MIC-BEV: Infrastructure-Based Multi-Camera Bird's-Eye-View Perception Transformer for 3D Object Detection

Yun Zhang Zhaoliang Zheng Johnson Liu Zhiyu Huang*
Zewei Zhou Zonglin Meng Tianhui Cai Jiaqi Ma
University of California, Los Angeles
{yun666, zhz03, jwu7, zhiyuh}@ucla.edu

Abstract

*Infrastructure-based perception is critical to intelligent transportation systems, offering global situational awareness and enabling cooperative autonomy. However, existing camera-based detection models often underperform in such scenarios due to challenges such as multi-view infrastructure setup, diverse camera configurations, degraded visual inputs, and various road layouts. We introduce **MIC-BEV**, a Transformer-based bird's-eye-view (BEV) perception model for infrastructure-based multi-camera 3D object detection. MIC-BEV is designed to support a variable number of cameras with heterogeneous extrinsic and intrinsic parameters, and maintains robustness under sensor degradation. MIC-BEV features a graph-enhanced fusion module that aggregates multi-view image features into the BEV space by leveraging geometric relations between cameras and BEV cells and latent visual cues. Additionally, a BEV segmentation head is incorporated to enhance scene understanding. To support training and evaluation, we introduce **M2I**, a synthetic dataset for infrastructure-based object detection, featuring diverse camera configurations, road layouts, and environmental conditions. Extensive experiments on both M2I and the real-world dataset **RoScenes** demonstrate that MIC-BEV achieves state-of-the-art performance in 3D object detection. It also remains robust under challenging conditions, including extreme weather and sensor degradation, enabling real-world deployment.*

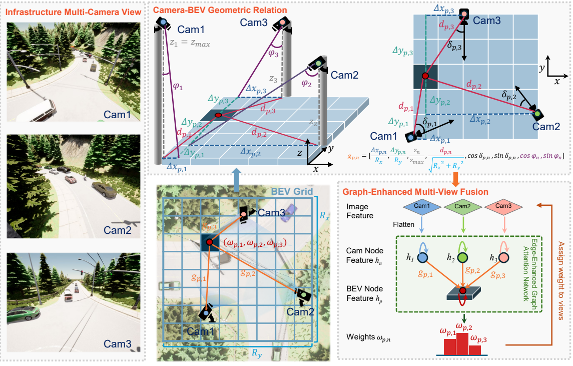


Figure 1. Illustration of infrastructure perception and the multi-camera BEV fusion module in MIC-BEV. Multiple infrastructure-mounted cameras observe the scene from different viewpoints. MIC-BEV encodes the geometric relations between each camera and the BEV grid by capturing spatial distances, heading angles, and height differences. These features are used to build a relation-edge-enhanced graph, enabling a graph attention network to predict view-specific importance weights for multi-view fusion. This enables MIC-BEV to be adaptable to various road layouts and heterogeneous camera configurations. The grid size shown is intended for illustration and not to scale.

1. Introduction

Infrastructure-based perception is a key enabler for intelligent transportation systems, providing critical support for traffic monitoring, situational awareness, and cooperative autonomy in urban environments [17, 31, 48]. Sensors deployed at intersections, crosswalks, and merging zones offer a strategic advantage for observing traffic participants from elevated viewpoints, providing broader and more stable observations. This spatial advantage facilitates long-term monitoring and enhances the ability to detect dynamic objects [2]. While LiDAR has been widely adopted for infrastructure-based object detection due to its accurate 3D measurements, it remains costly, maintenance-intensive, and sensitive to mounting constraints [25]. For instance, mounting LiDAR at higher positions reduces sensing resolution near the ground, while lower placements increase vulnerability to occlusion and physical damage [12, 14]. In contrast, cameras are significantly more affordable, scalable, and easier to deploy, making them an attractive alternative for large-scale infrastructure sensing [15].

While single-camera infrastructure perception systems are easier to deploy and have been widely explored in prior work [19, 35, 42], they suffer from limited spatial coverage and decreased robustness under occlusion or in complex scenes. In contrast, multi-

*Corresponding author

camera infrastructure sensing offers significant advantages by aggregating visual information from multiple viewpoints, leading to improved scene understanding and occlusion resilience [40]. However, infrastructure multi-camera systems also introduce several critical challenges. **1) Spatially distributed sensors.** Cameras deployed at large spatial distances often have overlapping fields of view with significant perspective differences and occlusions. These multi-view conditions make spatial alignment and feature fusion across views challenging. **2) High variability in camera configurations.** Unlike vehicle-mounted sensors that follow consistent mounting patterns, infrastructure cameras are deployed with diverse quantities, spatial layouts, orientations, fields of view (FoV), and degrees of overlap. Each intersection has a distinct design, requiring models to adapt to a wide range of installation configurations. **3) Sensor reliability and robustness.** Infrastructure cameras may degrade over time or fail without immediate detection or repair. Hence, perception models must be resilient to missing or low-quality inputs during deployment.

To address these challenges, we propose **MIC-BEV**, an effective 3D object detection model for infrastructure-based multi-camera systems using a Bird's-Eye-View (BEV) representation. MIC-BEV extends BEVFormer [20] with a relation-enhanced spatial cross-attention mechanism that fuses multi-view features through camera-specific features and their geometric relations for each BEV cell using a graph neural network (GNN) [37], as illustrated in Fig. 1. It also incorporates map-level and object-level BEV segmentation to improve spatial reasoning. MIC-BEV adapts to diverse camera and road layouts and applies camera masking strategies such as random dropout and Gaussian blur during training to enhance robustness to camera failure. To support training and evaluation, we introduce **M2I**, a large-scale synthetic dataset for Multi-camera, Multi-layout Infrastructure perception. M2I features a wide range of infrastructures with variability in camera quantity, spatial layout, heading angle, FoV, and degrees of overlap, as well as challenging conditions such as adverse weather and varying lighting, providing a comprehensive benchmark. The main contributions of this paper are summarized as follows:

1. We propose **MIC-BEV**, a BEV-based 3D object detection model for infrastructure multi-camera perception that fuses multi-view observations using spatial cross-attention enhanced with graph-based relation modeling.
2. We present **M2I**, a comprehensive dataset featuring diverse and realistic multi-camera settings and scene conditions, enabling model training and evaluation of generalization and robustness.
3. We demonstrate that MIC-BEV achieves strong performance and robustness on M2I and RoScenes datasets, validating its effectiveness under varying camera configurations, road layouts, and sensor degradation.

