TRB Annual Meeting

Multi-View BEV Fusion for Enhanced 3D Object Detection in Vehicle-Infrastructure Cooperative Systems

--Manuscript Draft--

Full Title:	Multi-View BEV Fusion for Enhanced 3D Object Detection in Vehicle-Infrastructure Cooperative Systems
Abstract:	3D object detection plays a critical role in connected and autonomous vehicles (CAVs), as it enables precise localization and classification of surrounding obstacles by estimating their spatial properties from 2D images. This capability effectively addresses the limitation of traditional 2D detectors, which lack depth information. However, image-based 3D detection remains challenging in complex scenarios, particularly under occlusion or limited fields of view, such as at intersections or on curved roads. To address these issues, vehicle-to-infrastructure (V2I) cooperative perception frameworks leverage vehicle-to-everything (V2X) communication to transmit sensor data from roadside units to CAVs, enabling multi-view feature fusion from both vehicle-mounted and infrastructure-side cameras. In this paper, we propose V2IFormer, a novel V2I cooperative 3D object detection framework based on Bird's-Eye View (BEV) representations. Instead of transmitting raw images or high-level detection outputs, V2IFormer transmits BEV features to reduce bandwidth while preserving essential spatial information. On the infrastructure side, we introduce a HeightNet module with a linearly-increasing discretization (LID) strategy to predict adaptive height distributions, improving long-range depth perception. On the vehicle side, multi-view image features are lifted into the BEV space using depth distribution prediction, enhancing depth accuracy in close-range regions. A deformable mutual-attention module is further used to adaptively fuse BEV features from both sides, selectively focusing on informative regions while suppressing irrelevant content. Extensive experiments on the DAIR-V2X benchmark demonstrate that V2IFormer achieves state-of-the-art performance, particularly under challenging conditions with severe occlusion and limited visibility.
Additional Information:	
Question	**Response**
The total word count limit is 7500 words including tables. Each table equals 250 words and must be included in your count. Papers exceeding the word limit may be rejected. My word count is:	6932
Manuscript Classifications:	Traffic Operations and Management; Automated Vehicles; Connected Vehicles; ITS Systems; Self-Driving; V2X; Critical Infrastructure
Manuscript Number:	
Article Type:	Presentation and Publication
Order of Authors:	Guoqiang Mao
	Tianxuan Fu
	Xiaojiang Ren
	Keyin Wang

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript

Multi-View BEV Fusion for Enhanced 3D Object Detection in Vehicle-Infrastructure Cooperative Systems

Guoqiang Mao, ***Fellow, IEEE***
School of Transportation
Southeast University, Nanjing, China, 210096
Email: g.mao@ieee.org

Tianxuan Fu (Corresponding Author)
School of Telecommunications Engineering
Xidian University, Xi'an, China, 710071
Email: futianxuan@stu.xidian.edu.cn

Xiaojiang Ren
Guangzhou Research Institute
Xidian University, Xi'an, China, 710071
Email: xjren@xidian.edu.cn

Keyin Wang
School of Telecommunications Engineering
Xidian University, Xi'an, China, 710071
Email: keyinwang@stu.xidian.edu.cn

Word Count: 6182 words + 3 table(s) × 250 = 6932 words

Submission Date: July 30, 2025

Under Review

ABSTRACT

3D object detection plays a critical role in connected and autonomous vehicles (CAVs), as it enables precise localization and classification of surrounding obstacles by estimating their spatial properties from 2D images. This capability effectively addresses the limitation of traditional 2D detectors, which lack depth information. However, image-based 3D detection remains challenging in complex scenarios, particularly under occlusion or limited fields of view, such as at intersections or on curved roads. To address these issues, vehicle-to-infrastructure (V2I) cooperative perception frameworks leverage vehicle-to-everything (V2X) communication to transmit sensor data from roadside units to CAVs, enabling multi-view feature fusion from both vehicle-mounted and infrastructure-side cameras. In this paper, we propose V2IFormer, a novel V2I cooperative 3D object detection framework based on Bird's-Eye View (BEV) representations. Instead of transmitting raw images or high-level detection outputs, V2IFormer transmits BEV features to reduce bandwidth while preserving essential spatial information. On the infrastructure side, we introduce a HeightNet module with a linearly-increasing discretization (LID) strategy to predict adaptive height distributions, improving long-range depth perception. On the vehicle side, multi-view image features are lifted into the BEV space using depth distribution prediction, enhancing depth accuracy in close-range regions. A deformable mutual-attention module is further used to adaptively fuse BEV features from both sides, selectively focusing on informative regions while suppressing irrelevant content. Extensive experiments on the DAIR-V2X benchmark demonstrate that V2IFormer achieves state-of-the-art performance, particularly under challenging conditions with severe occlusion and limited visibility.

Keywords: Vehicle-infrastructure cooperative, Bird's-eye view, Connected and autonomous vehicles, 3D object detection

Under Review

INTRODUCTION

Accurate 3D object detection is a critical component in autonomous driving. It enables precise localization and classification of surrounding entities, including vehicles, pedestrians, and cyclists. Unlike 2D detection, which only identifies objects on a plane, 3D detection estimates their position, size, and orientation from 2D images—thus providing a more comprehensive understanding of the driving environment. However, monocular vehicle-mounted cameras face inherent limitations due to their low installation height, which leads to frequent occlusion and a restricted field of view. According to Waymo's safety report, occlusions account for 25% of major accidents involving autonomous vehicles (*1*), highlighting the need for advanced perception systems.

Roadside infrastructure, equipped with elevated cameras, offers a complementary perspective with a wider field of view, reducing occlusion and extending detection range. Existing 3D detection approaches can be broadly divided into LiDAR-based and camera-based methods. While LiDAR offers high geometric precision, its high cost and limited performance in detecting small or distant objects hinder large-scale deployment. In contrast, camera-based systems—especially monocular or multi-view configurations—offer a scalable and cost-effective solution, as evidenced by Tesla's vision-only Full Self-Driving pipeline (*2*). Vision-based roadside systems are preferred over LiDAR due to lower costs and seamless integration with urban infrastructure, such as traffic poles (*3, 4*). However, roadside monocular systems face inherent depth ambiguity, particularly over long distances. Recent methods like BEVHeight (*5*) and MonoGAE (*6*) improve depth estimation by leveraging ground-relative constraints, but still face challenges in detecting distant and occluded objects. Vehicle-to-Infrastructure (V2I) cooperative perception offers a promising solution by transmitting high-viewpoint sensor data to CAVs in real-time.

