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

Anonymous ICCV submission

Paper ID 15

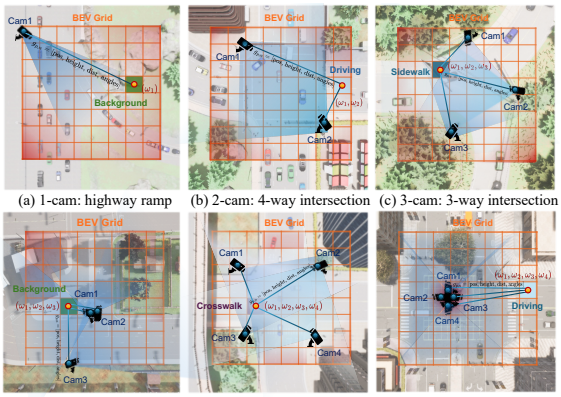
Abstract

*Infrastructure-based perception plays a pivotal role in intelligent transportation systems by providing global situational awareness and enabling cooperative autonomy. However, existing models struggle with the challenges of infrastructure settings, including diverse camera poses and configurations, significant perspective variation from wide-baseline viewpoints, and practical issues such as sensor degradation. To address these limitations, we introduce **MIC-BEV**, a Transformer-based Bird's-Eye-View (BEV) perception model for multi-camera infrastructure environments. MIC-BEV supports a variable number of camera inputs and includes a graph-based feature fusion module that captures geometric relationships between cameras. It also features a BEV semantic map prediction head to enhance scene understanding. To improve robustness, MIC-BEV is trained with random camera masking and Gaussian blur, simulating partial sensor failure and degraded image quality. Furthermore, we present the **M2I** dataset, a new benchmark on multi-view infrastructure perception featuring diverse infrastructure configurations and road geometries. Experiments on M2I demonstrate that MIC-BEV consistently outperforms existing state-of-the-art methods in infrastructure-based 3D object detection. It also maintains robustness under simulated sensor failures, demonstrating strong performance even in challenging test conditions.*

1. Introduction

Infrastructure-based perception is a key enabler for intelligent transportation systems, providing critical support for traffic monitoring [1, 52, 57], situational awareness [8, 12, 61], and cooperative autonomy [29, 36, 53] in urban environments. 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 [3, 49, 55]. While LiDAR has been widely adopted for infrastructure-based object detection due to its accurate 3D measurements [36, 42], it remains costly, maintenance-intensive, and sensitive to mounting constraints [10, 30]. For instance, mounting LiDAR at higher positions reduces sensing resolution near the ground, while lower placements increase vulnerability to occlusion and physical damage [18, 20]. In contrast, cameras are significantly more affordable, scalable, and easier to deploy, making them an attractive alternative for large-scale infrastructure sensing [4, 21].

While single-camera infrastructure perception systems are easier to deploy and have been widely explored in prior work [34, 44, 45], they suffer from limited spatial coverage and decreased robustness under occlusion or in complex scenes. In contrast, multi-camera infrastructure sensing offers significant advantages by aggregating visual information from multiple



(a) 1-cam: highway ramp (b) 2-cam: 4-way intersection (c) 3-cam: 3-way intersection
(d) 3-cam: T-intersection (e) 4-cam: 4-way intersection (f) 4-cam: (centralized placement)

Figure 1. Representative scenarios illustrating various infrastructure-mounted camera layouts at intersections. Each setup overlays one to four infrastructure-mounted cameras onto a predefined BEV perception grid. In **MIC-BEV**, a relation-enhanced spatial cross-attention module employs GNN to assign geometry-aware, per-view fusion weights to each camera based on camera node features and spatial edge relations for each BEV cell. Beyond 3D object detection, MIC-BEV predicts semantic maps, labeling each BEV cell with classes such as driving lane, parking area, sidewalk, or background. Note: The grid size shown is not to scale and is intended for illustrative purposes only.

viewpoints, leading to improved object coverage and scene understanding [11, 40]. However, multi-camera systems also introduce several critical challenges. **1) High variability in camera poses and configurations.** Unlike vehicle-mounted sensors that follow consistent mounting patterns, infrastructure cameras are deployed with diverse poses, orientations, fields of view, spatial layouts, and quantities. Each intersection has a distinct design, requiring models to adapt to a wide range of installation geometries and camera configurations. **2) Wide-baseline viewpoints.** Cameras deployed at large spatial distances often have overlapping fields of view with significant perspective differences and occlusions. These wide-baseline conditions make spatial alignment and feature fusion across views challenging. **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**, a robust and effective 3D object detection model designed for infrastructure-based multi-camera systems using a Bird's-Eye-View (BEV) representation. MIC-BEV extends BEVFormer [25] by incorporating 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). This enables adaptation to diverse camera and road layouts, as illustrated in Fig. 1. We utilize random camera view dropout or corruption during model training, enhancing robustness to camera failure at inference time. To support training and evaluation, we introduce **M2I**, a large-scale dataset for Multi-camera, Multi-configuration Infrastructure perception. M2I features diverse traffic scenarios in simulated environments, encompassing variations in the quantity, position, orientation, and field-of-view of cameras. It offers a challenging benchmark across realistic deployment settings. The main contributions of this paper are summarized as follows:

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

2. Related Work

2.1. Camera-based BEV Perception

Bird's-eye-view (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 [33] and CADDN [54] project monocular camera image features into BEV space for 3D object detection. Lift-Splat-Shoot [32] 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 [16] optimizes this process for multi-view efficiency. Transformer-based methods further advance BEV detection. DETR3D [39] and PETR [27] avoid explicit depth estimation by leveraging object queries and 3D reference points for cross-view feature aggregation, inspired by DETR [7] and Deformable DETR [58]. They introduce 3D reference points to guide multi-view feature aggregation via cross-attention. BEVFormer [25] introduces a learnable BEV query grid and applies spatiotemporal deformable attention for dense BEV fusion. BEVDet4D [15] and PETRv2 [28] incorporate temporal cues to enhance consistency and performance. Despite their success in vehicle-mounted applications, most BEV methods assume static, full observability with fixed camera configurations, which do not hold in infrastructure-mounted applications. This motivates the development of BEV perception models for infrastructure-centric environments.

2.2. Infrastructure-based 3D Perception

Infrastructure-based perception systems often rely on LiDAR [31, 47, 60, 62] or LiDAR-camera fusion for 3D object detection [2, 23, 51, 61]. However, due to the high deployment cost of LiDAR [6, 14, 26], camera-only approaches are gaining growing interest. Early efforts focused on monocular 3D detection using datasets such as Rope3D [48] and DAIR-V2X [50]. Methods like BEVDepth [24] improve depth estimation through LiDAR supervision, while BEVHeight [44], BEVHeight++ [46], and CoBEV [34] enhance spatial understanding by leveraging depth-height cues. More recently, MonoUNI [19] 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. RCooper [13] focuses on multi-camera perception in a four-way intersection and corridors, while RoScenes [59] covers long-range highway scenes. RoBEV [59] and RopeBEV [17] 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 lack interpretability at the

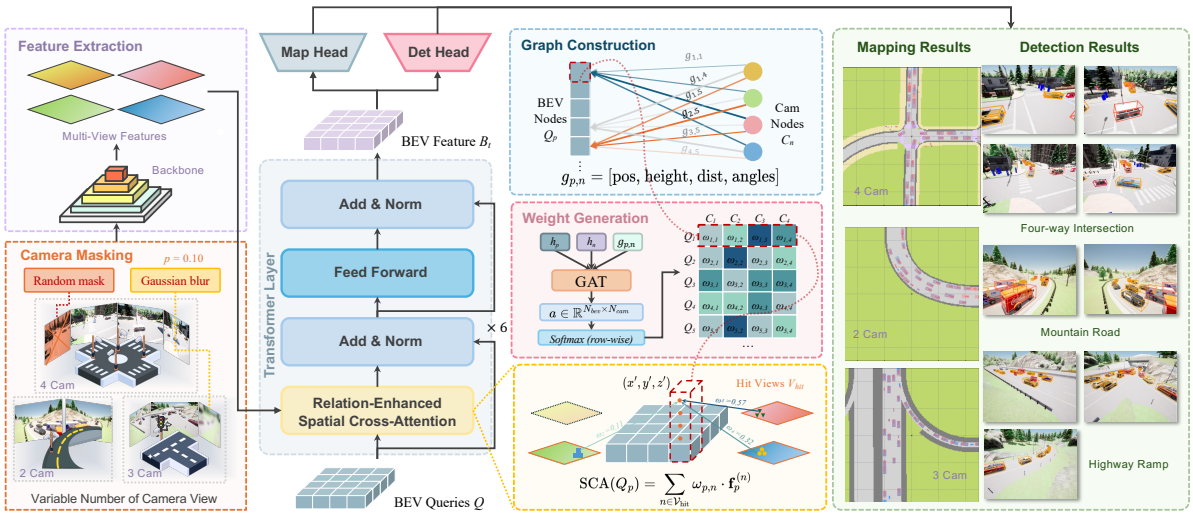


Figure 2. Overview of the **MIC-BEV** architecture. The framework takes multi-view images from a variable number of infrastructure-mounted cameras as input and extracts features through a shared backbone. A camera masking module applies random dropout or Gaussian noise to simulate degraded views. The extracted features are fused into a BEV representation via Transformer layers with the 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 map prediction tasks.