2. Related Work

Camera-based BEV Perception. BEV representations have become a dominant paradigm in camera-based 3D perception, offering a unified spatial abstraction across multi-view inputs. Early works such as OFT [29] and CADDN [49] project monocular camera image features into BEV space for 3D object detection. Lift-Splat-Shoot [28] extends this by lifting multi-view image features into a 3D voxel space using predicted depth and splatting them into a dense BEV plane. BEVDet [10] optimizes this process for multi-view efficiency. Transformer-based methods further advance BEV detection. DETR3D [36] and PETR [22] avoid explicit depth estimation by leveraging object queries and 3D reference points for cross-view feature aggregation, inspired by DETR [4] and Deformable DETR [51]. BEVFormer [20] introduces a learnable BEV query grid and applies spatiotemporal deformable attention for dense BEV fusion. BEVDet4D [9], SoloFusion [27], StreamPETR [34], and PETRv2 [23] incorporate temporal cues to enhance consistency and performance. While existing methods perform well in vehicle-mounted settings, they typically assume ego-centric cameras with fixed layouts. In contrast, infrastructure deployments involve dispersed cameras positioned around different intersections with varying viewpoints. This fundamental difference calls for BEV perception models tailored to infrastructure-based environments.

Infrastructure-based 3D Perception and Datasets. Infrastructure-based perception systems often rely on LiDAR [44, 47, 54] or LiDAR-camera fusion for 3D object detection [1, 24, 26, 53]. However, due to the high deployment cost of LiDAR, camera-only approaches are gaining growing interest [8, 19, 50]. Early efforts focused on monocular 3D detection using datasets such as Rope3D [45] and DAIR-V2X [46]. Methods like BEVDepth [18] improve depth estimation through LiDAR supervision, while BEVHeight [41], BEVHeight++ [43], and CoBEV [30] enhance spatial understanding by leveraging depth-height cues. More recently, MonoUNI [13] introduces normalized depth features to reduce reliance on explicit height cues, achieving better generalization from infrastructure to vehicle perspectives. While monocular setups have shown promise, multi-camera configurations offer broader spatial coverage and more robust performance. V2X-Real [38] and RCooper [6] focus on multi-camera perception in a four-way intersection and corridors, while RoScenes [52] covers highway scenes. RoBEV [52] and RopeBEV [11] establish strong baselines by fusing multi-view features using feature-guided queries and rotation-aware embeddings, respectively. However, these fusion strategies are largely implicit and

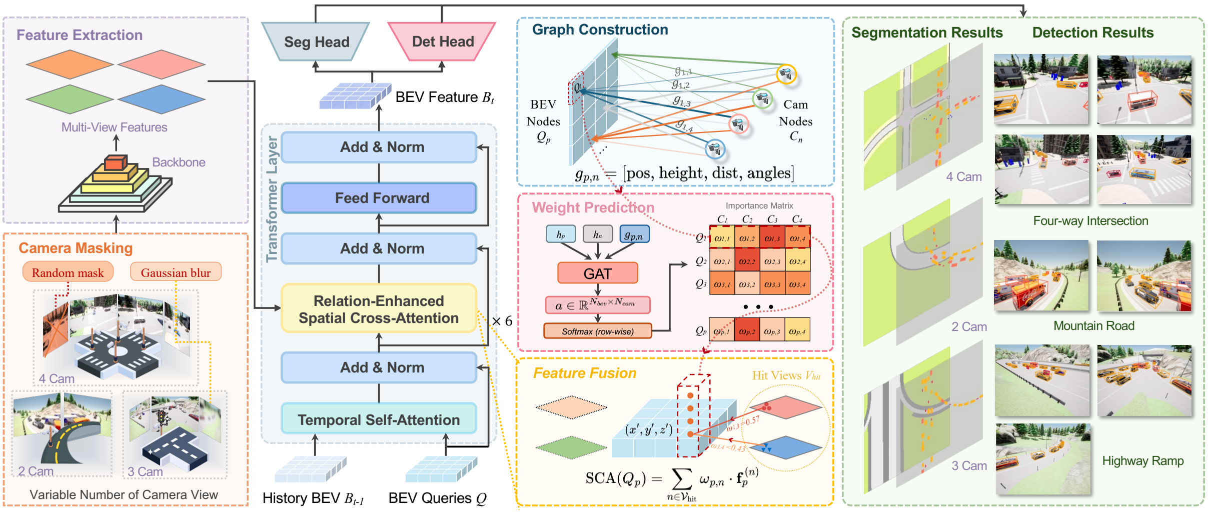


Figure 2. Overview of the **MIC-BEV** architecture. The model takes multi-view images from a variable number of infrastructure-mounted cameras as input and extracts features through a shared backbone. A camera masking module applies random dropout or Gaussian noise to simulate degraded views. The extracted features are fused into a BEV representation via Transformer layers with temporal self-attention and our proposed Relation-Enhanced Spatial Cross-Attention. GAT networks are used to dynamically assign view-dependent weights based on camera node features and geometric relations between the camera and its visible BEV cells. The resulting BEV features are used for both object detection and BEV segmentation tasks.

lack interpretability at the per-view level. Furthermore, the limited scene diversity in these datasets hampers generalization to more complex layouts. To address these limitations, we introduce the **M2I** dataset, which encompasses a wide variety of intersection types and infrastructure configurations. We propose **MIC-BEV**, which integrates a GNN to dynamically infer geometric-relation-aware, per-view fusion weights. This enables interpretable multi-view fusion that adapts to diverse scenarios.

3. Method

3.1. Problem Definition

The objective is to develop a multi-camera 3D object detection model for infrastructure-mounted sensors, enhanced by BEV segmentation as an auxiliary task to facilitate spatial reasoning. Given a set of synchronized multi-view RGB images, the model $\mathrm{Det}(\cdot)$ jointly predicts a set of 3D bounding boxes $\hat{B}$ and a BEV segmentation $\hat{M}$:

$$\hat{B}, \hat{M} = \mathrm{Det}_{\phi}\left(\{(I_n, E_n, K_n)\}_{n=1}^{N}\right) \tag{1}$$

where I_n is the RGB image from the n-th camera, E_n and K_n are the corresponding extrinsic and intrinsic matrices, and ϕ denotes the learnable parameters of the model. The number of cameras N varies across different scenes.

