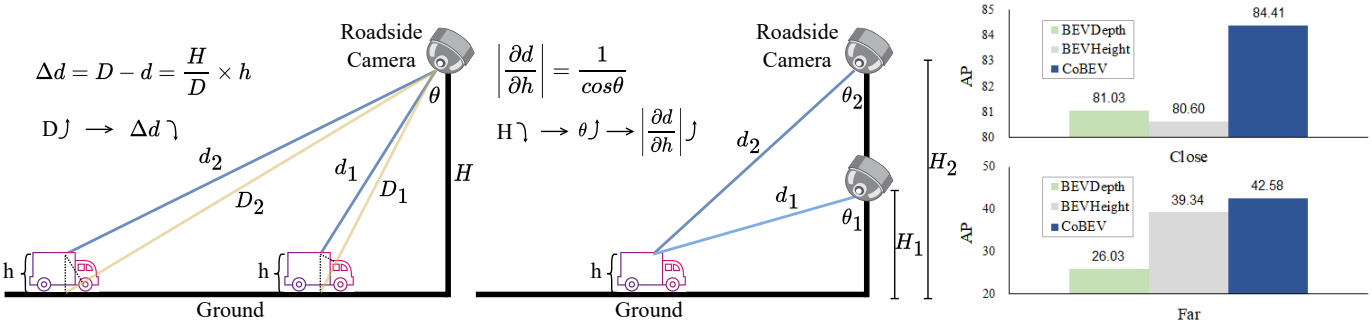
CoBEV: Elevating Roadside 3D Object Detection with Depth and Height Complementarity

Hao Shi†, Chengshan Pang†, Jiaming Zhang†, Kailun Yang*, Yuhao Wu, Huajian Ni, Yining Lin, Rainer Stiefelhagen, and Kaiwei Wang*



(a) Depth-based deficiency **(b) Height-based deficiency** **(c) CoBEV performance**

Fig. 1: (a) As the distance (d) between the vehicle and the roadside camera increases, the depth difference (Δd) between the vehicle and the ground decreases, leading to the unreliability of the depth-based lifting acquisition of the BEV features. (b) As the camera's height (H) decreases, the depth's partial differential with respect to the height ($|\frac{\partial d}{\partial h}|$) of the vehicle increases, leading to the unreliability of the height-based BEV features. (c) Vehicle monocular 3D detection results are presented for both close (range interval $\sim 20m$) and far (range interval $\sim$ 100m) scenarios on the DAIR-V2X-I validation set [1]. Our CoBEV leverages the fusion of complementary BEV features derived from both height and depth information, consistently achieving state-of-the-art performance across target-camera distances.

***Abstract*—Roadside camera-driven 3D object detection is a crucial task in intelligent transportation systems, which extends the perception range beyond the limitations of vision-centric vehicles and enhances road safety. While previous studies have limitations in using only depth or height information, we find both depth and height matter and they are in fact complementary. The depth feature encompasses precise geometric cues, whereas the height feature is primarily focused on distinguishing between various categories of height intervals, essentially providing semantic context. This insight motivates the development of Complementary-BEV (CoBEV), a novel end-to-end monocular 3D object detection framework that integrates depth and height to construct robust BEV representations. In essence, CoBEV estimates each pixel's depth and height distribution and lifts the camera features into 3D space for lateral fusion using the newly proposed two-stage complementary feature selection (CFS) module. A BEV feature distillation framework is also seamlessly integrated to further enhance the detection accuracy from the prior knowledge of the fusion-modal CoBEV teacher. We conduct extensive experiments on the public 3D detection benchmarks of roadside camera-based DAIR-V2X-I and Rope3D, as well as the private Supremind-Road dataset, demonstrating that CoBEV not only achieves the accuracy of the new state-of-the-art, but also significantly advances the robustness of previous methods in challenging long-distance scenarios and noisy camera disturbance, and enhances generalization by a large margin in heterologous settings with drastic changes in scene and camera parameters. For the first time, the vehicle AP score of a camera model reaches 80% on DAIR-V2X-I in terms of easy mode. The source code will be made publicly available at CoBEV.**

***Index Terms*—Roadside 3D object detection, BEV scene understanding, autonomous driving.**

I. Introduction

VEHICLE-CENTRIC 3D object detection has made significant strides in recent years [2], [3], [4], [5], [6], [7], [8]. However, it still faces significant challenges with regard to the occlusion of blind spots and limited long-distance perception capability due to the restricted observation of sensors mounted atop vehicles. Conversely, infrastructure-centric 3D object detection exhibits a natural ability to capture long-range information, and is less susceptible to occlusion by vehicles, making it a crucial technology for the realization of safer and smarter transportation systems. While LiDAR-based detectors located on the vehicle side continue to outperform camera-based detectors in terms of accuracy performance, camera-based methods have gained increased attention in roadside perception schemes due to their higher flexibility to adapt to existing traffic infrastructure, their lower cost, and ease of maintenance, compared to the expensive LiDAR alternatives [9], [10].

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 12174341, in part by Shanghai SUPREMIND Technology Company Ltd., and in part by Hangzhou SurImage Technology Company Ltd.

H. Shi and K. Wang are with the State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University, Hangzhou 310027, China (email: haoshi@zju.edu.cn; wangkaiwei@zju.edu.cn).

J. Zhang and R. Stiefelhagen are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (email: jiaming.zhang@kit.edu; rainer.stiefelhagen@kit.edu).

J. Zhang is also with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK.

K. Yang is with the School of Robotics, Hunan University, Changsha 410012, China (email: kailun.yang@hnu.edu.cn).

K. Yang is also with the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.

H. Shi, C. Pang, Y. Wu, H. Ni, and Y. Lin are with Shanghai SUPREMIND Technology Co., Ltd, Shanghai 201210, China.

*corresponding authors: Kaiwei Wang and Kailun Yang.

†denotes equal contribution.

In camera-based methods, the Bird's-Eye-View (BEV) space is extensively utilized because it not only encodes rich spatial context information but also provides a unified global-view feature space that is advantageous for the integration of various perception tasks such as 3D object detection [11], [12], [13] and semantic segmentation [14], [15], [16], mapping [17], [18], *etc*. The BEV space also facilitates the promotion of Vehicle-Infrastructure Cooperative Autonomous Driving (VIC-AD) [1], [19]. As shown in previous work [5], [6], [7], the primary challenge for BEV detection is how to "lift" the perspective data of the camera to the BEV space. Previous state-of-the-art detection schemes typically rely on explicitly estimating the depth or height distribution of pixels [7], [10] and utilizing camera parameters to perform feature projection.

While vehicle-side perception has made remarkable progress recently [6], [7], [20], roadside camera-based detectors face a unique set of challenges: (1) Roadside cameras typically operate individually with a monocular setup, making it difficult to integrate multi-view images to compute accurate depth. (2) The observation distance of the roadside camera is much longer than the vehicle view, as depicted in Fig. 1(a), making it difficult for a depth-based detector that operates effectively at the vehicle end, to distinguish between distant vehicles and road surface features. (3) The camera parameters of the roadside camera can be various. For instance, the camera located on the vehicle is usually parallel to the direction of motion and has a constant installation height, while the extrinsic parameters of the roadside camera at different intersections vary widely. As shown in Fig. 1(b), we observe that the differential of depth calculated from height increases for previous height-based detectors, as the camera mounting height decreases, implying that inaccurate height estimates will introduce larger errors in object detection.

On the other hand, we find that the features of depth and height detectors exhibit complementary properties. As shown in Fig. 1(c), in long-distance scenes that are challenging for depth-based methods, it is more precise to regress height to calculate BEV features, due to the constant height of vehicles. Conversely, when the distance is closer, directly regressing depth becomes comparatively easier. In terms of feature encoding, we argue that depth detectors rely more on precise geometric cues, whereas height detectors learn distinct height distribution intervals across different categories, thus placing greater emphasis on the semantic context of targets. We substantiate this observation through ablation in Tab. VIII.

Based on the above observations, we hereby propose **Co**mplementary-**BEV** (**CoBEV**), a roadside monocular 3D object detection framework that seamlessly integrates the complementary depth and height cues to establish robust BEV representations for elevating traffic scene understanding. Specifically, we introduce a Camera-aware Hybrid Lifting (CHL) module (Sec. III-B) to independently estimate the depth and height distribution for each pixel. This module lifts camera features into the BEV space by deriving a unified context. Subsequently, the heterogeneous 3D features undergo partial-pillar voxel pooling, reducing computational complexity while preserving the free flow of information in the height dimension. These condensed heterogeneous features are then put into the novel Complementary Feature Selection (CFS) module (Sec. III-C), where global pattern and local cues are selectively fused to generate a robust BEV representation. Furthermore, we present a new BEV Feature Distillation framework (Sec. III-D) capable of enhancing detection accuracy agnostic to various target sizes, without introducing additional computational complexity. Thanks to these core designs, our CoBEV achieves state-of-the-art performance, as shown in Fig. 1(c). Considering the diversity and variability of the infrastructure scenarios, generalization ability is crucial for the improvement of intelligent transportation systems. Therefore, we first present a thorough analysis of how the BEV representation quality affects the mono3D object detection's robustness and generalization across heterologous roadside datasets and camera parameters disturbance.