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 geometry-aware, per-view fusion weights. This design enables robust and interpretable multi-view fusion, offering adaptability to diverse layouts and situations.

3. Method

In this section, we present MIC-BEV, a Transformer-based framework for 3D object detection and semantic map prediction from infrastructure-mounted cameras. We first outline the problem statement and our overall architecture, then we present our model in detail.

3.1. Problem Definition

The objective is to develop a multi-camera 3D object detection model for infrastructure-mounted sensors, enhanced by semantic map prediction as an auxiliary task. The auxiliary supervision facilitates spatial reasoning and improves detection robustness.

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 semantic map $\hat{M}$:

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

where $I_n \in \mathbb{R}^{H \times W \times 3}$ is the RGB image from the n-th camera, $E_n \in \mathbb{R}^{3\times4}$ and $K_n \in \mathbb{R}^{3\times3}$ are the corresponding extrinsic and intrinsic matrices, and ϕ denotes the learnable parameters of the model. The quantity of cameras N varies across different scenes.

The primary task is 3D object detection, which is predicting a set of bounding boxes $\hat{B}$ in a shared BEV coordinate frame, where each box $\hat{B}_i$ is parameterized as $\hat{B}_i = (x, y, z, l, w, h, \psi)$, representing the object's position, dimensions, and yaw orientation. To support spatial understanding, we introduce semantic map prediction as an auxiliary objective. The model predicts a BEV semantic map $\hat{M} \in \mathbb{R}^{N_{\text{class}} \times H_{\text{bev}} \times W_{\text{bev}}}$, where N_{class} is the number of semantic classes (e.g., background, driving, crosswalk). Each grid cell (u, v) contains a per-class probability distribution $\hat{M}_{:,u,v}$.

3.2. Overall Architecture

Our framework builds upon BEVFormer [25], extending its capabilities to accommodate infrastructure-mounted camera setups with varying road layouts. As shown in Fig. 2, the model comprises four components : (1) an image encoder for feature extraction, (2) a BEV feature generator that lifts and aggregates image features into a unified top-down space, where a relation-enhanced spatial attention module is embedded within each Transformer layer to fuse multi-view features, and (3) task-specific decoding heads for 3D object detection and semantic map prediction.

3.3. Variable Multi-Camera Inputs

Infrastructure deployments often require a different quantity 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 (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. 3.5).

To improve robustness, we apply random view masking and noise injection during training. With a probability of $p_{\text{mask}} = 0.1$, one randomly selected camera view is either replaced with a dummy tensor or corrupted using Gaussian blur, simulating sensor degradation or camera dropout. This augmentation strategy encourages the model to maintain performance under partial observability. No masking or noise is applied when only a single view is present.

3.4. Encoder and BEV Queries

We adopt a ResNet backbone coupled with a Feature Pyramid Network (FPN) 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. Following BEVFormer [25], 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 spatial resolution, and C is the feature dimension. Each cell $Q_p \in \mathbb{R}^C$ serves as a latent query corresponding to a spatial location p 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.

3.5. Relation-Enhanced Transformer

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 (x, y) 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) [58] 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 $\omega_{p,n}$:

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

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) [38].

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 $(n \rightarrow p)$ is annotated with a geometry-aware descriptor $\mathbf{g}_{p,n} \in \mathbb{R}^8$, consisting of:

$$\begin{aligned} \mathbf{g}_{p,n} = \Big[&\tfrac{\Delta x_n}{R}, \tfrac{\Delta y_n}{R}, \tfrac{z_n}{H}, \tfrac{|\Delta \mathbf{x}|_2}{R\sqrt{2}}, \\ &\cos\delta_{p,n}, \sin\delta_{p,n}, \sin\phi_n, \cos\phi_n \Big], \end{aligned} \tag{5}$$

where $(\Delta x_n, \Delta y_n) = (x_p - x_n, y_p - y_n)$ is the 2D planar offset between the BEV grid and the camera center. R is a normalization constant corresponding to the sensing range, 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 H. $\delta_{p,n}$ 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 are computed via the softmax function:

$$\omega_{p,n} = \frac{\exp(s_{p,n})}{\sum_{m \in \mathcal{V}_{\text{cam}}} \exp(s_{p,m})}. \tag{7}$$

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.

BEV Transformer Layer. Each BEV Transformer layer integrates a relation-enhanced spatial cross-attention module to fuse multi-view image features into the BEV space in a geometry- and content-aware manner. This is followed by standard residual connections and layer normalization. A total of six such Transformer layers are stacked, allowing the model to progressively refine the BEV feature.

3.6. Object Detection and Map Prediction

The BEV Transformer layers output a BEV feature map $F \in \mathbb{R}^{C \times H_{\text{bev}} \times W_{\text{bev}}}$, which serves as a shared representation for both object detection and semantic map prediction. This design enables joint optimization, where supervision from one task can benefit the other by improving shared features.