The intermediate fusion strategy integrates sensor data from both onboard vehicle sensors and roadside infrastructure into a unified bird's eye view (BEV) representation. This design enables extended perception coverage, mitigates occlusion effects, and ensures reliable performance under limited communication bandwidth. Owing to their ability to encode rich spatial context within a globally consistent coordinate system, BEV representations have become a widely adopted paradigm in camera-based autonomous driving systems. A fundamental challenge in BEV-based perception lies in transforming perspective-view image features into accurate and dense BEV features. Recent advances—such as *BEVDepth* (*7*) and *BEVFormer* (*8*)—address this by learning geometric projections and spatiotemporal associations, respectively. These frameworks have significantly advanced 3D scene understanding and have been successfully applied to core autonomous driving tasks, including 3D object detection (*9*), semantic segmentation (*10*), and map construction (*11, 12*). Additionally, BEV-based representations provide a common spatial foundation for V2I cooperative perception (*13, 14*). Despite these advancements, existing V2I perception methods still face notable limitations when applied to real-world multi-agent collaboration. For instance, *BEVFusion* (*7*) fuses multi-modal sensor inputs (e.g., camera and LiDAR) within the BEV space and achieves strong detection performance. However, its substantial computational cost and bandwidth demand hinder its suitability for distributed V2I scenarios. Meanwhile, *ImVoxelNet* (*15*), a monocular-based 3D detection model, lifts 2D features into voxel space using predicted depth. While computationally efficient, its performance deteriorates in long-range detection and under occlusions due to limited depth precision. Other cooperative frameworks, such as *V2X-ViT* (*16*) and *VIMI* (*14*), adopt attention-based or concatenation-based fusion mechanisms. Although these strategies offer initial improvements, they struggle to resolve the spatial and semantic inconsistencies inherent in cross-view data. Global attention mechanisms, such as

Under Review

cross-attention, assign equal importance to all spatial locations, leading to inefficiencies in filtering irrelevant information and amplifying sensor noise. Similarly, naive feature concatenation lacks the capacity to adaptively align heterogeneous inputs, resulting in poor geometric consistency and suboptimal fusion.

V2I cooperative perception has emerged as a promising paradigm for enhancing environmental awareness. By enabling the real-time sharing of high-elevation sensor data from infrastructure to CAVs, V2I systems can significantly extend perception capabilities beyond the inherent limitations of onboard sensors. Enabled by 5G and V2X communication technologies (*17*), such systems support multi-view data fusion to improve situational awareness. However, a fundamental trade-off exists between detection accuracy and communication bandwidth. Transmitting raw sensor data retains rich detail but demands high bandwidth; transmitting only detection results reduces bandwidth but sacrifices spatial fidelity (*13, 14, 16*). To balance this trade-off, many recent methods adopt intermediate feature representations for transmission—particularly BEV features, which provide a unified spatial framework for multi-sensor fusion. BEV-based approaches offer strong spatial context and support efficient fusion of vehicle-side and infrastructure-side features. Nevertheless, transforming image-view features into accurate BEV features remains a major challenge, especially under monocular setups.

To address the challenges of limited bandwidth and multi-source feature fusion in V2I cooperative 3D object detection, we propose V2IFormer, a novel framework based on BEV representations. Instead of transmitting raw images or high-level predictions, V2IFormer transmits intermediate BEV features to reduce bandwidth while preserving spatial detail. On the infrastructure side, we introduce a HeightNet module with a linearly-increasing discretization (LID) strategy to predict adaptive height distributions and improve long-range depth perception. On the vehicle side, image features from multiple views are lifted into the BEV space using depth distribution prediction based on the LSS framework (*18*), which enhances accurate depth perception at close range. A deformable mutual-attention module then adaptively aligns and fuses dual-view BEV features by focusing on informative regions and suppressing irrelevant noise, ensuring robustness under occlusion and varying viewpoints. Experiments on the DAIR-V2X benchmark demonstrate that V2IFormer achieves state-of-the-art performance in complex driving scenarios. The main contributions of this paper are summarized as follows:

1. We propose V2IFormer, a novel V2I cooperative 3D object detection framework that transmits BEV features to balance perception accuracy and bandwidth efficiency. By integrating vehicle-mounted and infrastructure-side monocular data, it enhances robustness under occlusion and limited visibility.
2. Specialized BEV generation modules are designed on both the infrastructure and vehicle sides. On the infrastructure side, we introduce a HeightNet module with a linearly-increasing discretization (LID) strategy to predict adaptive height distributions, enhancing long-range depth perception. On the vehicle side, image features from multiple views are lifted into BEV space using depth distribution prediction based on the LSS framework, which improves depth accuracy in close-range regions.
3. A deformable mutual-attention module is employed to adaptively fuse dual-source BEV features by selectively focusing on informative regions while suppressing irrelevant noise. The effectiveness of V2IFormer is validated on the DAIR-V2X benchmark, where it achieves state-of-the-art 3D detection performance in complex driving scenarios.

Under Review

The structure of the paper is as follows. Section II surveys the state of the art in 3D object detection, collaborative perception, and BEV scene understanding, outlining recent advancements and identifying current limitations. Section III details the proposed V2IFormer architecture, which features dual-side BEV generation and a deformable mutual-attention fusion mechanism. Section IV reports the experimental evaluation on the DAIR-V2X benchmark, highlighting the method's accuracy and robustness. Finally, Section V summarizes the findings and explores directions for future research.

LITERATURE REVIEW

Camera-Based 3D Object Detection

Current 3D object detection techniques are generally classified into two main categories based on the type of sensing modality: image-based methods and LiDAR-based point cloud methods. While LiDAR offers high detection accuracy and precise spatial measurements, its high cost, bulky hardware, and limited effectiveness in detecting small or distant objects hinder its widespread deployment in real-world applications. In contrast, camera-based 3D object detection estimates 3D bounding boxes using image data from monocular cameras. This approach is increasingly favored over LiDAR for several compelling reasons. First, cameras are significantly more cost-effective and widely available, making them suitable for large-scale deployment in both consumer vehicles and roadside infrastructure (*19, 20*). For example, Tesla's FSD system demonstrates the viability of a vision-only solution by achieving advanced driver assistance capabilities using a suite of eight cameras (*2*). Second, cameras capture rich visual cues—including color, texture, and contextual information—that are essential for accurate object classification and comprehensive scene understanding, especially in complex environments (*8*). Third, recent advances in deep learning, particularly in monocular depth estimation and neural network architectures, have significantly improved the performance of image-based 3D detection systems. Tesla's FSD performance in urban scenarios with minimal human intervention highlights this progress (*21*). Despite challenges such as depth ambiguity, occlusion, and sensitivity to environmental conditions, camera-based 3D object detection strikes a favorable balance between cost, scalability, semantic richness, and system integration feasibility. These advantages make it a promising solution not only for autonomous driving but also for infrastructure-assisted perception and broader intelligent transportation applications.

Camera-based 3D object detection methods can be categorized into vehicle-mounted and infrastructure-based approaches. Early works extend 2D detectors to jointly predict both 2D and 3D object attributes. Notable examples include FCOS3D (*22*), DETR3D (*23*), and the PETR series methods (*24, 25*), which incorporate geometric priors or temporal context to enhance spatial reasoning. Alternatively, some methods operate directly in 3D space by lifting 2D image features into structured 3D representations. For instance, OFT (*26*) introduces orthographic feature transforms, while LSS (*18*) predicts depth distributions to generate BEV features. BEVDepth (*7*) leverages LiDAR supervision to refine depth prediction, and CrossDTR (*27*) improves spatial understanding through depth-aware embeddings and transformers. However, these methods are constrained by occlusions and limited fields of view inherent to onboard sensors.

Compared to vehicle-mounted systems, roadside cameras provide elevated viewpoints and broader scene coverage, effectively compensating for the limited field of view and occlusion issues inherent to onboard sensors. Despite these advantages, infrastructure-based detection faces challenges such as depth ambiguity, reduced accuracy for distant objects, and persistent occlusions.