To validate the power of the proposed CoBEV, we conduct extensive experiments on two public roadside datasets DAIR-V2X-I [1] and Rope3D [9], and a private dataset Supremind-Road. CoBEV achieves the new state-of-the-art performance of 69.57% / 47.21% / 66.17% in 3D Average Precision ($AP_{3D|R40}$) across Vehicle, Pedestrian and Cyclist on the DAIR-V2X-I validation set, outperforms the second-best BEVHeight [10] with a large margin of 3.80% / 7.92% / 6.09%, respectively. On the Rope3D, CoBEV surpasses all other methods with $AP_{3D|R40}$ of 52.72% / 29.28% for Car and Big Vehicle detection under the difficult $IoU = 0.7$ setting, perform a significant improvement of 6.99% / 6.21% compared with the previous best-published method. In the real world, the vision capture and camera parameters of different intersections vary greatly, and it is impractical and unrealistic to collect and annotate data for every intersection in each city. Therefore, we also compare the robustness of different BEV representations by studying heterologous settings and camera parameters disturbance and observe a significant drop in the accuracy of the previous state-of-the-art approach. Compared with contemporary methods [10], CoBEV achieves an average 1.61% improvement on the challenge Rope3D heterologous setting across all categories, and 2.56% improvement on Supremind-Road heterologous setting, thanks to the complementary robust BEV representations, which further demonstrates the effectiveness of the proposed CoBEV solution. In scenarios where camera parameters are directly perturbed by noise affecting focal length, roll, and pitch, CoBEV exhibits an enhanced detection capability, surpassing BEVHeight by an average of 3.93%. This once again demonstrates CoBEV's notable resistance to interference and robustness.

In summary, we deliver the following contributions:

- We present a comprehensive theoretical examination of the shortcomings of prior approaches that exclusively rely on either depth or height for constructing BEV features in infrastructure-centric environments.
- We propose CoBEV, a novel roadside monocular 3D detection framework that generates robust BEV representations by seamlessly incorporating complementary geometry-centric depth and semantic-centric height cues.
- An in-depth study is conducted to examine the impacts of various feature fusion methods on constructing robust BEV features through extensive ablation experiments.

- The superior generalization ability of CoBEV in diverse and heterologous settings, as well as its performance on noisy camera parameters, proves its suitability for roadside scenarios for elevating traffic scene understanding.
- Extensive experiments on two public and one private dataset demonstrate that CoBEV achieves state-of-the-art detection accuracy, establishing it as a robust and reliable solution for roadside 3D detection.

II. Related Work

A. Camera-based 3D Object Detection

Camera-based 3D object detection aims to predict the 3D bounding box of the object from an image. According to the deployment scenarios, they can be divided into two types: vehicle-centric and infrastructure-centric methods.

Vehicle-centric methods. Vehicle-centric methods have been extensively explored in previous works. One branch of methods directly uses 2D detectors and slightly modifies the detection head to achieve 3D detection [21], [22], [23], [24]. FCOS3D [25] adapts 2D detectors by predicting both 2D and 3D attributes. DETR3D [26] introduces 3D object queries to directly regress 3D bounding boxes. Building upon this work, PETR [27] introduces 3D position-aware representation to further improve accuracy. Another branch of vehicle-centric methods performs detection directly in 3D feature space. OFT [28] is the first to use the orthographic feature transform to transform 2D image features into 3D space for monocular 3D detection. LSS [5] achieves image-based 3D detection by regressing the depth distribution to transform image features into BEV space. ImVoxelNet [29] projects a voxel grid into the image feature space to generate a voxel representation. BEVDepth [7] introduces the depth of LiDAR for supervision on the basis of LSS, which improves the accuracy of depth estimation and achieves state-of-the-art detection accuracy. CrossDTR [30] builds depth-aware embedding from depth estimation to improve performance using transformer backbones. Different from vehicle-centric works, we focus on infrastructure-centric roadside-camera-driven 3D object detection for elevating traffic scene understanding.

Infrastructure-centric methods. While roadside devices have a wider range of perception than vehicle end and can achieve long-term observation, it has been under-explored in the literature. Recently, V2X-SIM [19] is the first to use the CARLA simulator [31] to collect a public vehicle-road collaboration dataset. DAIR-V2X [1] and Rope3D [9] have also released vehicle-road perception data in the real scene, establishing a benchmark for roadside detection. Yet, the accuracy of state-of-the-art vehicle detection algorithms [7] at the roadside is limited [10]. To address this issue, BEVHeight [10] first focuses on roadside detection and proposes predicting the height distribution of the scene instead of the depth distribution, which improves the performance. In contrast, CBR [32] maps image features to BEV space based on Multi-Layer Perceptrons (MLPs), bypassing the step of extrinsic calibration, but the accuracy is limited. Different from these methods, we propose CoBEV, which fully considers the challenges of long-distance targets and diverse camera parameters unique to roadside 3D detection for elevating perception performance and robustness by seamlessly fusing geometry-centric depth and semantic-centric height cues.

B. BEV Scene Understanding

The concept of Bird's-Eye-View (BEV) perception has gained increasing attention due to its ability to provide a unified feature space and rich spatial context, making it particularly well-suited for traffic scenarios. A crucial question in this field is how to construct BEV features from the images captured by pinhole cameras. The existing methods can be categorized into two categories: implicit and explicit methods.

Implicit lifting methods. The implicit methods for constructing BEV features utilize MLPs or transformers. The MLP-based approach offers a straightforward mapping strategy where image features are fully connected to BEV features. The VPN [15] initially utilizes the global mapping capability of MLP to map the multi-view camera input to the top-down feature map for semantic segmentation. Subsequent methods [33], [34], [35] also employ MLP for view transformation to implement road layout estimation, semantic segmentation, *etc.* In contrast, transformer-based methods define BEV grid queries or object queries and perform cross attention with image features to construct BEV representations [6], [26], [27], [36], [37]. The most well-known of these methods, BEVFormer [6], employs spatial cross-attention for perspective transformation and considers feature aggregation based on temporal attention. However, these implicit methods do not fully consider the corresponding camera parameters, which limits their generalization to diverse roadside scenarios.

Explicit lifting method. The explicit approach for constructing BEV features is based on geometric principles. The earliest Inverse Perspective Mapping (IPM) [38] is a straightforward method for achieving BEV perception, which utilizes horizontal plane constraints and physical mapping for perspective transformation. However, directly applying IPM transformation on images is hindered by sampling errors and quantization effects. On the other hand, a cluster of methods attempts to incorporate depth estimation to mitigate the ill-posed issues associated with monocular 3D perception. The Pseudo LiDAR [39] pioneers to employ off-the-shelf depth estimation techniques to convert image pixels into pseudo-LiDAR-point-clouds, which can then be processed in 3D space. Recently, LSS [5] stands out as the dominant paradigm, where depth or height distributions in 3D space are explicitly estimated and camera parameters are utilized for feature projection [7], [10], [16], [17], [18], [40], [41] Due to the incorporation of camera parameters, explicit methods are more suitable for BEV feature generation in roadside scenes. In this work, depth and height are leveraged to achieve view transformation, and their distinctions and complementarities are thoroughly considered to extract robust BEV features from monocular cameras. Consequently, satisfactory monocular 3D detection can be achieved in diverse heterologous intersection data that is not included in the training set, as well as coping with the noisy camera parameters.

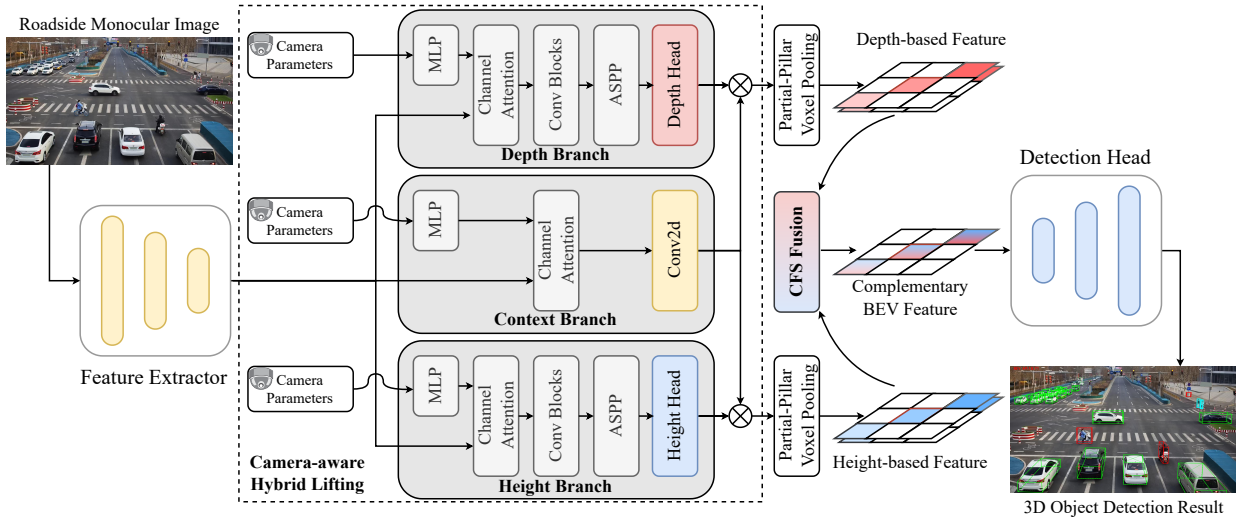


Fig. 2: *Overview of the Complementary-BEV (CoBEV) architecture.* Firstly, the monocular image on the roadside is fed into the feature extractor to encode high-dimensional features. Image features are then sent to the Camera-aware Hybrid Lifting (CHL) module that consists of a depth branch, a context branch, and a height branch, and fused with the camera parameters encoded by the MLP. The pixel distribution of the depth and height branches are integrated with the context feature via an outer product to obtain a frustum-shaped point cloud. This point cloud is then splatted to the depth-based and height-based compressed 3D features by partial-pillar voxel pooling. Finally, the multi-source BEV features are fused via the Complementary Feature Selection (CFS) module to develop robust BEV features for 3D object detection.