For object detection, we adopt a DETR-style decoder [7] with $N_q = 200$ object queries. Each query outputs a class probability vector $\hat{y} \in \mathbb{R}^{n_{\text{obj}}+1}$ and bounding box attributes $\hat{b} = (x, y, z, l, w, h, \psi)$. We use Hungarian matching to assign predictions to the ground truth. The detection loss combines a focal classification loss $\mathcal{L}_{\text{cls}}$ and an L1 regression loss $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{8}$$

For semantic map prediction, we apply a decoder composed of Conv-GN-ReLU blocks, followed by a 1×1 convolutional classifier, which transforms the BEV feature map F into dense semantic logits $\hat{M} \in \mathbb{R}^{C_{\text{map}} \times H_{\text{bev}} \times W_{\text{bev}}}$. The map prediction loss is defined as pixel-wise cross-entropy:

$$\mathcal{L}_{seg} = \frac{1}{HW} \sum_{u,v} \mathbf{CE}(\hat{M}_{:u,v}, M^*_{:u,v}). \tag{9}$$

The model is trained with a combined loss:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda \mathcal{L}_{seg}, \tag{10}$$

where λ is the task balance weight.

Joint training with a map segmentation head enhances detection performance in several ways. First, map prediction encourages the BEV feature map F to capture geometry priors (e.g., road boundaries, sidewalks, parking zones), allowing object queries to focus on semantically meaningful regions and reducing false positives in background areas. Second, dense supervision across the entire BEV grid enhances the contrast between foreground and background, leading to more accurate bounding box localization. Finally, in cases of partial occlusion, semantic context from the map (e.g., road type or crosswalk borders) provides cues that help recover missing object evidence.

4. Experiments

4.1. Datasets

Most existing infrastructure-based perception datasets are limited in scope, typically capturing a single intersection or highway segments with uniform and constrained camera setups. In many cases, cameras are co-located on a single pole, resembling vehicle-mounted configurations [13, 23, 43]. Such arrangements often introduce blind spots below the pole [56] and fail to reflect the challenges of real-world deployments with varied spatial layouts. Furthermore, the quantity of cameras required for sufficient coverage varies significantly across different intersection geometries, rendering fixed configurations impractical for large-scale or cost-sensitive deployments.

To overcome these limitations, we introduce the *Multi-camera, Multi-configuration Infrastructure (M2I) Perception Dataset*. M2I is the first benchmark designed for 3D perception in diverse roadside environments with variable and realistic camera configurations. Built using the high-fidelity CARLA simulator [9], M2I spans 29 distinct environments across 7 different towns. It includes not only conventional intersections but also complex roadside areas such as blind zones near sharp turns, gas stations, and occlusion-heavy regions. Each scene is equipped with 1 to 4 cameras sampled from 8 diverse configurations, varying in position, orientation, and field of view (ranging from 100° to 120°) [37, 42]. Camera placements are manually curated to reflect real-world deployments, including those from V2X-Real [42], RoScense [59], Rcooper [13], and layouts specific to complex road types like T-junctions and 5-way intersections.

M2I contains over 610,000 images and 200,000 annotated frames, each with synchronized LiDAR, 3D bounding boxes, and semantic BEV maps. To model diverse traffic scenarios, we simulate three levels of traffic density (low, medium, and high) across sequences of 200-300 frames each. On average, each frame includes around 40 dynamic agents comprising cars, pedestrians, trucks, and cyclists. The dataset reflects realistic agent distribution, with an average composition of 65% cars, 20% trucks, 10% pedestrians, and 5% cyclists, closely aligned with statistics from established benchmarks such as nuScenes and Waymo [5, 35]. The dataset contains 844 scenario clips and is partitioned into training, validation, and test sets using a 7:1:2 ratio. In addition to object-level annotations, M2I provides semantic BEV maps for fine-grained scene understanding. These maps include seven semantic classes: `background`, `driving`, `sidewalk`, `crosswalk`, `shoulder`, `border`, and `parking`. These annotations support multiple tasks, including 3D detection, semantic segmentation, tracking, and temporal modeling.

Tab. 1 compares infrastructure-based and V2X perception datasets in terms of scale, camera configu-

Table 1. Comparison of Infrastructure Components in V2X and Infrastructure-Based datasets. Previous simulation datasets adopt fixed, centered, vehicle-style camera placements for infrastructure, which limit spatial diversity. In contrast, our **M2I** dataset introduces 10 diverse camera configurations across a wide range of roadside environments. It supports varied FOV settings and scene types, enabling more robust and generalizable benchmarking for infrastructure-based 3D perception.