The primary task is 3D object detection, which involves predicting a set of bounding boxes $\hat{B}$ in a shared BEV coordinate frame. Each box $\hat{B}_i$ is parameterized as $\hat{B}_i = (x, y, z, l, w, h, \psi)$, where (x, y, z) denotes the object's position, (l, w, h) its bounding box dimensions, and ψ its yaw angle. To support spatial understanding, we formulate a BEV semantic segmentation task with two complementary levels: map-level and object-level. The model predicts a semantic map $\hat{M} = [\hat{M}_{\text{map}}, \hat{M}_{\text{object}}]$, where $\hat{M}_{\text{map}} \in \mathbb{R}^{N_{\text{map}} \times H_{\text{bev}} \times W_{\text{bev}}}$ captures static map semantics (e.g., road, crosswalk), and $\hat{M}_{\text{object}} \in \mathbb{R}^{N_{\text{object}} \times H_{\text{bev}} \times W_{\text{bev}}}$ captures dynamic objects (e.g., vehicle, pedestrian). Each BEV cell predicts a class-wise probability distribution for both static and dynamic elements, allowing the model to reason jointly about the environment and the spatial distribution of objects.

3.2. Overall Architecture

Our framework builds upon BEVFormer [20], extending to support infrastructure-mounted cameras with varying road layouts. As shown in Fig. 2, the model comprises three components : (1) an image backbone network for feature extraction, (2) a Transformer encoder with temporal and relation-enhanced spatial attention to aggregate image features into BEV space, and (3) task-specific decoding heads for 3D object detection and BEV segmentation.

3.3. Variable Multi-Camera Inputs

Infrastructure deployments often require different quantities of infrastructure-mounted cameras with varying fields of view. To ensure adaptability, our framework supports a variable number of input cameras. If fewer than the maximum number ($\mathcal{N}_{\max}$) are available, we pad the input with dummy images (zero-valued tensors) and assign identity matrices as their calibration parameters. These padded views are excluded from downstream spatial attention and graph computations by ensuring their 3D projections yield non-positive depths, preventing them from contributing to the set of effective views $\mathcal{V}_{\text{hit}}$ (see Sec. 3.4).

For M2I, we further apply a view-masking augmentation strategy to simulate partial sensor degradation during training. Specifically, we introduce a RandomMaskMultiView module that randomly masks exactly one real camera view per training sample with a probability of 0.25. The selected view is either zeroed out or blurred using a Gaussian kernel. To ensure reproducibility, the augmentation is applied deterministically, using a hash of the sample ID to seed the random operations. This setup promotes robustness to view missingness while maintaining consistent augmentation per frame.

3.4. Relation-Enhanced Transformer

Encoder and BEV Queries. We adopt a ResNet-101 [7] backbone coupled with a Feature Pyramid Network (FPN) [21] to extract multi-scale features from each camera image. The BEV representation is defined as a 2D grid anchored to the ground plane and centered at the scene. We initialize a learnable tensor $\mathbf{Q} \in \mathbb{R}^{H_{\text{bev}} \times W_{\text{bev}} \times C}$ to represent the grid, where H_{bev} and W_{bev} denote the grid resolution, and C is the feature dimension. Each cell $Q_p \in \mathbb{R}^C$ serves as a latent query corresponding to a grid location $p = (x, y)$ in the BEV space. These BEV queries interact with multi-view image features via spatial cross-attention and are iteratively refined to capture spatial cues encoded by the infrastructure-mounted cameras.

Temporal Self-Attention. To capture object dynamics, temporal self-attention allows current BEV queries Q to attend to the previous BEV map B_{t-1}. In static infrastructure setups, this is simplified without ego-motion compensation. Incorporating temporal context improves detection stability and recovers occluded or intermittent objects.

Spatial Cross-Attention (SCA). Given a set of multi-view camera feature maps $\{F^{(n)}\}_{n=1}^{N}$, SCA aggregates them into a unified BEV representation $F \in \mathbb{R}^{C \times H_{\text{bev}} \times W_{\text{bev}}}$. For each BEV query Q_p located at p in the BEV grid, we generate a vertical stack of N_{ref} 3D reference points $\mathbf{r}_{p,j} = (x, y, z_j)$ using a predefined set of anchor heights $\{z_j\}_{j=1}^{N_{\text{ref}}}$. These pillars help capture semantic features across different heights. Each 3D reference point $\mathbf{r}_{p,j}$ is projected onto the n-th camera view as 2D coordinates $\mathbf{u}_{p,j}^{(n)}$. Only camera views where the projected points fall within valid image bounds are included in the hit-view set $\mathcal{V}_{\text{hit}} \subseteq 1, \ldots, N$.

For each hit view $n \in \mathcal{V}_{\text{hit}}$, we apply deformable attention (DeformAttn) [51] around the projected locations $\{\mathbf{u}_{p,j}^{(n)}\}_{j=1}^{N_{\text{ref}}}$ of 3D reference points associated with BEV query Q_p. This produces a per-view feature $\mathbf{f}_p^{(n)} \in \mathbb{R}^C$. The final BEV feature is computed by fusing all visible views with learned weights $\omega_{p,n}$:

$$\begin{aligned} \mathrm{SCA}(Q_p) &= \sum_{n \in \mathcal{V}_{\text{hit}}} \omega_{p,n} \cdot \mathbf{f}_p^{(n)}, \ \sum_{n} \omega_{p,n} = 1, \\ \mathbf{f}_p^{(n)} &= \sum_{j=1}^{N_{\text{ref}}} \mathrm{DeformAttn}(Q_p,\, \mathbf{u}_{p,j}^{(n)},\, F_t^{(n)}). \end{aligned} \tag{2}$$

Relation-Enhanced Fusion via GAT. The conventional way of uniformly averaging the camera contributions ignores how *informative* or *reliable* each view is for a specific BEV cell. To address this limitation, we learn the fusion weights $\omega_{p,n}$ in Eq. (2) using a graph attention network (GAT) [33].