III. CoBEV: Proposed Framework

In this section, the problem formulation of the roadside monocular 3D object detection is first provided (Sec. III-A). Next, the core designs of CoBEV are described in detail, including the *Camera-aware Hybrid Lifting* module (Sec. III-B), *Complementary Feature Selection (CFS)* module (Sec. III-C), and the *BEV Feature Distillation* (Sec. III-D).

A. Problem Formulation

In this work, we aim to build a robust roadside 3D monocular detector:

$$\mathbf{B_{ego}} = Det(X, E, I|\phi), \tag{1}$$

where $Det(\cdot)$ is the detector model, ϕ is the learned parameters, $X \in \mathbb{R}^{3\times H\times W}$ is the input single frame capture of the infrastructure camera, $E \in \mathbb{R}^{3\times 4}$ and $I \in \mathbb{R}^{3\times 3}$ are the extrinsic and intrinsic matrix of the image X, respectively. H and W denote the input image's height and width. $\mathbf{B_{ego}}$ is a series of output 3D bounding box:

$$\mathbf{B_{ego}} = \{\hat{B}_1, \hat{B}_2, ..., \hat{B}_n\}, \tag{2}$$

where n is the number of foreground targets and $\hat{B}_i$ is a vector with 7 degrees of prospects, which are composed of position, dimension, and orientation:

$$\hat{B}_i = (x, y, z, l, w, h, \theta), \tag{3}$$

where (x, y, z) is the position of each 3D bounding box, (l, w, h) is the three-dimensional size of its corresponding cuboid, and θ is the yaw angle relative to the ground plane.

For roadside cameras, the extrinsic and intrinsic parameters can be obtained through calibration. Considering the ownership of data between different camera devices and the privacy regulation, we focus on the 3D detection of single-frame monocular cameras, which will cover a vast majority of intersection scenes and have high potential practical value for the construction of intelligent transportation systems.

B. Camera-aware Hybrid Lifting

Lifting 2D features to 3D space is crucial for BEV perception [5], [7], [10]. As shown in Fig. 1, by analyzing the variation in depth distribution and height distribution on the roadside, we observe that the error increases when the distance between the vehicle and the facility increases based on depth lifting, and the error in depth estimation increases when the height decreases based on height lifting. In other words, the geometric representations of depth and height are actually complementary in the transportation infrastructure scene as they are mutually beneficial in terms of accuracy. This is attributed to the fact that 1) when depth is ambiguous, the height of the camera and the car remain constant, 2) as the camera height decreases, the depth itself decreases, making it easier to estimate the geometry directly, and, 3) depth information encodes precise geometric cues, whereas height distributions better capture distinct classes of semantic context.

Motivated by these key observations, we introduce the *Camera-aware Hybrid Lifting (CHL)* to obtain multi-source 3D features leveraging both depth and height information. As shown in Fig. 2, this novel module is distinctly compartmentalized into three branches, including the depth branch, the context branch, and the height branch. Camera parameters $\{E, I\}$ are first encoded as camera-aware features F_{cam} by three individual Multi-Layer Perceptron (MLP):

$$F_{cam} = MLP(\xi(E) \oplus \xi(I)). \tag{4}$$

To elucidate further, both the depth D and height H distributions involve a discretization process. The depth is discretized by performing uniform discretization (UD), which entails partitioning it into a pre-defined number of bins N_D and the certain range of depth $(D_0 \sim D_{max})$:

$$\begin{cases} \delta D = \dfrac{D_{max} - D_0}{N_D}, \\ D_i = i \cdot \delta D, \end{cases} \tag{5}$$

where the index $i \in [1, N_D]$. Considering the variability in camera height fluctuations observed within different infrastructure facilities, the number of height bins N_H is also fixed, yet the height range represented by each bin is subject to dynamic adjustments in accordance with the specific scene:

$$H_j = \lfloor H_0 + (\frac{j}{N_H})^{\alpha} \cdot (H_{max} - H_0) \rfloor, \tag{6}$$

where the index $j \in [1, N_H]$, α denotes the hyperparameter that controls the dynamic expansion of each height bin. We empirically set $\alpha = 1.5$ in all experiments.

Given the inherent similarity of the regression tasks, the depth branch and the height branch exhibit a completely symmetric structure, and the prediction of depth and height distribution can be represented via the following formula:

$$\begin{cases} D_i^{pred} = & \psi_d(CA_d(F|F_{cam})), \\ H_j^{pred} = & \psi_h(CA_h(F|F_{cam})), \end{cases} \tag{7}$$

where $\psi_d(\cdot)$ and $\psi_h(\cdot)$ represent the height- and the depth network, respectively. $CA(\cdot)$ denotes the channel attention layer. $D_i^{pred} \in \mathbb{R}^{N_D \times \frac{H}{16} \times \frac{W}{16}}$ and $H_j^{pred} \in \mathbb{R}^{N_H \times \frac{H}{16} \times \frac{W}{16}}$ are the predicted depth and height distribution along the entire ray of viewpoint.

Upon obtaining the pixel-wise distributions of depth and height, the subsequent step in our methodology involves the projection of image features into the volumetric space of the encompassing frustum-shaped structure. To commence this process, we initiate with the computation of spatial context information $F_{context}$ derived from the image features F, which can be represented as follows:

$$F_{context} = Conv_{2d}(CA_c(F|F_{cam})), \tag{8}$$

where $Conv_{2d}(\cdot)$ denotes the 2D convolutional block. Subsequently, we employ the outer product operation to derive the three-dimensional feature representations, yielding:

$$\begin{cases} F_{depth}^{3d} & = F_{context} \otimes D_{pred}, \\ F_{height}^{3d} & = F_{context} \otimes H_{pred}, \end{cases} \tag{9}$$

where $\otimes$ represents the outer product operation. This intricate procedure facilitates the transformation of image-based data into a comprehensive 3D context, thereby enhancing the depth and height information integration within our framework.

Subsequently, the two heterologous three-dimensional volume $F_{depth}^{3d} \in \mathbb{R}^{N_D \times C \times \frac{H}{16} \times \frac{W}{16}}$ and $F_{height}^{3d} \in \mathbb{R}^{N_H \times C \times \frac{H}{16} \times \frac{W}{16}}$ undergo a meticulous transformation into the unified ego coordinate system. This ego coordinate system is meticulously centered on the ground plane, a visual representation of depth-based and height-based coordinate transformation is illustrated in Fig. 3. For the transformation of depth-based features, as shown in Fig. 3(a), a direct conversion is executed utilizing both intrinsic and extrinsic matrices:

$$P_{ego}^{depth} = RI^{-1}[ud, vd, d]^T + t, \tag{10}$$

where each 2D point on the image plane is represent with $p_{img} = [u, v, 1]^T$, u and v are the pixel index. d denotes the corresponding depth. R and t is the rotation matrix and translation matrix with $E = [R, t]$. In contrast, when handling

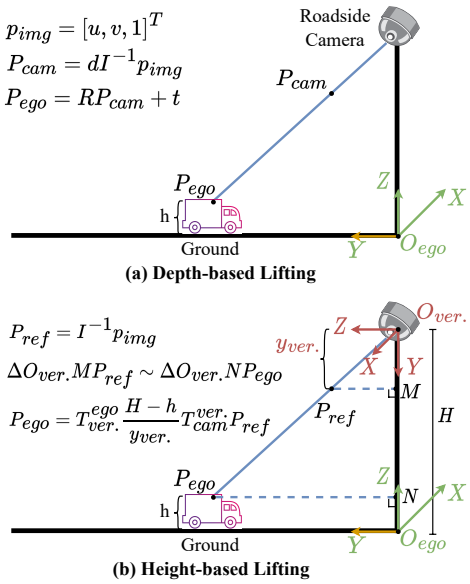


Fig. 3: *Camera-based Hybrid Lifting* includes (a) Explicit lifting based on the depth distribution and camera parameters. (b) Explicit lifting with similar triangles based on the height distribution and camera parameters.

height-based features, a two-step process is implemented as shown in Fig. 3(b). Initially, pixels p_{img} from the image I are transformed into the normalized coordinate system of the camera, *i.e.*, the reference plane:

$$P_{ref}^{height} = I^{-1}[u, v, 1]^T. \tag{11}$$

Subsequently, a vertical coordinate system is employed to compute the indices of the ego point P_{ego}, whose origin $O_{ver.}$ is the camera optical center, and its Y axis is perpendicular to the ground. This calculation is performed by constructing a pair of similar triangles:

$$\Delta O_{ver.}MP_{ref} \sim \Delta O_{ver.}NP_{ego}. \tag{12}$$

Concretely, the ego point can be obtained by:

$$P_{ego}^{height} = T_{ver.}^{ego} \frac{H - h}{y_{ver.}} T_{cam}^{ver.} I^{-1}[u, v, 1]^T, \tag{13}$$

where $T_{cam}^{ver.} \in \mathbb{SE}(3)$ is the transformation matrix from camera coordinate to vertical coordinate. $T_{ver.}^{ego} \in \mathbb{SE}(3)$ denotes the transformation matrix from the vertical coordinate to the uniformed ego coordinate system. $y_{ver.}$ represents the distance between P_{ref} and camera center point $O_{ver.}$ along the vertical coordinate's Y axis. This rigorous transformation process serves to align the volumetric data within a consistent ego coordinate system, thereby facilitating subsequent perception in a unified manifold.