Dataset	Type	Year	Frames	Boxs	# Cams	FoV	Map	Environment
V2X-Sim-I [23]	Sim	2022	60K	26.6K	4 (fixed layout)	Constant		Urban
V2XSet-I [43]	Sim	2022	44K	233K	4 (fixed layout)	Constant		Urban
DAIR-V2X-I [50]	Real	2022	10K	493K	1	Constant		Intersection
V2X-Real-I [41]	Real	2023	171K	1.2M	4 (fixed layout)	Constant		Intersection
V2X-Seq-I [41]	Real	2023	39K	464K	2 (fixed layout)	Constant	✓	Intersection
V2XPnP-Seq-I [57]	Real	2024	208K	1.45M	4 (fixed layout)	Constant	✓	Intersection
Rope3D [48]	Real	2022	50K	1.5M	1	Constant		Intersection
RCooper [13]	Real	2024	50K	242K	2–4	Constant		Intersection
RoScenes [59]	Real	2024	215K	21.13M	6–12	Varied		Highway
M2I	Sim	2025	610k	7M	1–4 (10 layout)	Varied	✓	Diverse

ration, field-of-view, map support, and scene diversity. Existing datasets often rely on fixed, limited camera layouts and are focused primarily on intersection scenes. In contrast, our proposed M2I dataset introduces 10 diverse camera configurations with varied FoVs across a wide range of simulated roadside environments. It provides 610k frames and 7 million annotated 3D boxes, along with detailed map data, making it one of the largest and most versatile datasets for infrastructure-based 3D perception.

4.2. Implementation Details

All models use a ResNet-101 backbone with deformable convolutions (ResNet101-DCN) as the image encoder, followed by an FPN producing feature maps at four scales: 1/16, 1/32, 1/64, and 1/128, each with an embedding dimension of 256. We train for 10 epochs and evaluate on the validation set after each epoch, selecting the checkpoint with the highest mAP. For our model, we use 3 layers with 4 attention heads per layer in GAT in spatial cross-attention, and the hidden dimension is 128. The map prediction head consists of four Conv-GN-ReLU blocks, and the balance weight λ in the loss function is set to 2.0. The object detection head is a DETR-style decoder with six Transformer layers.

The BEV grid is configured as 200×200 with a resolution of $0.512\,\text{m}$ per cell, covering a perception area of $[-51.2\,\text{m}, 51.2\,\text{m}]$ along both the X and Y axes. All models are trained for 10 epochs using 4 NVIDIA L40S GPUs with a batch size of 2 per GPU. We employ the AdamW optimizer with a learning rate of 2×10^{-4}, weight decay of 0.01, and a cosine annealing learning rate schedule. Input images have a resolution of 800×600, and standard multi-view photometric augmentations are applied during training. All models are trained to detect 4 object categories: pedestrian, car, cyclist, and truck, using consistent annotations across all baselines.

4.3. Evaluation Metrics and Baselines

We evaluate 3D object detection performance using two standard metrics: mean Average Precision (mAP) and nuScenes Detection Score (NDS) [5]. The mAP metric measures detection accuracy across multiple object classes and distance thresholds. Unlike conventional AP that uses fixed Intersection-over-Union (IoU) thresholds, the nuScenes benchmark defines true positives based on center distance thresholds (e.g., 0.5m, 1.0m, 2.0m, and 4.0m), which better accounts for annotation uncertainty in LiDAR-based datasets. mAP is computed as the average over all class-distance pairs:

$$\text{mAP} = \frac{1}{|\mathcal{C}| \cdot |\mathcal{D}|} \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} \text{AP}_{c,d}, \tag{11}$$

where $\mathcal{C}$ is the set of object classes, $\mathcal{D}$ is the set of distance thresholds, and $\text{AP}_{c,d}$ is the average precision for class c at distance threshold d.

NDS is a composite score that integrates mAP with five True Positive metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE). It provides a balanced evaluation of detection accuracy and localization fidelity:

$$\text{NDS} = \frac{1}{10}\left[5 \cdot \text{mAP} + \sum_{\text{mTP}} (1 - \min(1, \text{mTP}))\right], \tag{12}$$

where $\text{mTP} \in \{\text{mATE}, \text{mASE}, \text{mAOE}, \text{mAVE}, \text{mAAE}\}$.

We compare our method against state-of-the-art BEV-based models, including Lift-Splat-Shoot (LSS) [32], BEVFormer [25], DETR3D [39], PETR [27], and UVTR [22]. These models vary in camera configurations, feature lifting strategies, and types of supervision, offering a comprehensive benchmark for

Table 2. Performance comparison of BEV-based perception models on the M2I testing set.