We construct a bipartite graph $\mathcal{G} = (\mathcal{V}_{\text{cam}}, \mathcal{V}_{\text{bev}}, \mathcal{E})$, where each camera node $C_n \in \mathcal{V}_{\text{cam}}$ represents a pooled image feature map from camera n, and each BEV grid cell node $Q_p \in \mathcal{V}_{\text{bev}}$ is represented by a BEV query located at p. The node features are defined as:

$$\mathbf{h}_p = Q_p \in \mathbb{R}^C \quad \text{for BEV nodes,} \tag{3}$$

$$\mathbf{h}_n = \frac{1}{K} \sum_{k=1}^{K} f_{n,k}^{(t)} \in \mathbb{R}^C \quad \text{for camera nodes,} \tag{4}$$

where $K = H \times W$ is the number of tokens from the camera feature map $F^n \in \mathbb{R}^{C \times H \times W}$, with H and W denoting the height and width of the feature map, respectively. $f_{n,k}^{(t)}$ denotes the k-th token feature from camera n.

Edges $\mathcal{E}$ are directed from cameras to visible BEV nodes, $\mathcal{E} = \{(n, p) \mid Q_p \text{ is visible from camera } C_n\}$. Each edge $(n \to p)$ is annotated with a geometric relation descriptor $\mathbf{g}_{p,n} \in \mathbb{R}^8$, consisting of:

$$\begin{aligned} \mathbf{g}_{p,n} = \Big[& \frac{\Delta x_{p,n}}{R_x}, \ \frac{\Delta y_{p,n}}{R_y}, \ \frac{z_n}{z_{max}}, \ \frac{||\mathbf{d}_{p,n}||_2}{\sqrt{R_x^2 + R_y^2}}, \\ & \cos\delta_{p,n}, \ \sin\delta_{p,n}, \ \cos\varphi_n, \ \sin\varphi_n \Big], \end{aligned} \tag{5}$$

where $\mathbf{d}_{p,n} = (\Delta x_{p,n}, \Delta y_{p,n}) = (x_p - x_n, y_p - y_n)$ is the 2D planar offset between the BEV grid and the camera center. R_x and R_y are normalization constants corresponding to the sensing range in x and y directions, used to scale spatial offsets to a consistent range within $[-1, 1]$. Similarly, z_n is the camera's height, normalized by the maximum camera height z_{max}. $\delta_{p,n}$ is the heading difference between the camera's yaw and the angle from camera n to the BEV cell at location p, and φ_n is the pitch angle of camera n. To ensure rotational continuity and avoid discontinuities near $\pm\pi$, we use heading with its sine and cosine components, i.e., $\cos\delta_{p,n}$ and $\sin\delta_{p,n}$. By jointly normalizing geometric features, we ensure that the network is invariant to map scale, BEV resolution, and elevation difference, enabling generalization across scenes with different layouts or camera setups.

We employ a GAT network f_θ to process the BEV node, camera node, and their geometric relation:

$$s_{p,n} = f_\theta(\mathbf{h}_p, \mathbf{h}_n, \mathbf{g}_{p,n}), \tag{6}$$

where $s_{p,n}$ denotes the raw importance score for the camera node n contributing to the BEV node p. For views not in the visible set, we enforce $s_{p,n} \leftarrow -\infty$ to exclude them. The fusion weights $\omega_{p,n}$ are obtained by applying a softmax operation over the valid scores $s_{p,n}$ across all camera nodes:

$$\omega_{p,n} = \frac{\exp(s_{p,n})}{\sum_{n' \in \mathcal{N}_{\text{max}}} \exp(s_{p,n'})}. \tag{7}$$

This geometry- and content-aware fusion strategy enables the model to selectively emphasize the most informative and geometrically favorable camera views, while suppressing occluded or degraded inputs. As a result, the fused BEV representation becomes more robust, interpretable, and reliable across a wide range of camera configurations.

3.5. Object Detection and BEV Segmentation

The BEV Transformer layers output a shared BEV feature map $B_t \in \mathbb{R}^{C \times H_{\text{bev}} \times W_{\text{bev}}}$ for both object detection and BEV segmentation, enabling joint optimization.

For object detection, we adopt a DETR-style decoder [4] with N_q object queries. Each query outputs a class probability vector $\hat{y} \in \mathbb{R}^{n_{\text{obj}}+1}$ and bounding box attributes $\hat{b}$. The detection loss $\mathcal{L}_{det}$ combines a focal classification loss $\mathcal{L}_{\text{cls}}$ and an L1 regression loss $\mathcal{L}_{\text{reg}}$. For BEV segmentation, Each query outputs class probability maps for both map-level and object-level segmentation, denoted as $\hat{M} = [\hat{M}_{\text{map}}, \hat{M}_{\text{object}}]$. The loss $\mathcal{L}_{\text{seg}}$ is defined as pixel-wise cross-entropy for $\hat{M}$ and ground truth $\hat{M}^*$.

The model is trained with a combined loss:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda \mathcal{L}_{seg}, \tag{8}$$

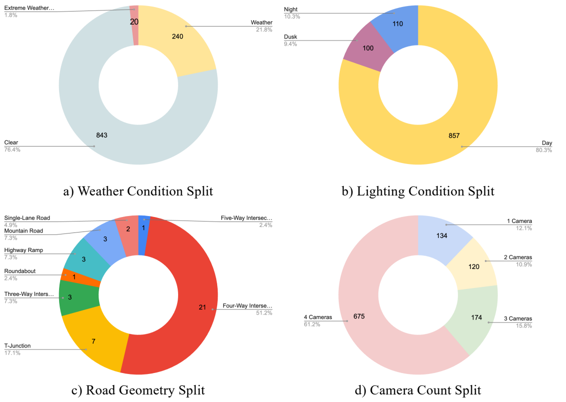
where λ is the task balance weight.

Joint training with map-level and object-level BEV segmentation boosts detection by enhancing spatial priors and refining class boundaries. Map-level segmentation provides high-level scene context that helps localize objects in semantically relevant regions (e.g., cars on roads, pedestrians on sidewalks), while object-level segmentation sharpens foreground separation and enforces shape continuity. This complementary supervision leads to more accurate predictions.