C. Complementary Feature Selection

After the lifting process, two large 3D point clouds unified in the ego coordinate system can be obtained. To save subsequent calculations, a partial-pillar voxel pooling operation is performed on the point cloud. In the context of PointPillars [42], a 'pillar' refers to an infinite height voxel grid. Different from previous approaches [5], [7], [10], we segment the pillar into several partial pillar-shape voxels,

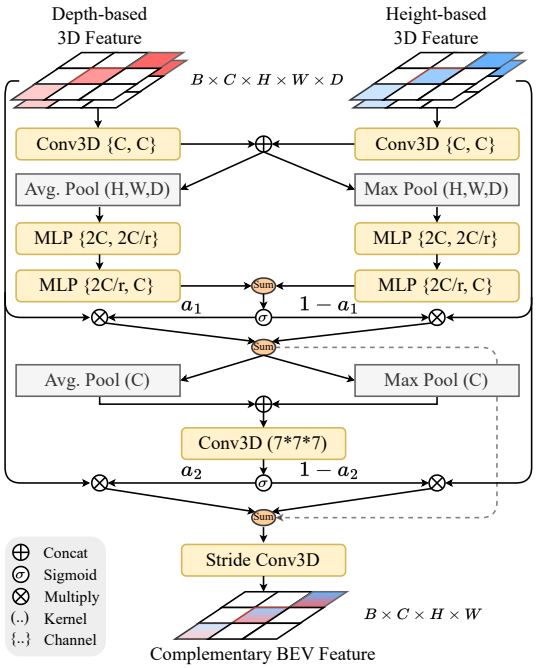


Fig. 4: *The proposed Complementary Feature Selection (CFS) module.* Different from previous works [10], we maintain a low-scale vertical axis of depth and height BEV features after voxel pooling to promote information flow in the feature fusion process. CFS consists of two cascade feature selection processes, with the first stage for selecting complementary features in the column-shape channels, and the second stage for selecting features in the BEV plane, which are ultimately compressed to two-dimensional complementary BEV features through the stride 3D convolutional compression.

thereby preserving the height dimension after voxel pooling rather than discarding it entirely:

$$\hat{F}^{3d}(\hat{p},i,j,k)=\sum_{p=r\cdot\hat{p}}^{r\cdot(\hat{p}+1)}F^{3d}(p,i,j,k), \tag{14}$$

where $F^{3d}\in\mathbb{R}^{D\times C\times\frac{H}{16}\times\frac{W}{16}}$ denotes the large 3D point cloud. $\hat{F}^{3d}\in\mathbb{R}^{\frac{D}{r}\times C\times\frac{H}{16}\times\frac{W}{16}}$ represents the output compressed feature with reduction factor r. This design aims to ensure that visual cues related to detection can freely flow in all dimensions during the subsequent feature selection process, thereby achieving task-centric fusion features. Furthermore, preserving the height axis also mitigates the information bottleneck before and after fusion operations.

Next, the Complementary Feature Selection (CFS) module will be discussed in detail. The key idea is to select and construct the most relevant BEV features for the detection task from two heterogeneous compressed 3D features.

As shown in Fig. 4, in the first stage of CFS, we concatenate the depth and height 3D features in parallel, and activate the grid after spatial pooling to obtain the affinity of first-stage selection a_1 and complete the first feature selection:

$$g=\mathcal{W}_2(\mathcal{W}_1(\mathcal{P}_{3d}(\hat{F}^{3d}_{depth}\oplus\hat{F}^{3d}_{height}))), \tag{15}$$

where g denotes the global feature context. $\oplus$ is the concatenation. $\mathcal{P}_{3d}(\cdot)$ represents 3D global pooling layer. $\mathcal{W}_1$ is a dimension reduction layer, and $\mathcal{W}_2$ is a dimension increasing layer with the channel reduction ratio r, both of which are implemented with MLP. Then, we calculate the affinity a_1 of the first-stage feature selection:

$$a_1=\sigma(g_a\uplus g_m), \tag{16}$$

where $\uplus$ denotes the element-wise summation, $\sigma(\cdot)$ is the $Sigmoid$ function. g_a and g_m are the global feature context from 3D average pooling branch and 3D max pooling branch, respectively. The first-stage selection feature F^1_s can be obtained by:

$$F^1_s=a_1\odot\hat{F}^{3d}_{depth}+(1-a_1)\odot\hat{F}^{3d}_{height}, \tag{17}$$

where $\odot$ denotes an element-wise product. The first-stage selection operates on the global spatial manifold, which squeezes each feature map of size $H\times W\times D$ into a scalar. The fusion affinity a_1 consists of real numbers between 0 and 1, so are the $1-a_1$, which enable the module to conduct a soft selection between $\hat{F}^{3d}_{depth}$ and $\hat{F}^{3d}_{height}$ in a weighted averaging fashion. This operation emphasizes the large objects like cars and trucks that are distributed globally. However, the detection of small objects like pedestrians and cyclists is vital for safety-critical intelligent transportation systems. Therefore, we introduce the second-stage feature selection to aggregate fine-grained local features and alleviate the problem that ignoring small instances. Concretely, in the second stage of CFS, we perform channel pooling and activation on the fused features F^1_s to obtain the affinity a_2 in the three-dimensional space and complete the second feature selection:

$$a_2=\sigma(\mathcal{G}(\mathcal{P}_c(F^1_s))), \tag{18}$$

where $\mathcal{P}_c(\cdot)$ represents the pooling layer across channel dimension, $\mathcal{G}(\cdot)$ denotes the 3D convolution layer with kernel size $7\times7\times7$. The second-stage selection feature F^2_s can be obtained by:

$$F^2_s=a_2\odot\hat{F}^{3d}_{depth}+(1-a_2)\odot\hat{F}^{3d}_{height}, \tag{19}$$

The second-stage feature selection operates on every voxel, thus integrating fine-grained local features adaptively. Following this, the results of the two feature selections are added through skip connections. To construct the final complementary BEV features F^{bev}_{com}, a stride 3D convolution layer is applied to fuse them and reduce the dimension to a 2-dimensional space:

$$F^{bev}_{com}=\mathcal{G}_{fuse}(F^1_s+F^2_s), \tag{20}$$

where $\mathcal{G}_{fuse}(\cdot)$ denotes the output 3D convolution layer with stride on the height axis. Different from the first-stage selection, the second-stage selection focuses on 'where' is an informative voxel for the heterogeneous 3D features, therefore complementary to the first stage.

D. BEV Feature Distillation

To fully unleash the potential of the CoBEV model, we introduce the BEV Feature Distillation Framework to further explore the boundaries of monocular camera 3D roadside detection. The framework is demonstrated in Fig. 5. Leveraging the unified BEV feature representation, CoBEV

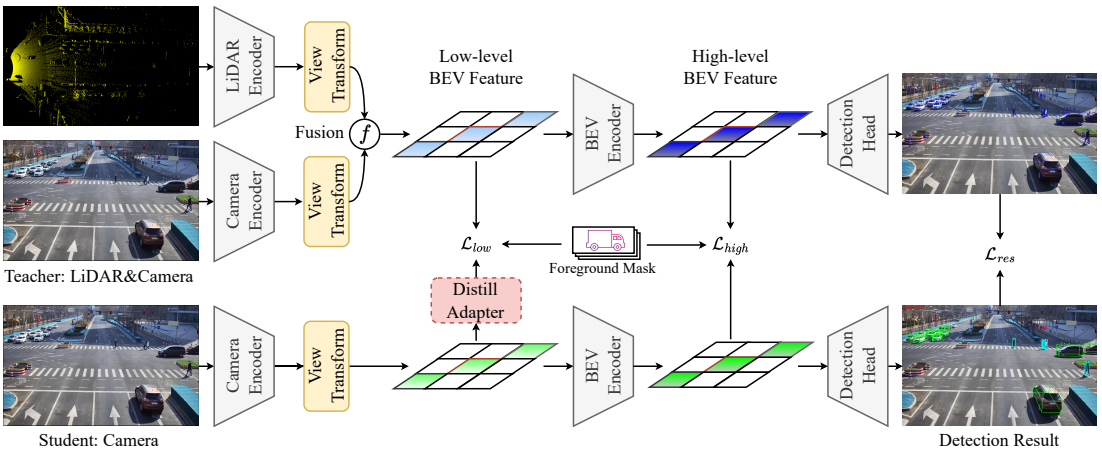


Fig. 5: *Illustration of the BEV Feature Distillation.* It employs a fusion-to-camera paradigm, aligning the student CoBEV detector with the LiDAR-camera fusion version of the CoBEV teacher across three stages: low-level BEV feature, high-level BEV feature, and the response. Different from previous work [43], we only apply supervision signal adaption at low-level features to emphasize valuable street-view structural knowledge of all categories at high-level features and therefore achieve performance improvement agnostic to the object size.