Method	Normal					Robust				
	mAP ↑	**NDS** ↑	mATE ↓	mASE ↓	mAOE ↓	**mAP** ↑	**NDS** ↑	mATE ↓	mASE ↓	mAOE ↓
LSS [32]	0.446	0.407	0.742	0.489	0.194	0.336	0.337	0.781	0.510	0.224
DETR3D [39]	0.601	0.453	0.685	0.615	0.624	0.461	0.371	0.701	0.620	0.638
PETR [27]	0.652	0.623	0.310	0.118	0.129	0.523	0.545	0.340	0.134	0.148
BEVFormer [25]	0.691	0.676	0.211	0.094	0.084	0.581	0.596	0.241	0.109	0.107
UVTR [22]	0.723	0.701	0.201	**0.061**	**0.054**	0.558	0.603	0.220	**0.061**	0.054
MIC-BEV	**0.767**	**0.726**	**0.179**	0.062	0.058	**0.647**	**0.654**	**0.215**	0.071	**0.067**

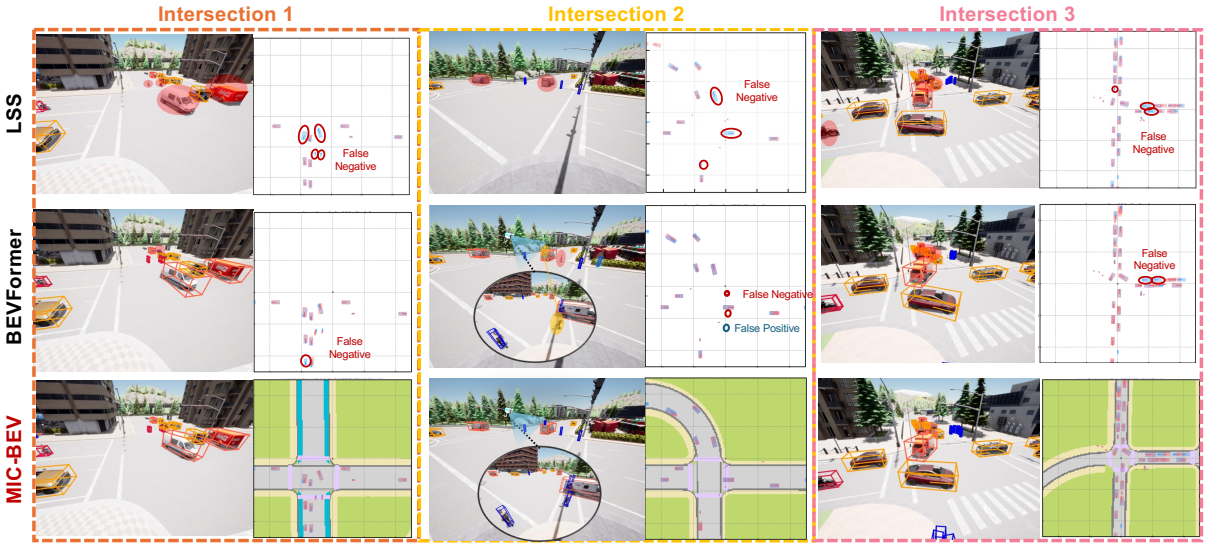


Figure 3. Qualitative comparison of MIC-BEV with baseline models (LSS and BEVFormer) across three intersections. MIC-BEV produces more accurate detections with fewer false negatives and false positives, especially in occluded or sparsely covered regions, by leveraging relation-aware multi-view fusion. In Intersection 2, a pedestrian partially occluded in one camera view is missed by BEVFormer but correctly detected by MIC-BEV.

infrastructure-based 3D perception. To ensure fair comparison under varying quantities of cameras, we introduced the same padding mechanism that enables the model to accept a variable quantity of camera inputs, similar to our method. This setup allows us to evaluate each model's robustness to camera sparsity consistently.

4.4. Main Results

We evaluate MIC-BEV under both standard and robust settings. In the robust setting (applied only when more than one camera is available), we randomly select one camera and, with 50% probability, either drop its input entirely or apply Gaussian blur with σ sampled uniformly from 3 to 10, simulating real-world sensor failures and distortions. As shown in Tab. 2, MIC-BEV achieves the highest performance across all metrics, with an mAP/NDS of 0.767/0.726 on the normal set and 0.647/0.654 on the robust set. Notably, MIC-BEV maintains strong accuracy under degraded conditions, outperforming the second best method (UVTR) by 4.4% mAP and 2.5% NDS in the robust setting. This highlights MIC-BEV's robustness to partial observability and sensor noise, which is a key advantage in infrastructure scenarios with diverse camera layouts and potential failures.