4. Experiments

4.1. Datasets

M2I. Existing infrastructure-based perception datasets typically cover a limited geographic area (e.g., a single intersection or short highway segment), with uniform camera setups and minimal variation in layout or weather. However, real-world deployments present far greater variability based on the unique geometry of each location. These constraints limit model generalization and hinder scalability in real-world applications. To bridge this gap, we propose the Multi-camera, Multi-layout Infrastructure (M2I) Perception Dataset, the first benchmark explicitly designed for 3D perception in heterogeneous roadside environments. Built with the CARLA simulator [5], M2I spans 41 locations across 7 towns, covering varied layouts such as intersections, roundabouts, blind corners, gas stations, and occlusion-prone areas. It also features diverse conditions, including heavy rain, dense fog, different times of day, and three traffic levels, offering a comprehensive benchmark for perception models. Each scene is manually equipped with 1 to 4 cameras selected from 11 real-world-inspired configurations [32, 39, 52], encompassing diverse quantities, spatial layouts, orientations, FoVs, and degrees of overlap. M2I contains over **926,950** images and **278,184** annotated frames, each equipped with synchronized LiDAR, 3D bounding boxes, and semantic BEV maps. The dataset includes **1,103** scenario clips from 41 diverse road geometries (e.g., highway ramps, T-junctions, 3-way and 4-way intersections) and varying traffic densities. Each clip spans 200-300 consecutive frames (10 Hz) with an average of 40 road users, covering multiple object classes such as cars, trucks, pedestrians, and cyclists. Fig. 3 summarizes the dataset composition across different splits, including variations in weather, lighting, and camera configurations.



a) Weather Condition Split

b) Lighting Condition Split

c) Road Geometry Split

d) Camera Count Split

Figure 3. **M2I Dataset Composition.** (a–b) Scenario-level splits by weather and lighting conditions. (c) Intersection-level split by road geometry. (d) Scenario-level split by number of cameras.

RoScenes. For real-world evaluation, we adopt RoScenes [52], the largest multi-view roadside perception dataset to date, featuring over 21 million 3D annotations captured across an 800m highway segment. Each scene is equipped with 6 to 12 synchronized infrastructure-mounted cameras. The dataset includes four object classes (car, van, bus, and truck) and covers daytime and nighttime conditions under normal and heavy traffic, collected in clear weather. The validation set includes clips from the top 10% most

Table 1. Comparison of 3D object detection performance on the **M2I** test set under three evaluation conditions: **Normal**, **Robust** (sensor dropout or degradation), and **Extreme Weather** (heavy rain or fog).

Method	Normal				Robust				Extreme Weather			
	mAP ↑	**NDS** ↑	mATE ↓	mASE ↓	**mAP** ↑	**NDS** ↑	mATE ↓	mASE ↓	**mAP** ↑	**NDS** ↑	mATE ↓	mASE ↓
Lift-Splat-Shoot	0.446	0.437	0.742	0.489	0.336	0.371	0.781	0.510	0.367	0.334	0.764	0.631
PETR	0.596	0.595	0.301	0.107	0.415	0.453	0.595	0.156	0.440	0.509	0.689	0.193
BEVFormer	0.637	0.678	0.235	0.072	0.513	0.643	0.288	0.084	0.445	0.560	0.535	0.131
PETRv2	0.651	0.689	0.213	0.093	0.505	0.582	0.295	0.156	0.584	0.623	0.530	0.127
UVTR	0.698	0.661	0.201	0.062	0.558	0.646	0.220	0.071	0.631	0.675	0.283	0.079
DETR3D	0.701	0.677	0.289	**0.056**	0.540	0.580	0.338	**0.063**	0.677	0.661	0.320	**0.069**
MIC-BEV (Ours)	**0.767**	**0.771**	**0.179**	0.061	**0.647**	**0.678**	**0.215**	0.065	**0.709**	**0.732**	**0.241**	0.077

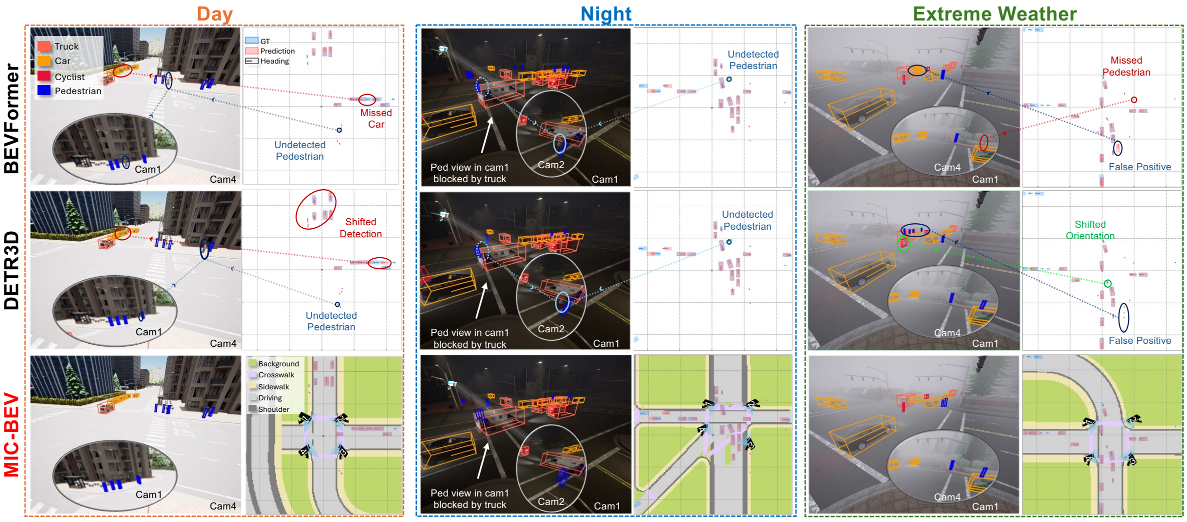


Figure 4. Qualitative comparison of MIC-BEV with baseline models (DETR3D and BEVFormer) across three scenarios. MIC-BEV produces more accurate detections with fewer false negatives and false positives, especially in occluded or sparsely covered regions, by leveraging relation-aware multi-view fusion. In Intersection 2, a pedestrian partially occluded by a truck in one camera view is missed by the baseline models but correctly detected by MIC-BEV.

occluded and bottom 10% least occluded scenes, selected as hard and easy cases, respectively.

4.2. Experimental Setup

Implementation Details. All models use a ResNet-101 backbone with deformable convolutions (ResNet101-DCN) as the image encoder. Training is conducted for 10 epochs on 8 NVIDIA L40S GPUs, using a batch size of 1 per GPU. We adopt the AdamW optimizer with an initial learning rate of 2×10^{-4}, scheduled via cosine annealing. For the M2I dataset, input images are resized to 800 × 600 pixels. The BEV grid is configured as 200×200, covering a square area of $[-51.2, \text{m}, 51.2, \text{m}]$ along both the X and Y axes. For the RoScenes dataset, input images have a resolution of 1920 × 1080 pixels. The BEV grid is set to 1000 × 250, covering a region of $[-400\text{m}, 400\text{m}]$ along the X-axis and $[-100\text{m}, 100\text{m}]$ along the Y-axis. Mixed-precision training (FP16) is used to reduce memory consumption. In both datasets, the origin of the BEV grid is aligned with either the center of the intersection or the position of a designated reference camera, depending on the scene configuration. BEV segmentation is disabled for RoScenes, as no semantic map or LiDAR data is provided to generate ground truth.