Under Review

To mitigate these issues, BEVHeight (*5*) predicts height distributions to enhance depth estimation from monocular inputs. CBR (*28*) removes extrinsic calibration dependence by using MLP-based feature transformation but at the cost of geometric precision. Furthermore, public benchmarks such as DAIR-V2X (*13*) and Rope3D (*29*) have been developed to evaluate the effectiveness of roadside perception systems.

Cooperative Perception

V2I cooperative perception enhances situational awareness by enabling information sharing between vehicles and infrastructure. It helps mitigate occlusions and extends the perception range. Based on the data format used for communication, fusion strategies can be classified into early, late, and intermediate fusion. Early fusion transmits raw sensor data, such as camera images, which preserves detailed information for aggregation but requires high communication bandwidth. Late fusion transmits only final detection outputs like 3D bounding boxes, reducing bandwidth but sacrificing contextual richness. For example, Cooper (*30*) uses early fusion to broadcast LiDAR data, which incurs substantial communication costs. In contrast, late fusion methods (*13*) transmit predictions from each CAV, but their performance suffers from the lack of shared context and reliance on individual detection quality.

Intermediate fusion offers a better trade-off by transmitting compact feature maps, such as BEV representations. These features retain spatial information while significantly lowering communication overhead. Typically, roadside units encode raw sensor inputs into intermediate features using local neural networks and transmit them to CAVs via low-latency V2X links. CAVs can then receive feature maps from multiple viewpoints, increasing spatial diversity. Effective fusion of these features is essential. Aggregation methods like feature-wise maximum (*31*) and summation (*32*) are commonly applied. Attention-based approaches, such as AttFuse (*4*), V2X-ViT (*16*), and CoBEVT (*33*), leverage transformer architectures for better alignment. Inspired by these, we propose an intermediate BEV fusion framework that employs deformable mutual-attention to adaptively combine vehicle and infrastructure features.

BEV Scene Understanding

BEV perception plays a key role in spatial reasoning by projecting sensor information into a unified top-down space. The core challenge lies in lifting perspective-view image features to the BEV domain. Existing methods fall into two categories: implicit and explicit lifting.

Implicit lifting methods learn image-to-BEV mappings using neural networks without enforcing geometric constraints. MLP-based methods, such as VPN, convert image features into BEV space for downstream tasks like segmentation and road layout estimation (*34*). Transformer-based approaches, such as PETR (*24*) and BEVFormer (*8*), use cross-attention between BEV queries and image features to learn spatial correspondence. BEVFormer further incorporates temporal fusion for improved temporal consistency. However, these methods often neglect camera intrinsics and extrinsics, which are crucial for precise spatial alignment.

Explicit lifting methods rely on camera geometry and estimated depth to project features accurately into BEV. Classical methods like IPM (*35*) assume a flat ground and fixed intrinsics, but are prone to sampling artifacts. To overcome the ill-posed nature of monocular depth recovery, Pseudo-LiDAR (*36*) reconstructs 3D point clouds from estimated depth. More recent models, such as LSS (*18*) and BEVDepth (*7*), predict dense depth distributions and use calibrated camera parameters for BEV transformation. BEVHeight (*5*) extends this by predicting height distributions

Under Review

rather than absolute depth, improving localization under uncertainty.

METHODOLOGY

V2IFormer leverages V2X communication to integrate perception from both vehicles and infrastructure. It consists of three key modules: an infrastructure-side BEV branch, a vehicle-side BEV branch, and a fusion module that combines the BEV features from both sources, as illustrated in Figure 1.

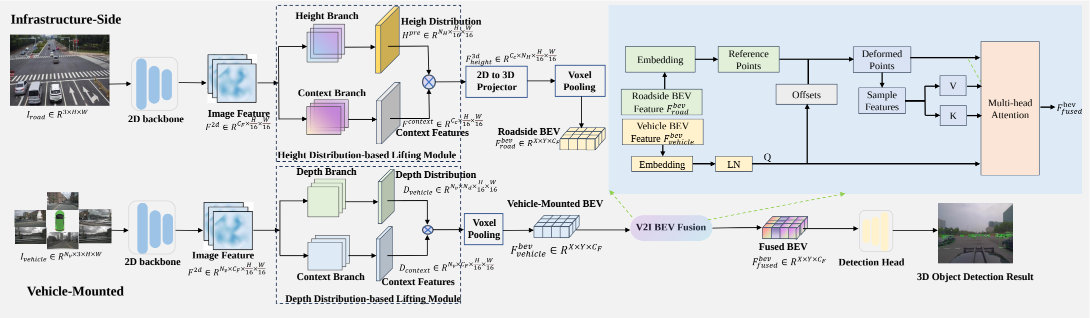
Infrastructure-Side
$I_{road} \in R^{3\times H\times W}$
2D backbone
Image Feature
$F^{2d} \in R^{C_F\times\frac{H}{16}\times\frac{W}{16}}$
Height Branch
Heigh Distribution
$H^{pre} \in R^{N_H\times\frac{H}{16}\times\frac{W}{16}}$
Context Branch
$F^{context} \in R^{C_c\times\frac{H}{16}\times\frac{W}{16}}$
Context Features
Height Distribution-based Lifting Module
$F^{3d}_{height} \in R^{C_c\times N_H\times\frac{H}{16}\times\frac{W}{16}}$
2D to 3D Projector
Voxel Pooling
Roadside BEV
$F^{bev}_{road} \in R^{X\times Y\times C_F}$
Embedding
Reference Points
Deformed Points
Roadside BEV Feature F^{bev}_{road}
Offsets
Sample Features
V
K
Multi-head Attention
F^{bev}_{fused}
Vehicle BEV Feature $F^{bev}_{vehicle}$
Embedding
LN
Q
$I_{vehicle} \in R^{N_v\times 3\times H\times W}$
2D backbone
Image Feature
$F^{2d} \in R^{N_v\times C_F\times\frac{H}{16}\times\frac{W}{16}}$
Depth Branch
Depth Distribution
$D_{vehicle} \in R^{N_v\times N_d\times\frac{H}{16}\times\frac{W}{16}}$
Context Branch
$D_{context} \in R^{N_v\times C_F\times\frac{H}{16}\times\frac{W}{16}}$
Context Features
Depth Distribution-based Lifting Module
Vehicle-Mounted
Voxel Pooling
Vehicle-Mounted BEV
$F^{bev}_{vehicle} \in R^{X\times Y\times C_F}$
V2I BEV Fusion
Fused BEV
$F^{bev}_{fused} \in R^{X\times Y\times C_F}$
Detection Head
3D Object Detection Result

Figure 1: The overall framework of V2IFormer. The proposed framework comprises three key components: the infrastructure-side branch, the vehicle-mounted branch, and the BEV feature fusion module. On the infrastructure side, a HeightNet module with a LID strategy predicts adaptive height distributions. These predictions are used to lift image-view features into height-based BEV features, enhancing long-range depth perception. On the vehicle-mounted side, image features from multiple views are lifted into BEV space using depth distribution prediction based on the LSS framework, which improves depth estimation accuracy in close-range regions. Finally, the BEV feature fusion module employs deformable mutual-attention to adaptively combine BEV features from both the infrastructure and vehicle branches. The resulting fused BEV features are then fed into the detection head for final 3D object prediction.