seamlessly implements the fusion-to-camera knowledge distillation paradigm, transitioning from the LiDAR-Camera Fusion modality to the Camera modality. Specifically, the fusion model, serving as the teacher, takes point clouds and image frames as input to provide rigorous supervision to the student CoBEV, across three levels: the low-level BEV features, the high-level BEV features, and the response. The key difference between the teacher and the student model lies in the sparse LiDAR encoder [44] and the fusion layer for point clouds and image features, which simply consists of two layers of 2D convolution. To ensure effective alignment, we strive to maintain as much structural consistency as possible between the teacher and the student models.

Low-level BEV features encapsulate valuable semantic knowledge. Therefore, during feature distillation at this stage, our focus primarily centers on imitating features of large objects, rather than small objects. Differing from [43], we adopt a straightforward MSE loss based on empirical experimentation to guide this process. Considering the inherent data disparities between sparse LiDAR and dense camera captures, we limit the supervision to task-relevant foreground regions to avoid mimicking irrelevant 3D features in background areas. To achieve this, we employ the bounding box as the center to generate a Gaussian Mask M, assigning weights to the loss in accordance with [45]. In light of occasional instances where the performance of the teacher model may lag behind that of the student model, we introduce an additional Distill Adapter during training to fine-tune the supervision signals. Importantly, this adapter can be discarded during testing, thus incurring no additional inference resource overhead. The BEV distillation loss at this stage can be expressed as:

$$\mathcal{L}_{low} = ||M \cdot (F^{low}_{teacher} - \mathcal{G}_{adapt}(F^{low}_{student}))||_2, \tag{21}$$

where $\mathcal{G}_{adapt}(\cdot)$ denotes the distill adapter that consists of three 2D convolutional layers with ReLU activation. The BEV convolutional encoder is realized through three stacked 2D residual blocks. The resultant high-level BEV features effectively encode the structural knowledge of the outdoor scene. Distillation at this stage grants the network an expanded capability to facilitate cross-modality alignment, while concurrently enabling further mimic of all categories of information irrelevant to the object size. In contrast to the prior work [43], we refrain from introducing an adapter at this stage because our objective is to maximize the alignment and refinement of all target features, especially those associated with smaller targets. The high-level BEV distillation loss can be expressed as follows:

$$\mathcal{L}_{high} = ||M \cdot (F^{high}_{teacher} - F^{high}_{student})||_2. \tag{22}$$

To align the final output of the student closely with that of the fusion detector, we incorporate additional supervision for both bounding box regression and classification at the response level. In this context, soft labels are employed instead of hard labels, as soft labels have the capacity to convey more informative nuances. Furthermore, soft labels leverage the high confidence associated with definitive outcomes in the teacher model, while attributing lower confidence to uncertain results, effectively serving as a natural filtering mechanism. Following [46], the response loss can be further dissected into two components: regression loss and classification loss:

$$\begin{aligned} \mathcal{L}_{res} &= \mathcal{L}_{reg} + \mathcal{L}_{cls} \\ &= s \times (SmoothL1(b_t, b_s) + QFL(c_t, c_s)), \end{aligned} \tag{23}$$

where s represents the IoU confidence score of the soft label bounding box given by the teacher. $SmoothL1(\cdot)$ is the Smooth L1 loss. b_t and b_s are the bounding box parameters of the soft label and prediction. c_t and c_s denote the classification parameters of the teacher and student. $QFL(\cdot)$ represents quality focal loss [47].

IV. Experiments

A. Datasets

DAIR-V2X-I. DAIR-V2X [1] is a real-world vehicle-infrastructure collaborative dataset, offering a multi-modal object detection resource within the context of intersection scenes. We focus on the roadside subset, denoted as DAIR-V2X-I. Comprising $10k$ images and corresponding LiDAR point clouds, this subset encompasses a total of $493k$ 3D

TABLE I: Comparison with the state-of-the-art on the DAIR-V2X-I validation set [1].

Method	Modality	Venue	Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
			Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
PointPillars [42]	LiDAR	CVPR' 19	63.07	54.00	54.01	38.53	37.20	37.28	38.46	22.60	22.49
SECOND [44]	LiDAR	Sensors	71.47	53.99	54.00	55.16	52.49	52.52	54.68	31.05	31.19
MVXNet [48]	LiDAR & Camera	ICRA' 19	71.04	53.71	53.76	55.83	54.45	54.40	54.05	30.79	31.06
ImVoxelNet [29]	Camera	WACV' 22	44.78	37.58	37.55	6.81	6.75	6.74	21.06	13.57	13.17
BEVFormer [6]	Camera	ECCV' 22	61.37	50.73	50.73	16.89	15.82	15.95	22.16	22.13	22.06
BEVDepth [7]	Camera	AAAI' 23	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
BEVHeight [10]	Camera	CVPR' 23	77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
CoBEV (Ours)	Camera	-	81.20	68.86	68.99	44.23	42.31	42.55	61.28	61.00	61.61
CoBEV$_{full}$ (Ours)	Camera	-	82.01	69.57	69.66	**49.32**	**47.21**	**47.48**	**66.13**	**66.17**	**66.69**
w.r.t. BEVHeight			+4.23	+3.80	+3.81	+8.10	+7.92	+8.02	+5.90	+6.09	+6.15
CoBEV* (Ours)	Camera	-	**82.52**	**81.48**	**81.60**	44.72	42.88	43.07	60.51	62.54	63.31

– Note: $full$ denotes using our BEV distillation. * denotes additionally covering the longer range between $100\sim200m$, while others only cover $0\sim100m$.

box annotations, spanning distances from 0 to 200 meters, distributed across ten distinct categories. Following previous work [10], we partition DAIR-V2X-I into a training set (50%) and a validation set (20%) to facilitate comparative analyses. Additionally, it is worth noting that the testing examples (30%) are not yet publicly disclosed.

Rope3D. Rope3D [9] is a large-scale multi-modal 3D detection dataset capturing roadside scenes, comprising a collection of $50k$ camera and LiDAR captures. This dataset offers annotations for 12 distinct categories within the $0-200m$ range, amounting to a substantial total of $1.5M$ 3D box annotations. Rope3D contains a diverse array of complex traffic scenes, including 26 different intersection scenarios, spanning conditions such as rainy days, nights, and dawn scenes. In accordance with the original partitioning strategy detailed in the Rope3D paper [9], we adopt a training-test split that allocates 70% of the images for training and reserves the remaining 30% for testing. Notably, as the LiDAR captures are not publicly accessible, we have refrained from conducting distillation experiments on this dataset.

Supremind-Road. The Supremind-Road dataset is a real-world roadside 3D detection dataset that captures authentic traffic scenes. This dataset comprises $16,210$ image frames and corresponding LiDAR scans, delivering annotations for $243k$ 3D object boxes across four classes: *vehicles*, *pedestrians*, *cyclists*, and *tricycles*. Supremind-Road also contains diverse road scenes across different camera parameters, including 130 different intersection scenarios. Our utilization of this dataset involves testing within genuine application scenarios, thereby validating the adaptability and performance of various algorithms in intelligent transportation systems. The training, validation, and test data undergo a certain division, following a ratio of 75% : 10% : 15%. In this setup, the training and validation come from 110 distinct scenarios, whereas the test set comprises the remaining 20 different intersections. Consequently, the training and validation set form the `homologous` setting. Conversely, the training and test set constitute the `heterologous` setting. We kindly note that this dataset is not intended for public distribution and does not constitute a contribution to this article.

TABLE II: Comparison with state-of-the-art methods on the Rope3D validation set [9] in `homologous` settings. +(G) denotes adapting the ground plane.

Method	$IoU = 0.5$				$IoU = 0.7$			
	Car		Big Vehicle		Car		Big Vehicle	
	AP	Rope	AP	Rope	AP	Rope	AP	Rope
M3D-RPN+(G) [49]	54.19	62.65	33.05	44.94	16.75	32.90	6.86	24.19
Kinematic3D+(G) [50]	50.57	58.86	37.60	48.08	17.74	32.90	6.10	22.88
MonoDLE+(G) [51]	51.70	60.36	40.34	50.07	13.58	29.46	9.63	25.80
BEVFormer [6]	50.62	58.78	34.58	45.16	24.64	38.71	10.05	25.56
BEVDepth [7]	69.63	74.70	45.02	54.64	42.56	53.05	21.47	35.82
BEVHeight [10]	**74.60**	**78.72**	48.93	57.70	45.73	55.62	23.07	37.04
CoBEV (Ours)	73.39	77.11	**52.77**	**60.28**	**52.72**	**60.57**	**29.28**	**41.52**
w.r.t. BEVHeight	-1.21	-1.61	+3.84	+2.58	+6.99	+4.95	+6.21	+4.48

TABLE III: Comparison with state-of-the-art methods on the Rope3D validation set [9] in `heterologous` settings.