Evaluation Metrics and Baselines. We evaluate 3D object detection performance using standard nuScenes metrics [3], including mean Average Precision (mAP), nuScenes Detection Score (NDS), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), and Average Attribute Error (AAE). We compare our method against well-established detection models, including Lift-Splat-Shoot (LSS) [28], BEVFormer [20], DETR3D [36], PETR [22], PETRv2 [23], and UVTR [16]. We do not include RoBEV[52] or RopeBEV[11]. in the M2I benchmark due to the lack of available open-source code. To ensure compatibility with variable camera inputs, we apply our padding mechanism to all baselines. The RoScenes benchmark used in our evaluation is adapted from the setup introduced in RopeBEV.

4.3. Main Results

M2I. Tab. 1 reports the 3D object detection performance on the M2I testing set under three settings

Table 2. Comparison of 3D object detection performance on the **RoScenes** validation set under **Easy** and **Hard** levels.

Method	Easy					Hard					Average
	mAP ↑	**NDS** ↑	mATE ↓	mASE ↓	mAOE ↓	**mAP** ↑	**NDS** ↑	mATE ↓	mASE ↓	mAOE ↓	**NDS** ↑
SOLOFusion	0.129	0.308	0.878	0.144	0.517	0.066	0.202	0.844	0.144	1.000	0.255
BEVDet4D	0.200	0.428	0.896	0.094	0.041	0.139	0.393	0.922	0.099	0.038	0.411
BEVDet	0.299	0.506	0.742	0.079	0.042	0.184	0.445	0.754	0.087	0.043	0.476
StreamPETR	0.513	0.619	0.690	0.102	0.032	0.284	0.496	0.739	0.107	0.031	0.558
PETRv2	0.587	0.674	0.590	0.090	0.032	0.414	0.580	0.633	0.100	0.029	0.627
BEVFormer	0.609	0.693	0.560	0.078	0.030	0.433	0.597	0.600	0.090	0.029	0.645
DETR3D	0.644	0.722	0.501	0.067	0.031	0.471	0.633	0.508	0.080	0.028	0.678
RoBEV	0.684	0.753	0.442	0.058	0.031	0.524	0.672	0.438	0.077	0.027	0.713
RopeBEV	0.721	0.786	0.435	**0.056**	0.030	0.545	0.685	0.416	0.078	0.027	0.736
MIC-BEV (Ours)	**0.742**	**0.799**	**0.422**	0.058	**0.028**	**0.561**	**0.714**	**0.255**	**0.077**	**0.042**	**0.757**

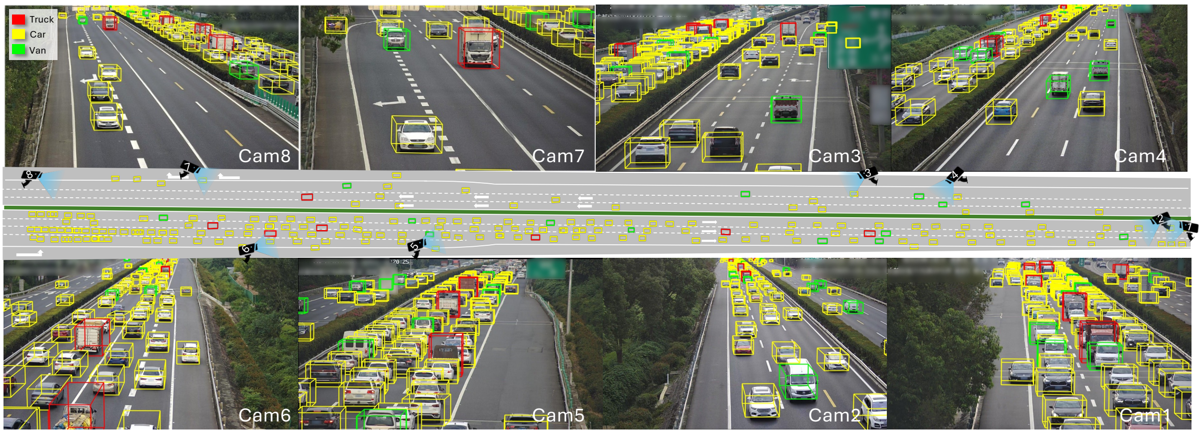


Figure 5. Detection results of MIC-BEV on the RoScenes validation set using 8 synchronized infrastructure-mounted cameras. Each panel shows per-camera predictions; the center shows the fused BEV output over the full 800 m range.

(Normal, Robust, and Extreme Weather). MIC-BEV achieves state-of-the-art performance across all settings. Under the Normal setting, MIC-BEV attains 0.767 mAP and 0.771 NDS, surpassing the strongest baseline (DETR3D) by over 6 points in mAP. The performance gap widens as input quality degrades. In the Robust setting (one camera is blurred or missing), MIC-BEV still achieves competitive results with 0.647 mAP and 0.678 NDS, corresponding to only a 15% performance drop. In contrast, DETR3D and BEVFormer exhibit performance degradation of up to 25%. Under Extreme Weather conditions, MIC-BEV still performs the best with 0.709 mAP and 0.732 NDS. Across all settings, MIC-BEV consistently outperforms the best competing method by 4-11 points in mAP, corresponding to relative improvements of 11-17%. These results highlight the efficacy of MIC-BEV's camera-grid relation-enhanced attention, which facilitates more effective multi-view feature fusion into the BEV space, thereby improving detection accuracy. The M2I benchmark validates MIC-BEV's robustness under diverse camera placements and degraded sensor configurations, and enhanced capability in detecting small and distant objects.

Fig. 4 illustrates the detection outputs of MIC-BEV. Compared to existing baselines, MIC-BEV reliably detects objects missed by other methods, corrects misaligned bounding boxes, and suppresses false positives. The BEV segmentation head in MIC-BEV provides spatial priors that effectively separate foreground instances from the background, thus benefiting detection. Moreover, camera masking during training simulates partial sensor failures, encouraging the model to leverage complementary views and enhancing its robustness to occlusions and degraded inputs.