Under Review

Infrastructure-Side Branch

In monocular 3D perception, depth prediction is a widely used approach for constructing BEV representations. At close range, objects occupy a relatively large number of pixels, providing rich visual cues such as texture and clear contours, which facilitate accurate depth estimation. In such conditions, depth-based methods can achieve reliable 3D reconstruction by exploiting strong geometric priors. However, in long-range scenarios, the effectiveness of depth prediction declines significantly. Distant objects appear much smaller in the image—often covering only a few pixels—resulting in insufficient spatial evidence for reliable depth inference. This leads to unstable or biased predictions, limiting the applicability of depth-based methods in V2I settings that require long-range and occlusion-resilient perception.

To address this, inspired by height-based modeling strategies such as BEVHeight (*5*), 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. In structured traffic environments, common object categories—such as vehicles and pedestrians—exhibit consistent height patterns, making height a more robust cue for 3D reasoning. Furthermore, infrastructure-mounted cameras—due to their elevated viewpoint—can better leverage height information to infer object location in 3D space with reduced ambiguity. This makes height-based modeling particularly suitable for roadside perception tasks that emphasize long-range detection and occlusion resilience. Based on these insights, we introduce the *HeightNet* module, a roadside BEV feature extraction module centered on height prediction. *HeightNet* employs a LID strategy (*37*) to model height distributions, enabling precise localization of distant targets and enhancing the geometric consistency of BEV features.

HeightNet

To enable robust and geometry-aware BEV feature extraction from roadside monocular images, we introduce *HeightNet*, a dedicated height prediction network that estimates pixel-wise height distributions from 2D image features. Unlike depth, object height exhibits greater invariance across distances and is less sensitive to scale distortions, making it a more reliable geometric cue for long-range perception.

As illustrated in Figure 1, HeightNet consists of two main components: a context branch and a height estimation branch. The context branch extracts semantic features from the input 2D image features using a series of residual blocks (*38*) and a channel attention module. These features are then refined by a deformable convolution (DCN) layer (*39*), which enhances spatial adaptability and focuses on object-relevant regions. In parallel, the camera's intrinsic parameters (I), such as focal length and principal point, and extrinsic parameters (E), including rotation and translation vectors that define the camera's position and orientation in 3D space, are encoded through two MLPs to generate camera-aware features F_{cam}. These features are then fused with the contextual features to improve the network's adaptability across different viewpoints and camera configurations.

Height prediction is formulated as a classification task by discretizing the continuous height range into a predefined number of bins. To ensure accurate and consistent modeling across varying distances, we adopt LID (*37*) to partition the height interval. LID increases the bin width linearly with height, which balances sampling resolution across short and long ranges. LID yields more consistent relative error across the full height spectrum, offering both precision for nearby objects and robustness for distant targets. This design choice significantly enhances the quality of height-aware BEV features, particularly under long-range and occluded conditions, which are prevalent in V2I scenarios.

Formally, the discretized bin center h_i in LID is defined as:

$$h_i = h_{\min} + (h_{\max} - h_{\min}) \cdot \frac{i(i+1)}{N_{\text{H}}(N_{\text{H}}+1)} \tag{1}$$

Where N_{H} is the total number of bins, and $[h_{\min}, h_{\max}]$ defines the target height interval. This formulation ensures that the discretization adapts smoothly to the distribution of object heights in real-world traffic scenes.

The height prediction module, as illustrated in Figure 1, is designed to estimate pixel-wise height distributions from image features. The network architecture begins with a set of residual blocks (*38*), which process the input 2D image features $\mathbf{F}^{2d}$. To enhance geometric adaptability, these features are conditioned on camera parameters $\mathbf{F}_{\text{cam}}$, which encode both intrinsic and

Under Review

extrinsic calibration data. The features are then refined using a deformable convolution (DCN) layer (*39*), which dynamically adjusts the sampling positions to better capture object boundaries and spatial context. Finally, the processed features are passed through a prediction head $\psi_h(\cdot)$ to output per-pixel height distributions. The entire height prediction process can be formulated as:

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

Where $\text{Res}(\cdot)$ denotes the residual block operations, $\text{DCN}(\cdot)$ represents the deformable convolution, and $\psi_h(\cdot)$ is the prediction network. The output $\mathbf{H}^{\text{pre}} \in \mathbb{R}^{N_H \times \frac{H}{16} \times \frac{W}{16}}$ represents the predicted discretized height distribution over N_H height bins, at a spatial resolution reduced by a factor of 16 relative to the input image.

Height-based 2D-to-3D Feature Projection

To construct a 3D representation from 2D image features, we leverage the predicted height distribution $\mathbf{H}^{\text{pre}}$ to lift pixel-level features into a frustum-shaped 3D space. This process enables spatial reasoning in the height dimension, which is particularly useful for generating BEV representations from monocular images.

We first extract semantic contextual features $\mathbf{F}^{\text{context}}$ from the 2D input features $\mathbf{F}^{2d}$ using a series of convolutional operations and channel attention mechanisms. These context features are then combined with the height distribution $\mathbf{H}^{\text{pre}}$ through an outer product operation:

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

Where $\otimes$ denotes the outer product across the channel and height dimensions. This results in a volumetric feature tensor $\mathbf{F}^{3d}_{\text{height}} \in \mathbb{R}^{C_c \times N_H \times \frac{H}{16} \times \frac{W}{16}}$, where C_c is the number of channels and N_H is the number of discretized height bins.

To enable downstream BEV-based processing, we project the frustum-shaped feature volume into the ego-vehicle coordinate system through a three-stage geometric transformation pipeline, as illustrated in Figure 2.

Step 1: Image Pixel to Camera Coordinates

For each image pixel $\mathbf{p}_{\text{imag}} = (u, v)$, we define a reference point $\mathbf{P}^{\text{cam}}_{\text{ref}}$ at unit depth (i.e., depth = 1) in the camera coordinate system. Using the camera intrinsic matrix $\mathbf{I}$, the mapping is computed as:

$$\mathbf{P}^{\text{cam}}_{\text{ref}} = \mathbf{I}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{4}$$

Where $\mathbf{I}$ is the intrinsic matrix:

$$\mathbf{I} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Here, f_x and f_y are the focal lengths in the horizontal and vertical directions, respectively, and (c_x, c_y) denotes the principal point (optical center) of the image.