Tab. 3 presents per-class results on the normal M2I testing set. MIC-BEV consistently surpasses prior methods across all object categories. For pedestrians, MIC-BEV achieves an mAP of 0.860, significantly outperforming the second-best method (BEVFormer at 0.814), highlighting its effectiveness in detecting small and dynamic agents. For trucks, it scores 0.777, higher than UVTR (0.740), demonstrating robustness to large objects with varying shapes. For cars, MIC-BEV leads with 0.806, exceeding UVTR (0.748) and BEVFormer (0.659), maintaining high precision in dense, structured traffic environments. Finally, despite the inherent challenges of cyclist detection, it achieves 0.626, outperforming PETR and UVTR (0.597), reflecting its ability to handle occluded or elongated instances. These consistent per-class performance gains underscore our model's reliability in varied layouts and incomplete sensor views.

As shown in Fig. 3, MIC-BEV produces more complete and accurate detection across multiple intersec-

Table 3. Per-class results on the M2I normal testing set, using mAP as the primary metric.

Method	Pedestrian	Truck	Car	Cyclist	Avg.
LSS [32]	0.444	0.397	0.562	0.379	0.446
DETR3D [39]	0.764	0.571	0.584	0.485	0.601
PETR [27]	0.805	0.656	0.550	0.597	0.652
BEVFormer [25]	0.814	0.695	0.659	0.596	0.691
UVTR [22]	0.807	0.740	0.748	0.597	0.723
MIC-BEV	**0.860**	**0.777**	**0.806**	**0.626**	**0.767**

tions, with fewer false negatives and false positives compared to baseline models. We observe that the model consistently attends to complementary views when an object is partially visible, reinforcing its spatial reasoning capability. This aligns with the observed performance gains in occlusion-heavy scenarios.

4.5. Ablation Studies

In Tab. 4, we analyze the contributions of camera masking, BEV map prediction, and relation-enhanced attention. Each component provides clear performance gains, with camera masking improving robustness to missing views, and BEV map supervision enhancing spatial consistency. Incorporating relation-aware attention yields the largest boost, with the full MIC-BEV model achieving the best performance at 0.767 mAP and 0.726 NDS, demonstrating the effectiveness of dynamic, geometry-aware fusion.

Table 4. Ablation study on M2I dataset showing impact of camera mask, semantic map generation as auxiliary task, and relation-enhanced attention.

Cam. Masking	BEV Map	Relation	mAP	NDS
✗	✗	✗	0.691	0.676
✓	✗	✗	0.705	0.684
✓	✓	✗	0.727	0.697
✓	✓	✓	**0.767**	**0.726**

To assess the necessity of temporal modeling in our setting, we remove the temporal self-attention module from the base BEVFormer architecture. The result in Tab. 5 shows that while temporal modeling offers a slight improvement in NDS (0.729 vs. 0.726), it results in a minor drop in mAP (0.765 vs. 0.767). This indicates that temporal reasoning provides limited gains in static infrastructure scenarios, where cameras are fixed and each frame already contains rich spatial information. MIC-BEV therefore omits the temporal module, achieving strong performance while reducing model complexity.

To better evaluate the balance between model complexity and performance, we compare the trainable parameter counts of different model variants in Tab. 6. MIC-BEV removes the temporal self-attention module from BEVFormer and introduces a graph-based spatial fusion module along with a BEV semantic segmentation head. Despite these additions, the total number of trainable parameters increases by only around 2M (from 67.33M to 69.32M), representing a modest 3% growth. This small increase in model size leads to notable performance gains, highlighting the effectiveness of spatial relation modeling and semantic supervision in infrastructure-based perception.

Table 5. Influence of temporal self-attention module

Method	mAP	NDS
W/ temporal module	0.765	**0.729**
W/o temporal (base)	**0.767**	0.726

Table 6. Trainable parameter count across different versions of the model.

Model Variant	Trainable Parameters
BEVFormer (w/ temporal)	68,706,681
BEVFormer (w/o temporal)	67,326,201
MIC-BEV (GAT + Map Head)	69,321,692

5. Conclusions

We present MIC-BEV, a Transformer-based BEV perception framework designed for multi-camera infrastructure scenarios. Built on our proposed M2I dataset, which captures a wide range of roadside geometries and camera configurations, MIC-BEV employs relation-aware attention to dynamically fuse multi-view features with enhanced spatial understanding and interpretability. Experiments demonstrate that MIC-BEV outperforms existing BEV-based baselines in both clean and noisy settings, surpassing the performance of state-of-the-art BEV perception models. Ablation studies confirm the effectiveness of key components, including camera masking for handling variable inputs, BEV map segmentation as auxiliary supervision, and relation-enhanced multi-view fusion. Overall, MIC-BEV delivers a robust and scalable solution for real-world infrastructure perception.