RoScenes. Tab. 2 summarizes the 3D object detection results on the RoScenes benchmark. MIC-BEV achieves the highest performance across both evaluation splits, obtaining 0.742 mAP and 0.799 NDS on the Easy set, and 0.561 mAP and 0.714 NDS on the Hard set, where challenges such as occlusion and limited viewpoint overlap are more pronounced. Compared to RopeBEV, MIC-BEV improves mAP by 2.1 points on the Easy split and 1.6 points on the Hard split, while significantly reducing mATE on the Hard scenes by 38.6%. Notably, MIC-BEV shows strong detection capability across the large-scale detection area (800-meter range across eight camera views), reliably detecting distant and densely distributed vehicles in complex traffic scenarios, which is an essential capability for infrastructure-based perception. As illustrated in Fig. 5, MIC-BEV produces dense and spatially coherent detection results with minimal category ambiguity, and maintains stable predictions across varying viewpoints. These results further confirm the effectiveness of the proposed relation-enhanced attention mechanism, which enables detection under occlusion, sparse views, and long-range areas.

4.4. Ablation Studies

Influence of Key Components. Tab. 3 analyzes the contribution of key components in MIC-BEV. Starting from a BEVFormer baseline, introducing camera masking in training improves mAP by 5.1 points under the Robust setting, underscoring its effectiveness in simulating sensor degradation during training. Incorporating BEV segmentation further enhances the mAP by 4.2 points under the Normal setting, reflecting improved spatial discrimination between foreground objects and background context. The most substantial gain results from integrating relation-enhanced multi-camera fusion through GNN, which improves mAP by 7.6 points under the Normal setting and 5 points under the Robust setting. These results validate the overall design of MIC-BEV, showing the strengths of the proposed components that improve detection performance.

Table 3. Ablation study on the M2I test set. “N” for Normal setting; “R” for Robust setting with sensor degradation.

Cam. Masking	BEV Seg.	GNN Rel.	mAP (N)	NDS (N)	mAP (R)	NDS (R)
✗	✗	✗	0.637	0.678	0.513	0.643
✓	✗	✗	0.649	0.689	0.564	0.571
✓	✓	✗	0.691	0.727	0.597	0.601
✓	✓	✓	**0.767**	**0.771**	**0.647**	**0.678**

Runtime performance. We evaluate the inference speed of MIC-BEV and baseline models on an NVIDIA L40S GPU. As shown in Tab. 4, MIC-BEV runs at 4.65 FPS, which is on par with BEVFormer (4.76 FPS) and suitable for near real-time applications that prioritize high accuracy. Although some lightweight models, such as LSS (17.52 FPS) and PETR (13.33 FPS), achieve higher throughput, they offer significantly lower detection performance. This trade-off highlights MIC-BEV's strength in delivering high detection accuracy with efficient runtime.

Influence of GNN Relation Encoding We conduct an ablation study to evaluate the contribution of latent camera features, distance-based relations, and angle-based relations within the graph-based fusion module. Starting from a baseline that excludes all three components, we observe a mAP of 0.691 and NDS of 0.727. Introducing only geometric relations (distance and angle) without camera features yields a notable improvement (mAP: 0.729, NDS: 0.745), highlighting the importance of spatial structure among infrastructure-mounted cameras. Using only camera features provides comparable gains (mAP: 0.725, NDS: 0.740), indicating that latent image features are beneficial but slightly less effective than geometric relations. Combining camera features with distance relations further boosts performance to 0.761 mAP and 0.765 NDS. The full model, which incorporates all three components, achieves the best results (mAP: 0.767, NDS: 0.771), demonstrating that latent camera features and geometric relations are complementary and essential for robust multi-view detection.

Table 4. Comparison of inference speed (frames per second) and mAP on the M2I test set under normal conditions.

Model	**mAP (M2I)**	**FPS**
LSS	0.446	17.52
PETR	0.596	13.33
BEVFormer	0.637	4.76
PETRv2	0.651	9.17
UVTR	0.698	9.09
DETR3D	0.701	11.76
MIC-BEV	0.767	4.65

Table 5. Influence of latent camera features, distance, and angle relations in multi-view fusion module on M2I dataset.

Cam.	**Dist.**	**Angle**	**mAP**	**NDS**
✗	✗	✗	0.691	0.727
✗	✓	✓	0.729	0.745
✓	✗	✗	0.725	0.740
✓	✓	✗	0.761	0.765
✓	✓	✓	**0.767**	**0.771**

5. Conclusions

We propose MIC-BEV, a Transformer-based framework for multi-camera infrastructure perception, along with M2I, a synthetic benchmark encompassing diverse road layouts, camera placements, and adverse weather. MIC-BEV introduces camera-BEV relation-aware attention for multi-view fusion, a BEV segmentation head for improved spatial reasoning, and targeted data augmentation to simulate sensor degradation. Extensive evaluations on both the synthetic M2I and real-world RoScenes datasets shows that MIC-BEV achieves SOTA accuracy and robustness across a wide range of conditions, including complex urban intersections and long-range highway scenarios. Future work will explore extending MIC-BEV to multi-object tracking and real-time deployment in more complex, dynamic environments.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 2

[2] Zhengwei Bai, Guoyuan Wu, Xuewei Qi, Yongkang Liu, Kentaro Oguchi, and Matthew J Barth. Infrastructure-based object detection and tracking for cooperative driving automation: A survey. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1366–1373. IEEE, 2022. 1

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5

[5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 5

[6] Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Rcooper: A real-world large-scale dataset for roadside cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22347–22357, 2024. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[8] Hanjiang Hu, Zuxin Liu, Sharad Chitlangia, Akhil Agnihotri, and Ding Zhao. Investigating the impact of multi-lidar placement on object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2550–2559, 2022. 2

[9] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection, 2022. 2

[10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv e-prints*, pages arXiv–2112, 2021. 2

[11] Jinrang Jia, Guangqi Yi, and Yifeng Shi. Ropebev: A multi-camera roadside perception network in bird's-eye-view. *arXiv preprint arXiv:2409.11706*, 2024. 2, 6