Under Review

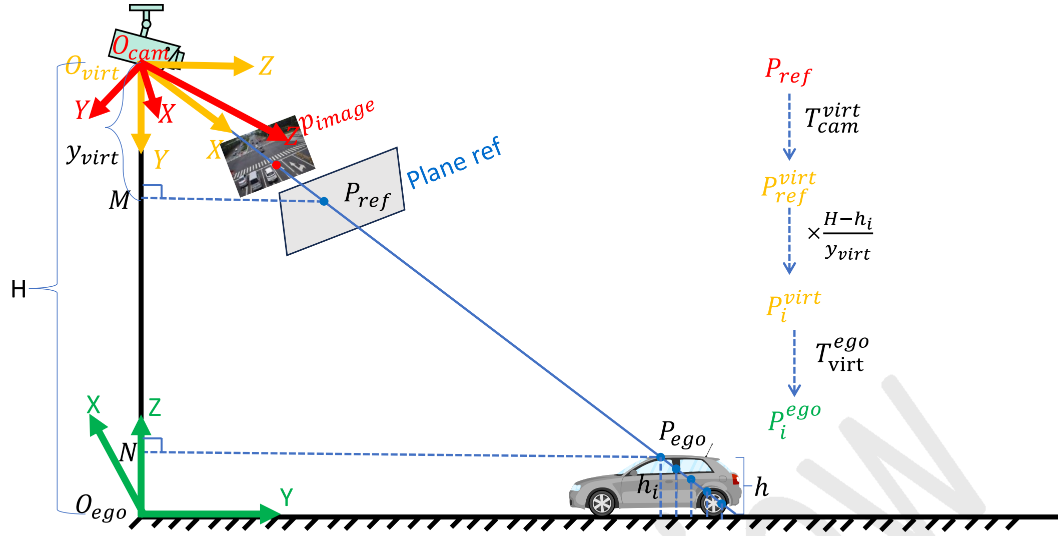

Figure 2: Height-based 2D–3D projector. $\mathscr{O}_{\text{ego}}XYZ$ denotes the ego-vehicle coordinate system. $\mathscr{O}_{\text{virt}}$ shares the same origin as $\mathscr{O}_{\text{cam}}$, but its Y-axis is perpendicular to the ground. A pixel in the image plane, with an estimated height h_i from the predicted distribution, is projected through this 2D-to-3D process into a 3D point $\mathbf{P}_{\text{ego}}$ in the ego coordinate system.

Step 2: Camera to Virtual Coordinate System

The reference point is transformed into the virtual coordinate system via:

$$\mathbf{P}^{\text{virt}}_{\text{ref}} = \mathbf{T}^{\text{virt}}_{\text{cam}} \mathbf{P}^{\text{cam}}_{\text{ref}} \tag{6}$$

Where $\mathbf{T}^{\text{virt}}_{\text{cam}}$ encodes the extrinsic transformation (rotation and translation) from the camera to the virtual coordinate system. The virtual coordinate frame is defined such that the origin lies at the camera's optical center, with the Y-axis pointing downward (toward the ground), the Z-axis pointing forward, and the X-axis to the right.

Step 3: Virtual to Ego Coordinates

For each height bin h_i, we compute the corresponding 3D point in the ego vehicle coordinate system using the principle of similar triangles:

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

Where $\mathbf{T}^{\text{ego}}_{\text{virt}}$ transforms coordinates from the virtual frame to the ego frame, H is the known camera height above the ground, and y_{virt} is the vertical coordinate of $\mathbf{P}^{\text{virt}}_{\text{ref}}$. The factor $\frac{H-h_i}{y_{\text{virt}}}$ scales the reference point to the appropriate height slice. This process assigns each pixel–height pair (u, v, h_i) to a unique 3D position in the ego coordinate system. The full 3D volume is then projected onto the BEV plane by applying sum-pooling along the height dimension N_H. The result is a compact 2D BEV feature map that retains semantic and geometric context, and is well-suited for downstream tasks such as object detection and multi-view feature fusion.

Under Review

Vehicle-Mounted BEV Features Extraction

In contrast to the infrastructure-side pipeline, which typically transforms 2D features into BEV representations using height distributions, the vehicle-mounted perception module independently generates BEV features by transforming image data captured from multiple onboard cameras. These features are projected into a unified top-down BEV space centered on the ego vehicle, enabling spatially consistent reasoning within the vehicle's surrounding environment.

The BEV space is discretized into a fixed-resolution 2D grid of size $X \times Y$, representing a BEV of the environment. Each grid cell corresponds to a fixed physical area of $d \times d$ meters on the ground plane (e.g., $d = 0.5\,\mathrm{m}$). To populate this grid with meaningful semantic features, raw image data must be lifted from the 2D image plane into 3D space using depth information, and then reprojected into the BEV plane based on their real-world positions. Given the absence of direct depth sensing in monocular inputs, we adopt the LSS paradigm (*18*), which estimates depth distributions for each pixel and uses them to construct a 3D volumetric representation before collapsing it into a 2D BEV feature map.

Let $I_{\text{vehicle}} \in \mathbb{R}^{N_v \times 3 \times H \times W}$ represent the multi-view RGB images captured by N_v vehicle-mounted cameras, each with spatial resolution $H \times W$. These images are first passed through a shared 2D convolutional backbone to extract visual features, followed by a feature pyramid network (FPN) to generate multi-scale representations. The output is then downsampled to $\frac{1}{16}$ of the original resolution, producing intermediate feature maps for each view.

This output is further processed through two parallel branches:

- **Depth Estimation Branch:** Outputs per-pixel discrete depth distributions $D_{\text{vehicle}} \in \mathbb{R}^{N_v \times N_d \times \frac{H}{16} \times \frac{W}{16}}$, where $N_d = \frac{r}{\Delta r}$ is the number of depth bins covering the range $[1, r]$ meters (typically $r = 100\,\mathrm{m}$, $\Delta r = 1\,\mathrm{m}$), modeling the probability of each pixel belonging to different depths along its viewing ray.
- **Context Feature Branch:** Outputs semantic features $D_{\text{context}} \in \mathbb{R}^{N_v \times C_F \times \frac{H}{16} \times \frac{W}{16}}$, where C_F is the channel dimension of the feature space.

For each pixel, N_d candidate 3D points are sampled along its camera ray by applying inverse projection using the known camera intrinsics and extrinsics. Each 3D point is assigned a semantic feature by weighting the corresponding context feature with the associated depth probability, resulting in a set of probabilistically weighted 3D point features across all camera views.

These 3D features are then projected onto the ground plane and mapped to the corresponding BEV grid cells. To aggregate information from multiple overlapping points that fall into the same grid cell, a pooling operation is applied, producing the final BEV feature map $F_{\text{vehicle}}^{\text{bev}} \in \mathbb{R}^{X \times Y \times C_F}$. This representation encodes high-level semantic and geometric information aligned in a globally consistent spatial frame centered on the ego vehicle. Importantly, the BEV features generated from vehicle-mounted cameras can be seamlessly integrated with those produced by the infrastructure-side pipeline. This fusion enables complementary advantages: vehicle-side sensors provide dense near-field coverage, while infrastructure-side sensors offer broader and more stable field-of-view, especially beneficial in complex occlusion scenarios or when cooperative perception is required.

Vehicle-Infrastructure BEV Feature Fusion

The BEV features $\mathbf{F}_{\text{road}}^{\text{bev}}$ and $\mathbf{F}_{\text{vehicle}}^{\text{bev}}$ exhibit fundamentally different error characteristics due to their observation perspectives. Simple feature stacking fails to align features from the same spatial

Under Review

region accurately. As depicted in Figure 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. This approach, inspired by deformable mutual-attention in (*40*), achieves efficient and precise feature integration.

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

$$\mathbf{Q}_p = \text{conv}_{1\times 1}\left(\left[\mathbf{F}_{\text{road}}^{\text{bev}};\mathbf{F}_{\text{vehicle}}^{\text{bev}}\right]\right) \tag{8}$$

Where $\mathbf{Q}_p$ 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:

$$\text{DeformAttn}(\mathbf{Q}_p, p, \mathbf{F}_{\text{road}}^{\text{bev}}) = \sum_{m=1}^{M} \mathbf{W}_m \left[\sum_{k=1}^{K} A_{mqk}\cdot \mathbf{W}'_m \mathbf{F}_{\text{road}}^{\text{bev}}(p+\Delta p_{mqk})\right] \tag{9}$$

Where M is the number of attention heads, m indexes the attention head, K is the number of sampled keys, and k indexes each key. $A_{mqk}\in[0,1]$ is the attention weight, and $\sum_{k=1}^{K} A_{mqk}=1$. Δp_{mqk} denotes the learned offset from the reference point p, and $\mathbf{F}_{\text{road}}^{\text{bev}}(p+\Delta p_{mqk})$ is the feature at the sampled location. $\mathbf{W}_m$ aggregates the multi-head outputs and $\mathbf{W}'_m$ is the input projection weight.