While MIC-BEV demonstrates strong performance, it has several limitations. Its robustness under extreme weather or lighting conditions remains untested, and it assumes static, pre-calibrated camera setups. Additionally, the current model focuses primarily on object detection and does not yet address tasks such as tracking. Future work will investigate MIC-BEV's performance in adverse environmental conditions and evaluate its generalization on real-world infrastructure datasets.

References

[1] Eduardo Arnold, Mehrdad Dianati, Robert De Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1852–1864, 2020. 1

[2] 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

[3] 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

[4] Alberto Broggi, Paolo Grisleri, and Paolo Zani. Sensors technologies for intelligent vehicles perception systems: A comparison between vision and 3d-lidar. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 887–892. IEEE, 2013. 1

[5] 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. 5, 6

[6] Xinyu Cai, Wentao Jiang, Runsheng Xu, Wenquan Zhao, Jiaqi Ma, Si Liu, and Yikang Li. Analyzing infrastructure lidar placement with realistic lidar simulation library. *arXiv preprint arXiv:2211.15975*, 2022. 2

[7] 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

[8] Raul David Dominguez Sanchez, Xavier Diaz Ortiz, Xingcheng Zhou, Max Peter Ronecker, Michael Karner, Daniel Watzenig, and Alois Knoll. Lidar-guided monocular 3d object detection for long-range railway monitoring. *arXiv e-prints*, pages arXiv–2504, 2025. 1

[9] 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

[10] Mariella Dreissig, Dominik Scheuble, Florian Piewak, and Joschka Boedecker. Survey on lidar perception in adverse weather conditions. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8. IEEE, 2023. 1

[11] Siqi Fan, Zhe Wang, Xiaoliang Huo, Yan Wang, and Jingjing Liu. Calibration-free bev representation for infrastructure perception. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9008–9013, 2023. 2

[12] Ahmed Ghita, Bjørk Antoniussen, Walter Zimmer, Ross Greer, Christian Creß, Andreas Møgelmose, Mohan M Trivedi, and Alois C Knoll. Activeanno3d-an active learning framework for multi-modal 3d object detection. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1699–1706. IEEE, 2024. 1

[13] 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, 5, 6

[14] 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

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

[16] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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, 7, 8

[23] 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. 2, 5, 6

[24] 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

[25] 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, 4, 6, 7, 8

[26] Ciyun Lin, Yuying Wang, Bowen Gong, Hui Liu, and Hongchao Liu. Roadside lidar deployment optimization for vehicle-infrastructure cooperative perception in urban occlusion environments. *IEEE Transactions on Instrumentation and Measurement*, 2025. 2

[27] 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, 7, 8

[28] 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

[29] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 1

[30] 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

[31] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. *arXiv preprint arXiv:2211.07214*, 2022. 2

[32] 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, 7, 8

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

[34] 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. 1, 2

[35] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin 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*, pages 2446–2454, 2020. 5

[36] 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

[37] 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

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 4

[39] 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, 7, 8

[40] Zhe Wang, Siqi Fan, Xiaoliang Huo, Tongda Xu, Yan Wang, Jingjing Liu, Yilun Chen, and Ya-Qin Zhang. Vimi: Vehicle-infrastructure multi-view intermediate fusion for camera-based 3d object detection. *arXiv preprint arXiv:2303.10975*, 2023. 2

[41] 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. 6

[42] 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. 1, 5

[43] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 5, 6

[44] 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. 1, 2

[45] 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

[46] 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*, 2025. 2

[47] 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. *arXiv preprint arXiv:2411.14927*, 2024. 2

[48] 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, 6

[49] Guizhen Yu, Han Li, Yunpeng Wang, Peng Chen, and Bin Zhou. A review on cooperative perception and control supported infrastructure-vehicle system. *Green Energy and Intelligent Transportation*, 1(3):100023, 2022. 1

[50] 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, 6

[51] 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

[52] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 1

[53] 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

[54] 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

[55] Tianya Zhang, Lei Cheng, Tam Bang, Lihao Guo, Mustafa Hajij, Siyang Cao, Austin Harris, and Mina Sartipi. Roadside sensor systems for vulnerable road user protection: A review of methods and applications. *IEEE Access*, 2025. 1

[56] Zhaoliang Zheng, Yun Zhang, Zongling Meng, Johnson Liu, Xin Xia, and Jiaqi Ma. Inspe: Rapid evaluation of heterogeneous multi-modal infrastructure sensor placement, 2025. 5

[57] Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, et al. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. *arXiv preprint arXiv:2412.01812*, 2024. 1, 6

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 4

[59] 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

[60] Walter Zimmer, Marcus Grabler, and Alois Knoll. Real-time and robust 3d object detection within roadside lidars using domain adaptation. *arXiv preprint arXiv:2204.00132*, 2022. 2

[61] 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. 1, 2

[62] 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