[12] Wentao Jiang, Hao Xiang, Xinyu Cai, Runsheng Xu, Jiaqi Ma, Yikang Li, Gim Hee Lee, and Si Liu. Optimizing the placement of roadside lidars for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18381–18390, 2023. 1

[13] Jia Jinrang, Zhenjia Li, and Yifeng Shi. Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. *Advances in Neural Information Processing Systems*, 36:11703–11715, 2023. 2

[14] Tae-Hyeong Kim, Gi-Hwan Jo, Hyeong-Seok Yun, Kyung-Su Yun, and Tae-Hyoung Park. Placement method of multiple lidars for roadside infrastructure in urban environments. *Sensors*, 23(21):8808, 2023. 1

[15] Laurent Kloeker, Gregor Joeken, and Lutz Eckstein. Economic analysis of smart roadside infrastructure sensors for connected and automated mobility. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2331–2336. IEEE, 2023. 1

[16] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35: 18442–18455, 2022. 6

[17] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 1

[18] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1477–1485, 2023. 2

[19] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, and Junjun Jiang. Unsupervised domain adaptation for monocular 3d object detection via self-training. In *European conference on computer vision*, pages 245–262. Springer, 2022. 1, 2

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu10568349, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3, 6

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[22] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pages 531–548. Springer, 2022. 2, 6

[23] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3262–3272, 2023. 2, 6

[24] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE In-*

ternational Conference on Robotics and Automation (ICRA), pages 2774–2781. IEEE, 2023. 2

[25] Michael Lötscher, Nicolas Baumann, Edoardo Ghignone, Andrea Ronco, and Michele Magno. Assessing the robustness of lidar, radar and depth cameras against ill-reflecting surfaces in autonomous vehicles: An experimental study. In *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE, 2023. 1

[26] Zonglin Meng, Yun Zhang, Zhaoliang Zheng, Zhihao Zhao, and Jiaqi Ma. Agentalign: Misalignment-adapted multi-agent perception for resilient inter-agent sensor correlations. *arXiv preprint arXiv:2412.06142*, 2024. 2

[27] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection, 2022. 2

[28] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. 2, 6

[29] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection, 2018. 2

[30] Hao Shi, Chengshan Pang, Jiaming Zhang, Kailun Yang, Yuhao Wu, Huajian Ni, Yining Lin, Rainer Stiefelhagen, and Kaiwei Wang. Cobev: Elevating roadside 3d object detection with depth and height complementarity. *IEEE Transactions on Image Processing*, 2024. 2

[31] Miao Tang, Dianyu Yu, Peiguang Li, Chengwen Song, Pu Zhao, Wen Xiao, and Nengcheng Chen. A multi-scene roadside lidar benchmark towards digital twins of road intersections. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:341–348, 2024. 1

[32] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. *Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 4

[34] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3631, 2023. 2

[35] Wenjie Wang, Yehao Lu, Guangcong Zheng, Shuigen Zhan, Xiaoqing Ye, Zichang Tan, Jingdong Wang, Gaoang Wang, and Xi Li. Bevspread: Spread voxel pooling for bird's-eye-view representation in vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14718–14727, 2024. 1

[36] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on robot learning*, pages 180–191. PMLR, 2022. 2, 6

[37] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (1):4–24, 2021. 2

[38] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, et al. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, pages 455–470. Springer, 2024. 2

[39] Hao Xiang, Zhaoliang Zheng, Xin Xia, Seth Z. Zhao, Letian Gao, Zewei Zhou, Tianhui Cai, Yun Zhang, and Jiaqi Ma. V2x-realo: An open online framework and dataset for cooperative perception in reality. *ECCV*, 2024. 5

[40] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *6th Annual Conference on Robot Learning*, 2022. 2

[41] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng 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*, pages 21611–21620, 2023. 2

[42] Lei Yang, Xinyu Zhang, Jiaxin Yu, Jun Li, Tong Zhao, Li Wang, Yi Huang, Chuang Zhang, Hong Wang, and Yiming Li. Monogae: Roadside monocular 3d object detection with ground-aware embeddings. *IEEE Transactions on Intelligent Transportation Systems*, 25 (11):17587–17601, 2024. 1

[43] Lei Yang, Tao Tang, Jun Li, Kun Yuan, Kai Wu, Peng Chen, Li Wang, Yi Huang, Lei Li, Xinyu Zhang, et al. Bevheight++: Toward robust visual centric 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):5094–5111, 2025. 2

[44] Zhenwei Yang, Jilei Mao, Wenxian Yang, Yibo Ai, Yu Kong, Haibao Yu, and Weidong Zhang. Lidar-based end-to-end temporal perception for vehicle-infrastructure cooperation. *IEEE Internet of Things Journal*, 12(13):22862–22874, 2025. 2

[45] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21341–21350, 2022. 2

[46] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui 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*, pages 21361–21370, 2022. 2

[47] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Jirui Yuan, Ping Luo, and Zaiqing Nie. Vehicle-infrastructure cooperative 3d object detection via feature flow prediction, 2023. 2

[48] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. *AAAI*, 2025. 1

[49] Linping Zhang, Yu Liu, Xueqian Wang, You He, Gang Li, Yiming Zhang, Chang Liu, Zhizhuo Jiang, and Yang Liu. Caddn: A content-aware downsampling-based detection method for small objects in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025. 2

[50] Zhaoliang Zheng, Yun Zhang, Zongling Meng, Johnson Liu, Xin Xia, and Jiaqi Ma. Inspe: Rapid evaluation of heterogeneous multi-modal infrastructure sensor placement, 2025. 2

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 2, 4

[52] Xiaosu Zhu, Hualian Sheng, Sijia Cai, Bing Deng, Shaopeng Yang, Qiao Liang, Ken Chen, Lianli Gao, Jingkuan Song, and Jieping Ye. Roscenes: A large-scale multi-view 3d dataset for roadside perception. In *European Conference on Computer Vision*, pages 331–347. Springer, 2024. 2, 5, 6

[53] Walter Zimmer, Joseph Birkner, Marcel Brucker, Huu Tung Nguyen, Stefan Petrovski, Bohan Wang, and Alois C Knoll. Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8. IEEE, 2023. 2

[54] Walter Zimmer, Jialong Wu, Xingcheng Zhou, and Alois C Knoll. Real-time and robust 3d object detection with roadside lidars. In *Proceedings of the 12th International Scientific Conference on Mobility and Transport: Mobility Innovations for Growing Megacities*, pages 199–219. Springer, 2023. 2