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

$$\mathbf{F}_{\text{fused}}^{\text{bev}} = \text{BevFuse}(\mathbf{F}_{\text{road}}^{\text{bev}}, \mathbf{F}_{\text{vehicle}}^{\text{bev}}) = \sum_{p=0}^{HW}\left[\mathbf{Q}'_p + \sum_{V\in\{\mathbf{F}_{\text{road}}^{\text{bev}},\mathbf{F}_{\text{vehicle}}^{\text{bev}}\}} \text{DeformAttn}(\mathbf{Q}_p, p, V)\right] \tag{10}$$

Where H and W are the height and width of the BEV map, and $\mathbf{Q}'_p$ is the sampled query feature at position p from either the roadside or vehicle-mounted BEV branch.

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.

EXPERIMENTAL EVALUATION

We evaluate the proposed V2I collaborative perception framework on the DAIR-V2X dataset. We first outline the experimental setup and key characteristics of the dataset. We then conduct a comparative analysis between V2IFormer and representative state-of-the-art methods.

Dataset

Most existing V2I perception studies rely on simulated datasets with idealized V2X communication, limiting their applicability in real-world scenarios. In contrast, we conduct our evaluation on

Under Review

the *DAIR-V2X* dataset (*13*), which captures real-world data. Released in 2022 by Baidu Apollo and the Institute for AI Industry Research (AIR) at Tsinghua University, *DAIR-V2X* is the first publicly available benchmark designed for vehicle-infrastructure cooperative autonomous driving.

The dataset covers 10 km of urban roads, 10 km of highways, and 28 intersections within Beijing's High-Level Autonomous Driving Demonstration Zone. It encompasses diverse traffic scenarios under varying weather conditions (sunny, rainy, foggy) and lighting environments (day and night), offering comprehensive support for perception research. *DAIR-V2X* provides rich multi-modal data, including synchronized camera images, LiDAR point clouds, timestamped annotations, and calibration information. Specifically, the *DAIR-V2X-C* subset contains 38,845 annotated frames for cooperative detection. Our experiments use the *VIC-Sync* portion of *DAIR-V2X-C*, comprising 9,311 vehicle-infrastructure synchronized frame pairs. Annotations, originally given in world coordinates, are converted into vehicle coordinates for evaluation. We follow the official data split—4,822 training frames, 17,955 validation frames, and 2,694 test frames—and adopt the KITTI-style average precision metric for 3D object detection (*19*).

Evaluation Metrics

We adopt the official evaluation metrics defined by the *DAIR-V2X* dataset (*13*) to assess both detection accuracy and communication efficiency. Specifically, Average Precision (AP) (*20*) is used to evaluate 3D object detection performance, while Average Byte (AB) quantifies transmission cost. For 3D detection, AP is computed by comparing predicted results with ground-truth annotations provided in the *DAIR-V2X-C* subset. Following the VIC3D evaluation protocol, we evaluate objects from the vehicle's egocentric perspective, considering those located within a predefined 3D region in the vehicle coordinate frame: $x \in [x_{\min}, x_{\max}]$ m, $y \in [y_{\min}, y_{\max}]$ m, and $z \in [z_{\min}, z_{\max}]$ m. Detection performance is reported over the full 0–100 m range using a 3D Intersection-over-Union (IoU) threshold of 0.5. Two components are evaluated: AP_{3D} for full 3D bounding boxes and AP_{BEV} for BEV projections. In parallel, AB captures the average size (in bytes) of transmitted data, enabling an assessment of perception quality under communication constraints.

Implementation Specifics

We conduct comparative experiments between the proposed *V2IFormer* and several state-of-the-art baselines. For late fusion settings, we adopt *ImVoxelNet* (*15*) as the monocular detector applied to each view independently. For early fusion, *PointPillars* (*41*) serves as the LiDAR-based detection method. On the vehicle side, we benchmark multi-view camera fusion techniques, including *BEVFormer* (*8*), which projects image features into BEV space, and *VIMI* (*14*), a recent multi-view fusion approach. On the infrastructure side, *BEVDepth* (*7*) is used to aggregate multi-camera inputs from elevated perspectives. For fair comparison, all models use a ResNet-101 backbone for image encoding. Input images are resized to 864 × 1536 pixels during training. We follow the data augmentation strategy of BEVDepth, applying random cropping, scaling, flipping, and rotation to images, and random scaling, flipping, and rotation to BEV representations. All models are trained for 150 epochs using the AdamW optimizer (*42*), with an initial learning rate of 2×10^{-4}, distributed across eight NVIDIA RTX 3090 GPUs.

Quantitative Results

Quantitative evaluation results are reported in Table 1 and Table 2, covering both 3D object detection and BEV-based detection across multiple range intervals. Overall, the proposed *V2IFormer*

Under Review

Under Review

outperforms all baseline methods in terms of both detection accuracy and range robustness. Specifically, *V2IFormer* achieves an overall 3D detection performance of 55.64% in AP_{3D} and 59.63% in AP_{BEV}. Compared to the representative late fusion method *ImVoxelNet*, which yields 26.56% AP_{3D} and 31.40% AP_{BEV}, *V2IFormer* improves accuracy by 29.08 and 28.23 percentage points, respectively. Against the early fusion method *PointPillars*, which achieves 50.03% AP_{3D} and 53.73% AP_{BEV}, *V2IFormer* still gains an additional 5.61 and 5.90 points in 3D and BEV detection, respectively, showing its superiority even over strong baselines with direct access to LiDAR data.

The advantages of *V2IFormer* become even more evident under long-range detection (50–100 m), where accurate perception is typically more difficult due to reduced resolution, depth ambiguity, and increased occlusion. In this range, *V2IFormer* achieves 35.50% AP_{3D} and 44.80% AP_{BEV}. These values exceed those of *PointPillars* (33.05% / 36.17%) and *ImVoxelNet* (9.81% / 12.99%) by significant margins. This suggests that the intermediate fusion strategy not only effectively aggregates cross-view information but also enhances spatial consistency, especially in sparsely observed or occluded areas. Notably, *PointPillars* exhibits relatively strong performance in the 30–50 m mid-range interval (60.38% AP_{3D} and 64.08% AP_{BEV}), slightly outperforming *V2IFormer* in this specific range (59.06% / 63.04%). However, this advantage comes at the cost of higher bandwidth usage and performance drop at long range. In contrast, *V2IFormer* maintains consistent and well-balanced detection accuracy across all three intervals (short, mid, and long), demonstrating its robustness and adaptability to varying spatial configurations.

Table 1: 3D detection performance (AP_{3D}, IoU=0.5) on the DAIR-V2X-C dataset.

Fusion	**Model**	Overall	0–30	30–50	50–100
Only-Veh	VIMI (*14*)	8.66	19.11	4.33	0.20
Only-Inf	BEVFormer (*8*)	8.80	18.07	3.71	1.76
Only-Inf	BEVDepth (*7*)	7.36	16.23	1.79	0.18
Late Fusion	ImVoxelNet (*15*)	26.56	34.20	17.20	9.81
Early Fusion	PointPillars (*41*)	50.03	53.07	60.38	33.05
Intermediate-Fusion	V2IFormer	**55.64**	**72.36**	**59.06**	**35.50**

Table 2: BEV detection performance (AP_{BEV}, IoU=0.5) on the DAIR-V2X-C dataset.

