multisource information fusion.

Wei Xiang (Senior Member, IEEE) is a Cisco Research Chair of AI and IoT, the Director and the Chief Scientist of the Australian Centre for AI and Medical Innovation, and the Director of the Cisco La Trobe Centre for AI and IoT, La Trobe University, Melbourne, VIC, Australia. Previously, he was the Foundation Chair and the Head of the Discipline of IoT Engineering, James Cook University, Cairns, QLD, Australia. For the past five consecutive years from 2020 to 2024, he has consistently ranked among Stanford University's World's Top 2% Scientists for both his single-year and career-long impacts. Due to his instrumental leadership in establishing Australia's first accredited Internet of Things Engineering degree program, he was inducted into Pearcy Foundation's Hall of Fame in October 2018. He has published over 450 peer-reviewed papers, including three books and nearly 300 journal articles. His research interests include the Internet of Things, wireless communications, machine learning for IoT data analytics, and computer vision.

Mr. Xiang is an elected Fellow of the IET in U.K. and Engineers Australia. He received the TNQ Innovation Award in 2016, the Pearcey Entrepreneurship Award in 2017, and the Engineers Australia Cairns Engineer of the Year in 2017. He was a co-recipient of four best paper awards from WiSATS 2019, WCSP 2015, IEEE WCNC 2011, and ICWMC 2009. He has been awarded several prestigious fellowship titles. He was named as a Queensland International Fellow (2010–2011) by Queensland Government of Australia, an Endeavor Research Fellow (2012–2013) by the Commonwealth Government of Australia, a Smart Futures Fellow (2012–2015) by Queensland Government of Australia, and a JSPS Invitational Fellow jointly by the Australian Academy of Science and Japanese Society for Promotion of Science (2014–2015). He was the Vice Chair of the IEEE Northern Australia Section from 2016 to 2020. He has served in a large number of international conferences in the capacity of the general co-chair, the TPC co-chair, and the symposium chair. He is an Associate Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE ACCESS, and *Scientific Reports* (Nature). He is a TEDx Speaker.

Xiaohu Ge (Senior Member, IEEE) received the Ph.D. degree in communication and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003.

He was a Researcher with Ajou University, Suwon, South Korea, and the Politecnico di Torino, Turin, Italy, from 2004 to 2005. He has been with HUST since 2005. He is currently a Full Professor with the School of Electronic Information and Communications, HUST. He is also an Adjunct Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He has published over 200 papers in refereed journals and conference proceedings, and has been granted about 15 patents in China. His research interests include mobile communications, traffic modeling in wireless networks, green communications, and interference modeling in wireless communications.

Dr. Ge received the best paper award from IEEE GLOBECOM 2010. He served as the General Chair for the 2015 IEEE International Conference on Green Computing and Communications in 2015. He serves as an Associate Editor for IEEE WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE ACCESS.