Method	Lifting	$IoU = 0.5$				$IoU = 0.7$			
		Car		Big Vehicle		Car		Big Vehicle	
		AP	Rope	AP	Rope	AP	Rope	AP	Rope
BEVFormer [6]	Implicit	25.98	39.51	8.81	24.67	3.87	21.84	0.84	18.42
BEVDepth [7]	Explicit	9.00	25.80	3.59	20.39	0.85	19.38	0.30	17.84
BEVHeight [10]	Explicit	29.65	42.48	13.13	28.08	5.41	23.09	1.16	18.53
CoBEV (Ours)	Hybrid	**31.25**	**43.74**	**16.11**	**30.73**	**6.59**	**24.01**	**2.26**	**19.71**
w.r.t. BEVHeight	-	+1.60	+1.26	+2.98	+2.65	+1.18	+0.92	+1.10	+1.18

B. Comparison with the State-of-the-Arts

For a comprehensive evaluation, we compare the proposed CoBEV against state-of-the-art BEV detectors, including BEVFormer [6], BEVDepth [7], and BEVHeight [10]. The evaluation is conducted on three datasets described as follows.

Results on the DAIR-V2X-I dataset. In Tab. I, we report the results of LiDAR-based [42], [44] and multi-modal [48] detectors, reproduced by the original DAIR-V2X-I dataset [1]. Our results demonstrate that CoBEV outperforms all other methods across the board, with consistent and notable improvements in each category. Specifically, it achieves a performance boost of +3.09% (68.86% *vs.* 65.77%) on the *vehicle* category, +3.02% (42.31% *vs.* 39.29%) on the *pedestrain* category, and +0.92% (61.00% *vs.* 60.08%) on the *cyclist* category, as compared to the previously best detector [10]. Remarkably, CoBEV marks the first approach where a camera-based monocular 3D detector achieves *vehicle* detection accuracy exceeding 80% in the Easy setting. Moreover, the accuracy of CoBEV for *vehicles* and *cyclists* largely surpasses that of LiDAR-based and multi-modal methods by ∼10%. When our BEV feature distillation framework is fully equipped, the detection accuracy of CoBEV$_{full}$ has seen substantial

TABLE IV: Mono3D detection results on the Supremind-Road dataset in `homologous` and `heterologous` settings.

Method	Lifting	Validation Set (`homologous`)								Test Set (`heterologous`)							
		Vehicle		Pedestrian		Cyclist		Tricycle		Vehicle		Pedestrian		Cyclist		Tricycle	
		Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
BEVFormer [6]	Implicit	69.17	58.78	11.32	11.35	25.44	20.14	20.74	19.60	46.96	37.87	0.12	0.12	2.62	2.59	0.00	0.00
BEVDepth [7]	Explicit	73.74	63.21	14.71	14.25	24.73	19.52	16.89	18.58	12.95	10.22	0.06	0.06	1.04	1.00	0.83	0.83
BEVHeight [10]	Explicit	76.65	66.17	20.96	20.59	50.01	44.20	43.68	43.03	52.44	41.86	1.41	0.94	**18.61**	**16.74**	23.07	20.42
CoBEV (Ours)	Hybrid	**78.12**	**67.71**	**23.41**	**21.65**	**52.98**	**45.20**	**47.87**	**48.18**	**55.24**	**44.10**	**1.46**	**1.32**	18.32	16.46	**30.72**	**28.38**
w.r.t. BEVHeight	-	+1.47	+1.54	+2.45	+1.06	+2.97	+1.00	+4.19	+5.15	+2.80	+2.24	+0.05	+0.38	-0.29	-0.28	+7.65	+7.96

enhancements across all categories, especially on small targets, which is comparable to the LiDAR detectors (47.21% *vs.* 52.49%). This improvement is attributed to the richer knowledge possessed by the fusion-modal teacher when it comes to detecting small targets. Notably, CoBEV$_{full}$ establishes new state-of-the-art performance on the DAIR-V2X-I benchmark. We notice that previous methods were trained only using ground truth labels within the $0{\sim}100m$ range for supervision. This is due to the fact that labels within the $100{\sim}200m$ range mainly contain challenging long-distance *vehicles* and rarely consist of small targets like *pedestrians*. To address this, we for the first time, cover a longer range of BEV detection tasks, encompassing targets in the $100{\sim}200m$. The method is denoted as CoBEV* in Tab. I. In comparison to CoBEV, CoBEV* demonstrates improvements across all categories, especially in the *vehicle* category (81.48% *vs.* 68.86%). Obtaining better performance in a longer distance demonstrates the advance of our hybrid lifting method and the effectiveness of fusing complimentary features based on depth and height.

Results on the Rope3D dataset. On the Rope3D dataset [9], we conduct a comparative evaluation of CoBEV against state-of-the-art monocular 3D detectors, considering two distinct settings: `homologous` and `heterologous`. Notably, some detection methods incorporate the ground plane as an additional constraint, marked as $+(G)$. As shown in Tab. II, CoBEV achieves the second-best *car* detection accuracy (73.39% *vs.* 74.60%) and the top accuracy for *big vehicles* (52.77% *vs.* 48.93%) under the easy evaluation conditions characterized by $IoU = 0.5$. Under the more strict evaluation of $IoU = 0.7$, CoBEV outperforms BEVHeight by large margins of 6.99% / 4.95% and 6.21% / 4.48% in terms of $AP_{3D|R40}$ and $Rope_{score}$ for *car* and *big vehicle* categories, respectively. Moving to the challenging heterologous setting, as displayed in Tab. III, all monocular 3D detectors endure a significant performance drop. Consistent with the previous comparison, CoBEV maintains its overall superiority, exhibiting state-of-the-art accuracy under both easy and strict evaluation conditions. This exceptional performance is attributed to CoBEV's camera-aware hybrid lifting design and the complementary feature selection module, which contribute to the construction of robust BEV features. Consequently, in the heterologous setting with new viewing cameras and unseen intersection scenes, CoBEV shows enhanced generalization capabilities across camera parameters and traffic scenarios.

Results on the Supremind-Road dataset. To facilitate a comprehensive evaluation across diverse intersection scenarios, we reproduce the previous state-of-the-art BEV detectors on our new Supremind-Road dataset. As shown in Tab. IV, CoBEV emerges as the top-performing method across the *vehicle*, *pedestrian*, *cyclist*, and *tricycle* categories, encompassing various difficulty levels within the `homologous` setting. Notably, it outperforms the second-best BEVHeight by considerable margins, with improvements of +1.47% / +1.54%, +2.45% / +1.06%, +2.97% / +1.00%, and +4.19% / +5.15%, respectively. Under the challenging `heterologous` setting, although the performances of all detectors drop greatly, CoBEV still achieves state-of-the-art performance across most categories. Especially noteworthy is its performance in large targets of the *vehicle* and *tricycle*, where it shows significant advantages of +2.80% / +2.24% and +7.65% / +7.96%, respectively. This proves CoBEV's better robustness and generalization capabilities to previously unseen transportation scenarios. In cases involving smaller targets such as *pedestrians* and *cyclists*, CoBEV's accuracy is on par with BEVHeight, significantly surpassing BEVFormer and BEVDepth.

C. Robustness Analysis

In real-world transportation scenarios, cameras positioned at intersections are often subjected to variations in extrinsic parameters caused by factors such as wind, vibrations, human adjustments, and other environmental conditions. Additionally, the intrinsic parameters of different cameras will also change. This motivates us to investigate the robustness of different BEV detectors in the context of fluctuations in camera parameters. Following the simulation methodology outlined in [52], we introduce offset noise with a $N(0, 1.67)$ distribution to the *roll* and *pitch* angles associated with the extrinsic matrix. For the camera *focal* length, we introduce scale noise, with the scaling coefficient following a $N(1, 0.2)$ distribution. Corresponding rotation and scaling transformations are also applied to the camera capture to ensure that the image are accurately mapped to the reference coordinate system using the camera parameters, thus preserving the calibration relationship.

To ensure a fair and consistent comparison, BEVDepth [7], BEVHeight [10], and the proposed CoBEV all adapt exactly the same camera parameter perturbation during the training process. As detailed in Tab. V, CoBEV maintains the best accuracy across all test-time scenarios involving noisy camera parameters. When only the camera focal length is disturbed, CoBEV exhibits significantly enhanced robustness compared to BEVHeight and BEVDepth, with respective accuracies of 66.36% *vs.* 60.45% *vs.* 60.19%. Similarly, when noise perturbs solely the camera angles, CoBEV maintains its superiority in terms of accuracy (66.33% *vs.* 63.08% *vs.* 62.57%). Furthermore, in situations where all camera parameter noise is considered, CoBEV clearly stands out as the only method achieving Vehicle and Pedestrian detection accuracy exceeding 60% (63.46% *vs.* 59.92% *vs.* 59.94%) and 30% (30.08%

TABLE V: Robustness analysis on the DAIR-V2X-I validation set. Three disturbed factors of roadside cameras are investigated, including focal length, roll angle, and pitch angle.