Fusion	**Model**	Overall	0–30	30–50	50–100
Only-Veh	VIMI (*14*)	10.46	22.42	5.57	0.42
Only-Inf	BEVFormer (*8*)	13.45	24.76	6.46	4.63
Only-Inf	BEVDepth (*7*)	13.17	26.42	5.00	4.82
Late Fusion	ImVoxelNet (*15*)	31.40	37.75	21.21	12.99
Early Fusion	PointPillars (*41*)	53.73	55.80	64.08	36.17
Intermediate-Fusion	V2IFormer	**59.63**	**71.05**	**63.04**	**44.80**

Communication overhead is a critical constraint for real-time V2X cooperative perception systems, especially in practical deployments with limited wireless bandwidth. Table 3 presents the average communication bandwidth (AB) required by each method, revealing a clear efficiency-accuracy trade-off across fusion strategies. *V2IFormer* requires only 146.24 KB of data transmission per frame, which is an order of magnitude lower than the early fusion baseline *PointPillars* (1382.28 KB). Despite this reduction, *V2IFormer* not only retains but exceeds the detection performance of *PointPillars*, indicating its ability to minimize redundant information transfer by focusing on compact, semantically meaningful features. Compared with the late fusion approach *ImVoxelNet*, which transmits as little as 102.32 bytes per frame by independently processing monocular inputs, *V2IFormer* consumes moderately more bandwidth. However, this slight increase in data exchange yields substantial performance gains—+29.08% in AP_{3D} and +28.23% in AP_{BEV}—demonstrating that the cost is well-justified.

This evaluation underscores the efficiency of the proposed intermediate fusion strategy. It strikes a favorable balance between detection accuracy and communication cost, achieving near-optimal performance without incurring the excessive transmission burden typical of early fusion methods. Such a characteristic makes *V2IFormer* particularly suitable for scalable deployment in bandwidth-constrained vehicular environments, where minimizing latency and data traffic is essential.

Table 3: Average communication bandwidth (AB) required for each method.

Fusion	Model	AB (Byte)
Only-Veh	VIMI (*14*)	0
Only-Inf	BEVFormer (*8*)	0
Only-Inf	BEVDepth (*7*)	0
Late Fusion	ImVoxelNet (*15*)	102.32
Early Fusion	PointPillars (*41*)	1382.28K
Intermediate-Fusion	V2IFormer	**146.24K**

Under Review

We conduct a comprehensive evaluation 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. Among all compared methods, the proposed intermediate fusion approach, V2IFormer, consistently delivers the best performance across all distance intervals. Specifically, it achieves 35.50% AP_{3D} and 44.80% AP_{BEV} in the long-range (50–100 m) scenario—significantly surpassing early fusion (33.05% / 36.17%) and late fusion (9.81% / 12.99%) methods. Although early fusion (e.g., PointPillars) shows competitive accuracy at close and mid-ranges, it suffers from excessive communication overhead, transmitting approximately 1.38 MB of data per frame. In contrast, V2IFormer not only achieves higher detection accuracy, especially at long distances, but also maintains a much lower communication load of only 146.24 KB per frame, making it more suitable for real-time cooperative perception systems with limited bandwidth. ImVoxelNet, relying solely on monocular camera inputs, struggles with detecting distant objects due to insufficient depth information. PointPillars performs better for infrastructure-side perception by directly processing LiDAR point clouds, which offer precise spatial geometry. However, its performance declines at longer ranges. In summary, V2IFormer effectively addresses three key challenges in V2X cooperative perception: achieving high detection accuracy, ensuring reliable long-range performance, and maintaining efficient bandwidth usage for scalable deployment.

HandlingDynamicWeathersandLightingConditions

V2IFormer is then evaluated on real-world datasets that naturally encompass diverse and dynamically changing environmental conditions, including variations in challenging weather (e.g., rainy,

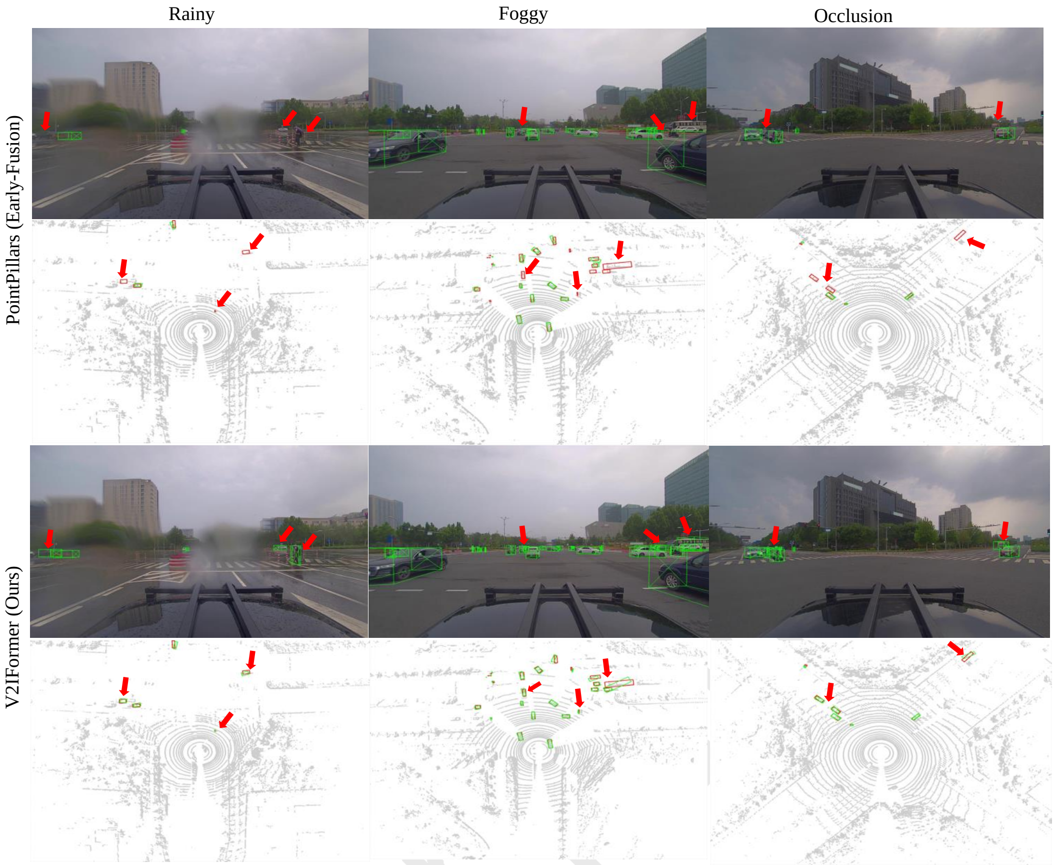

Figure 3: Visualization results under occlusion and adverse weather conditions show that V2IFormer accurately detects partially occluded vehicles, as well as distant vehicles and pedestrians in rainy and foggy scenes.

Under

foggy). While recent advances in 3D object detection have achieved remarkable performance on standard vehicle-centric benchmarks such as *KITTI* (*19*), *nuScenes* (*43*), and *Waymo* (*44*), their generalization capability under extreme or adverse conditions remains limited (*45, 46*). To address this, some prior works have curated dedicated datasets targeting challenging weather scenarios for robust vehicle detection (*47, 48*). In contrast, the cooperative vehicle-infrastructure datasets *DAIR-V2X* used in this work are collected in-the-wild under a wide spectrum of real-world conditions without frame-level annotation of specific weather types. The absence of fine-grained environmental annotations precludes rigorous quantitative benchmarking under specific conditions, limiting comprehensive performance evaluation. Consequently, we adopt qualitative assessments to validate *V2IFormer*'s robustness in challenging scenarios. As illustrated in Figure 3, we compare detection results under occlusion, foggy, and rainy conditions using *PointPillars* (*41*), and our *V2IFormer*. *V2IFormer* consistently outperforms baselines, accurately detecting partially occluded vehicles, small pedestrian targets, and distant objects even in low-visibility or degraded illumination conditions. This superior performance stems from *V2IFormer*'s enhanced BEV representation, which effectively fuses vehicle-side and infrastructure-side BEV features.

CONCLUSION

In this paper, we proposed V2IFormer, a novel V2I cooperative 3D object detection framework that integrates multi-view image features from both vehicle-mounted and infrastructure-side cameras into a unified BEV representation. To address the challenges of occlusion and limited field of view in complex driving scenarios, we introduced a HeightNet module on the infrastructure side to enhance BEV feature generation through height-aware modeling. On the vehicle side, depth distribution prediction based on the LSS framework is employed to lift perspective-view features into BEV space. Furthermore, a deformable mutual-attention module was designed to adaptively fuse features from both views, focusing on informative regions and suppressing irrelevant noise. Extensive experiments on the DAIR-V2X benchmark validate the effectiveness of the proposed approach. V2IFormer consistently outperforms existing state-of-the-art methods, particularly under conditions of heavy occlusion and degraded visibility. These results highlight the importance of cooperative perception and adaptive feature fusion in improving robustness and accuracy of 3D detection. Future work will explore the extension of this framework to more diverse sensor modalities and real-time deployment in large-scale autonomous driving systems.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grant U21A20446.

AUTHOR CONTRIBUTIONS

The authors' contributions to this paper are as follows: study conception and design: Guoqiang Mao, Tianxuan Fu; data collection: Tianxuan Fu, Keyin Wang; analysis and interpretation of results: Guoqiang Mao, Tianxuan Fu, Xiaojiang Ren; draft manuscript preparation: Tianxuan Fu, Guoqiang Mao. All authors reviewed the results and approved the final version of the paper.

Under Review

REFERENCES

1. Schwall, M., T. Daniel, T. Victor, F. Favaro, and H. Hohnhold, Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038*, 2020.
2. Ma, X., W. Ouyang, A. Simonelli, and E. Ricci, 3d object detection from images for autonomous driving: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 5, 2023, pp. 3537–3556.
3. Qiao, D. and F. Zulkernine, Adaptive feature fusion for cooperative perception using lidar point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 1186–1195.
4. Xu, R., H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 2583–2589.
5. Yang, L., K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, and P. Chen, Bevheight: A robust framework for vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21611–21620.
6. Yang, L., X. Zhang, J. Yu, J. Li, T. Zhao, L. Wang, Y. Huang, C. Zhang, H. Wang, and Y. Li, MonoGAE: Roadside monocular 3D object detection with ground-aware embeddings. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
7. Li, Y., Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, 2023, Vol. 37, pp. 1477–1485.
8. Li, Z., W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
9. Huang, J., G. Huang, Z. Zhu, Y. Ye, and D. Du, Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.
10. Peng, L., Z. Chen, Z. Fu, P. Liang, and E. Cheng, Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5935–5943.
11. Schramm, J., N. Vödisch, K. Petek, B. R. Kiran, S. Yogamani, W. Burgard, and A. Valada, Bevcar: Camera-radar fusion for bev map and object segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2024, pp. 1435–1442.
12. Wang, J., F. Li, Y. An, X. Zhang, and H. Sun, Towards robust lidar-camera fusion in bev space via mutual deformable attention and temporal aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
13. Yu, H., Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, et al., Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21361–21370.
14. Wang, Z., S. Fan, X. Huo, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, VIMI: Vehicle-infrastructure multi-view intermediate fusion for camera-based 3D object detection. *arXiv preprint arXiv:2303.10975*, 2023.

Under Review

15. Rukhovich, D., A. Vorontsova, and A. Konushin, Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2397–2406.

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

17. Noor-A-Rahim, M., Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, and H. V. Poor, 6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities. *Proceedings of the IEEE*, Vol. 110, No. 6, 2022, pp. 712–734.

18. Philion, J. and S. Fidler, Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, pp. 194–210.

19. Geiger, A., P. Lenz, and R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 3354–3361.

20. Geiger, A., P. Lenz, C. Stiller, and R. Urtasun, Vision meets robotics: The kitti dataset. *The international journal of robotics research*, Vol. 32, No. 11, 2013, pp. 1231–1237.

21. Brazil, G. and X. Liu, M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9287–9296.

22. Wang, T., X. Zhu, J. Pang, and D. Lin, Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 913–922.

23. Wang, Y., 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 *Conference on Robot Learning*, PMLR, 2022, pp. 180–191.

24. Liu, Y., T. Wang, X. Zhang, and J. Sun, Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, Springer, 2022, pp. 531–548.

25. Liu, Y., J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3262–3272.

26. Roddick, T., A. Kendall, and R. Cipolla, Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018.

27. Tseng, C.-Y., 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 4850–4857.

28. Fan, S., Z. Wang, X. Huo, Y. Wang, and J. Liu, Calibration-free bev representation for infrastructure perception. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 9008–9013.

29. Ye, X., M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, and E. Ding, Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In

Under Review

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21341–21350.

30. Chen, Q., S. Tang, Q. Yang, and S. Fu, Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2019, pp. 514–524.
31. Chen, Q., X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.
32. Marvasti, E. E., A. Raftari, and Y. P. Fallah, *Cooperative lidar object detection via feature sharing in deep networks*, 2023, uS Patent App. 17/928,473.
33. Xu, R., Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022.
34. Yang, W., Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15536–15545.
35. Mallot, H. A., H. H. Bülthoff, J. J. Little, and S. Bohrer, Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, Vol. 64, No. 3, 1991, pp. 177–185.
36. Wang, Y., 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8445–8453.
37. Reading, C., A. Harakeh, J. Chae, and S. L. Waslander, Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8555–8564.
38. He, K., X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
39. Zhu, X., H. Hu, S. Lin, and J. Dai, Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.
40. Zhu, X., W. Su, L. Lu, B. Li, X. Wang, and J. Dai, Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.
41. Lang, A. H., S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12697–12705.
42. Loshchilov, I. and F. Hutter, Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
43. Caesar, H., V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

Under Review

44. Sun, P., H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

45. Hahner, M., C. Sakaridis, M. Bijelic, F. Heide, F. Yu, D. Dai, and L. Van Gool, Lidar snowfall simulation for robust 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16364–16374.

46. Hahner, M., C. Sakaridis, D. Dai, and L. Van Gool, Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15283–15292.

47. Dong, Y., C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1022–1032.

48. Hendrycks, D., S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

Under Review