Method	Disturbed			Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
	focal	roll	pitch	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
BEVDepth [7]	-	-	-	75.31	65.24	65.32	32.68	31.01	31.33	46.96	50.88	51.44
	✓	-	-	72.17	60.19	60.20	25.75	25.16	24.35	40.65	47.09	47.21
	-	✓	-	74.78	62.72	62.81	30.80	30.20	30.43	45.58	50.07	50.72
	-	-	✓	74.83	62.76	62.85	30.21	28.62	28.91	46.07	50.15	50.85
	-	✓	✓	74.62	62.57	62.66	30.38	28.87	29.13	45.96	50.15	50.79
	✓	✓	✓	71.91	59.94	59.96	26.61	25.18	25.22	39.79	46.11	46.13
BEVHeight [10]	-	-	-	78.08	65.97	66.04	40.01	38.21	38.38	58.01	60.46	60.95
	✓	-	-	72.30	60.45	60.47	32.18	30.65	29.65	50.06	55.04	55.14
	-	✓	-	77.65	65.57	65.65	38.38	36.60	36.72	56.15	59.11	59.52
	-	-	✓	75.37	63.31	63.38	33.13	31.47	31.63	52.88	56.07	56.44
	-	✓	✓	75.06	63.08	63.16	33.67	31.19	31.30	51.65	54.93	56.83
	✓	✓	✓	71.71	59.92	59.96	27.81	26.43	26.36	47.42	51.19	51.26
CoBEV (Ours)	-	-	-	81.20	68.86	68.99	44.23	42.31	42.55	61.28	61.00	61.61
	✓	-	-	78.70	66.36	66.43	36.19	34.36	34.39	55.56	57.11	57.39
	-	✓	-	81.03	68.78	68.91	42.47	40.56	40.88	61.38	61.94	62.59
	-	-	✓	78.57	66.33	66.45	36.82	35.01	35.51	57.65	58.59	59.28
	-	✓	✓	78.60	66.34	66.47	37.38	35.57	35.89	56.44	57.44	58.10
	✓	✓	✓	**75.53**	**63.46**	**63.55**	**30.75**	**30.08**	**29.17**	**51.42**	**54.78**	**54.97**
	w.r.t. BEVHeight			+3.82	+3.54	+3.59	+2.94	+3.65	+2.81	+4.00	+3.59	+3.71

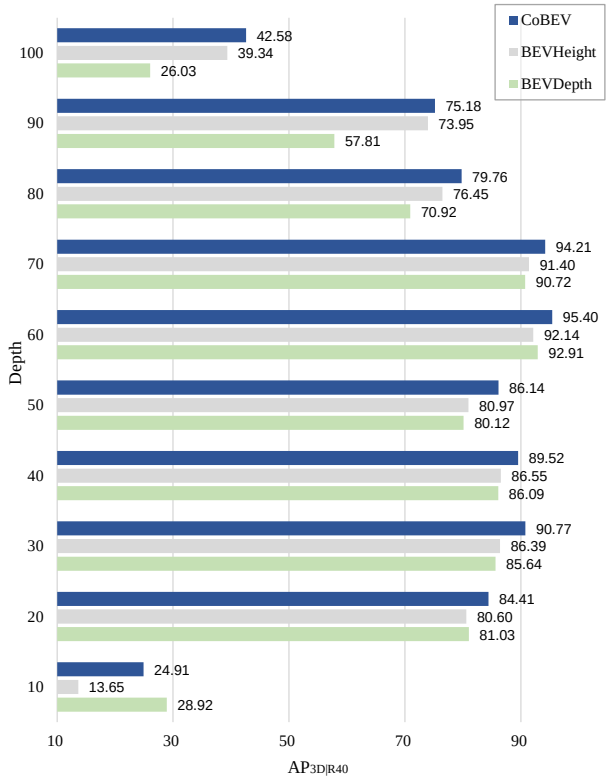


Fig. 6: *Range-wise evaluation on the DAIR-V2X-I validation set.* Metric is $AP_{3D|R40}$ of the Vehicle category under moderate setting. The sample interval is $10m$, *e.g.*, the value at vertical axis 50 indicates the overall performance of all samples between $45m$ and $55m$.

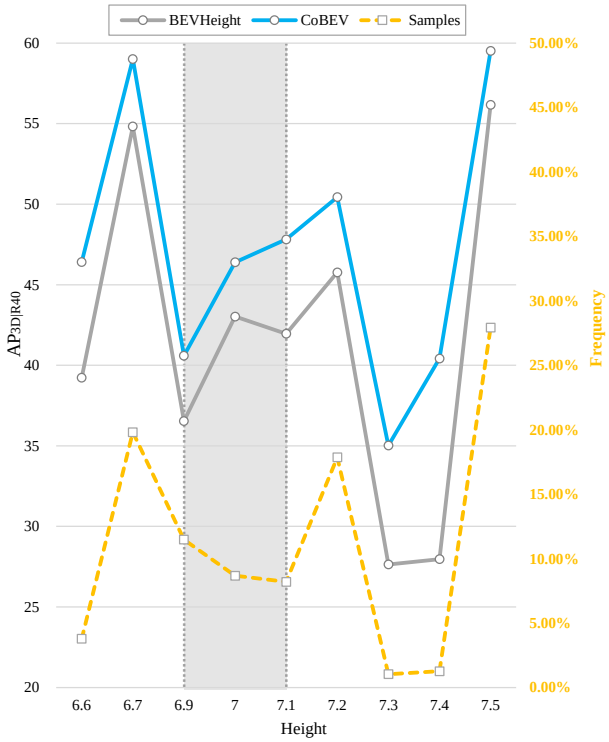


Fig. 7: *Height-wise evaluation on the Rope3D validation set.* Metric is $AP_{3D|R40}$ of all vehicles under $IoU = 0.7$ setting. The sample interval is $0.1m$, *e.g.*, the value at horizontal axis 7.0 indicates the overall performance between $6.95m$ and $7.05m$. The overall accuracy is associated with the training set's sample frequency. However, we observed a negative correlation within the height range of $6.9 \sim 7.1$ meters, suggesting that the detection error based on height increases as the camera height decreases.

vs. 26.43% *vs.* 25.18%), respectively. These results reveal CoBEV's excellent robustness and resistance to interference when confronted with varying camera parameters.

Delving into the complementary attribute of depth-based and height-based BEV detector. To delve deeper into the complementary attribute between depth- and height detectors, we present the accuracy distributions of BEVDepth [7],

TABLE VI: Ablations on BEV feature fusion on the DAIR-V2X-I dataset [1].

Source	Aggregation	Depth-based	Height-based	Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
				Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Vanilla	*w/o*	✓	-	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
		-	✓	77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
2D BEV	CBAM [53]	✓	✓	78.25	66.15	66.23	41.45	39.57	39.81	56.92	58.05	58.52
	Attention Feature Fusion [54]	✓	✓	78.61	66.41	66.49	42.06	40.16	40.36	59.35	60.15	60.72
3D Voxel	Add	✓	✓	78.55	68.53	68.65	**45.38**	**43.33**	**43.66**	60.25	58.65	59.26
	Concatenation	✓	✓	78.53	68.46	68.59	44.55	42.48	42.79	60.87	60.22	60.92
	Channel Attention$_{3D}$	✓	✓	80.87	68.59	68.71	44.04	42.00	42.35	60.85	58.08	60.06
	Spatial Attention$_{3D}$	✓	✓	81.00	68.74	68.86	43.56	41.63	41.91	60.68	60.36	61.13
	CFS (Ours)	✓	✓	**81.20**	**68.86**	**68.99**	44.23	42.31	42.55	**61.28**	**61.00**	**61.61**
	w.r.t. BEVHeight	-	-	+3.42	+3.09	+3.14	+3.01	+3.02	+3.09	+1.05	+0.92	+1.07

BEVHeight [10], and the proposed CoBEV within different range intervals. As shown in Fig. 6, it becomes apparent that depth-based detectors exhibit a notable advantage in short-range scenarios, particularly at distances of around 10 meters. We believe that this advantage stems from the fact that many targets within this range may be partially obstructed by the camera's Field-of-View (FoV). In contrast, BEVHeight prioritizes capturing the overall height of each target to facilitate classification, emphasizing semantic information. Thus, height-based detectors are more focused on the semantic aspect of targets. Estimating semantic information for truncated objects is more challenging than regressing depth at short distances, which accounts for the dominance of depth-based detectors in this range. CoBEV, however, attains a harmonious balance in accuracy.

As the target distance exceeds 70 meters, regressing depth becomes increasingly challenging, affording height-based detectors a significant advantage. The robust complementary BEV features constructed by CoBEV further amplify this advantage. Fig. 7 presents the accuracy distribution of BEVHeight and CoBEV within different camera height intervals, along with the frequency of training samples. In general, detection accuracy shows a positive correlation with the frequency of training samples. However, it is worth noting an abnormal negative correlation between accuracy and frequency in the $6.85 - 7.15$ meter interval. This is attributed to the gradual reduction in camera installation height from 7.15 meters to 6.85 meters, leading to increased height-based detection errors that intensify the difficulty of the task, even surpassing the influence of training sample frequency. CoBEV leverages depth-based complementary features and it continues to maintain a significant advantage within relatively challenging intervals.

D. Ablation Studies

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

Fusion Strategy. We explore different fusion strategies for depth-based and height-based features using the ResNet101 image backbone [55]. Tab. VI reveals that incorporating 3D features obtained via our partial-pillar voxel pooling for fusion outperforms the approach of obtaining 2D features directly through voxel pooling and then fusing. For instance, merely adding 3D features leads to significant enhancements in Vehicle (68.53% *vs.* 65.77%) and Pedestrian (43.33% *vs.* 39.29%) detection accuracy. The proposed Complementary Feature Selection (CFS) module achieves the best Vehicle (68.86% *vs.* 65.77%) and Cyclist (61.00% *vs.* 60.08%) results while also improving Pedestrian detection performance. In comparison to the baseline [10], the CFS fusion module constructs complementary BEV features through two-stage feature selections, thereby significantly improving detection results across all categories.

Hybrid-Lifting BEV Feature. Tab. VII shows our exploration of how heterogeneous BEV features reinforce each other. As for transformer-based BEV features, we employ the BEVFormer-style [6] using cross-attention mechanisms. For MLP-based BEV features, linear layers are used for mapping to achieve perspective transformation. Experimental results indicate that depth-based and height-based BEV features exhibit the best complementarity. Introducing additional transformer-based or MLP-based BEV features does not further enhance detection performance, aligning with our original motivation for CoBEV construction and observations of accuracy distributions across depth and height intervals.

Point Cloud Supervision for Depth or Height. Considering LiDAR's capacity to precisely measure the depth and height of roadside targets, we experiment with introducing point clouds for depth or height supervision. Surprisingly, contrary to prior work [7], we discover that refraining from introducing any form of point cloud supervision for depth and height effectively enhances detection accuracy across all categories. Introducing supervision solely for depth or height further improves vehicle detection performance but hinders the results on small targets. When both depth and height supervision are applied, vehicle detection accuracy drops significantly (66.85% *vs.* 68.86%). Our findings corroborate that in BEVHeight [10], which observes similar trends for height supervision. To delve into this phenomenon, we replace the estimated depth/height distribution with depth or height ground truth during test time after training and observe corresponding accuracy changes. Depth ground truth significantly improves detection accuracy (70.02% *vs.* 68.86%), while height ground truth weakens the detector's performance (53.25% *vs.* 68.86%). We argue that this is because the height detector tends to learn the overall height interval of vehicles, pedestrians, and other categories, rather than specific height values for each pixel, focusing more on semantic cues during training than geometric ones. CoBEV leverages the rich semantic context of the height detector and the precise geometric cues of the depth detector, therefore showing the ability to construct robust BEV features.

TABLE VII: Ablations on hybrid-lifting BEV feature on the DAIR-V2X-I dataset [1].

Depth-based	Height-based	Implicit-based		Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
		Transformer	MLP	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
✓	-	-	-	73.05	61.32	61.19	22.10	21.57	21.11	42.85	42.26	42.09
-	✓	-	-	76.61	64.71	64.76	27.34	26.09	26.33	49.68	48.84	48.58
-	-	✓	-	61.37	50.73	50.73	16.89	15.82	15.95	22.16	22.13	22.06
✓	✓	-	-	**78.02**	**65.90**	**68.09**	**32.72**	**31.06**	**31.29**	**55.56**	**56.31**	**57.05**
✓	✓	✓	-	76.65	64.79	64.89	31.56	30.00	30.27	52.23	52.88	53.61
✓	✓	-	✓	77.24	65.19	65.28	32.44	30.90	31.17	49.87	51.76	52.32

TABLE VIII: Ablations on point cloud supervision on the DAIR-V2X-I dataset [1].

Depth-supervision	Height-supervision	Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
		Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Using GT	-	82.27	70.02	70.06	64.19	61.83	62.05	68.88	69.63	70.09
-	Using GT	61.80	53.25	51.49	39.20	37.48	37.71	55.47	58.64	59.15
Using GT	Using GT	82.64	70.29	70.33	62.76	60.49	60.81	66.52	65.00	65.45
-	-	81.20	68.86	68.99	**44.23**	**42.31**	**42.55**	**61.28**	**61.00**	**61.61**
✓	-	**81.33**	**69.08**	**69.18**	42.86	40.89	41.22	59.62	58.32	58.94
-	✓	81.27	69.01	69.10	42.81	40.79	41.07	59.85	57.29	57.88
✓	✓	79.04	66.85	69.05	43.81	41.93	42.24	59.76	60.91	61.45

TABLE IX: Ablations on knowledge distillation on the DAIR-V2X-I dataset [1].

Case	Modality	Vehicle			Pedestrian			Cyclist		
		Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Teacher	C+L	82.57	70.28	70.33	73.01	70.61	70.66	73.95	76.31	76.69
Student	C	81.20	68.86	68.99	44.23	42.31	42.55	61.28	61.00	61.61
FitNet [56]	C	81.00	68.71	68.82	44.41	42.47	42.69	63.75	62.25	62.93
CMKD [46]	C	81.14	68.80	68.90	44.53	42.51	42.84	62.19	63.53	64.17
BEVDistill [57]	C	81.38	69.05	69.18	45.61	43.50	43.93	63.63	64.40	65.00
UniDistill [43]	C	81.71	69.33	69.43	48.78	46.57	46.86	64.34	65.32	65.77
Ours	C	**82.01**	**69.57**	**69.66**	**49.32**	**47.21**	**47.48**	**66.13**	**66.17**	**66.69**
w.r.t. CoBEV	-	+0.81	+0.71	+0.67	+5.09	+4.90	+4.93	+4.85	+5.17	+5.08

– Note: 'C': Camera is used, 'C+L': both Camera and LiDAR are used.

Knowledge Distillation. In Tab. IX, we compare the popular knowledge distillation frameworks [43], [46], [56], [57] with the proposed BEV feature distillation scheme. The teacher model, incorporating fused modalities, excels at handling small targets like pedestrians and cyclists. Consequently, the student model can more effectively emulate the teacher model's prior knowledge regarding small object detection problems. The proposed BEV feature distillation scheme adeptly handles erroneous supervision signals when the teacher model performs poorly, introducing the distill adapter at low-level features to enhance the detection accuracy of large targets like Vehicles ($+0.71\%$). The student model closely replicates the teacher model at high-level BEV features, thus capturing the structural knowledge of the traffic scene independent of target size. This approach avoids suppressing visual cues of small targets, resulting in substantial improvements in the detection accuracy of Pedestrians and Cyclists ($+4.90\%$ / $+5.17\%$), outperforming existing distillation methods.

V. Conclusion

In this paper, we introduce CoBEV, a monocular camera-based 3D detector designed for roadside scenes, which harnesses the complementary attributes of depth and height information to construct robust BEV representations for elevating traffic scene understanding. By combining accurate geometric cues from depth features with distinct semantic context categories from height features, CoBEV achieves state-of-the-art accuracy on public roadside 3D detection datasets DAIR-V2X-I and Rope3D. Notably, CoBEV exhibits exceptional adaptability to varying cameras and intersection scenes in heterologous settings and presents enhanced resilience against perturbations in camera parameters, thanks to the robust BEV features built by the complementary feature selection module. Furthermore, we introduce a BEV feature distillation method tailored for roadside detection scenarios, which further promotes the detector's comprehension of scene semantic information and structural relationships, leading to improved detection accuracy agnostic to the target size. We analyze how the detection performance varies across different depth and camera installation height intervals, validating the natural complementarity between depth and height detectors. Extensive experiments and ablation studies verify the effectiveness of CoBEV and its core components.

References

[1] H. Yu *et al.*, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 329–21 338.

[2] W. Bao, B. Xu, and Z. Chen, "MonoFENet: Monocular 3D object detection with feature enhancement networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 2753–2765, 2020.

[3] M. Feng, S. Z. Gilani, Y. Wang, L. Zhang, and A. Mian, "Relation graph network for 3D object detection in point clouds," *IEEE Transactions on Image Processing*, vol. 30, pp. 92–107, 2021.

[4] B. Xie, L. Yang, A. Wei, X. Weng, and B. Li, "MuTrans: Multiple transformers for fusing feature pyramid on 2D and 3D object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 4407–4415, 2023.

[5] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *European Conference on Computer Vision (ECCV)*, vol. 12359, 2020, pp. 194–210.

[6] Z. Li *et al.*, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European Conference on Computer Vision (ECCV)*, vol. 13669, 2022, pp. 1–18.

[7] Y. Li *et al.*, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, no. 2, 2023, pp. 1477–1485.

[8] X. He *et al.*, "SSD-MonoDETR: Supervised scale-aware deformable transformer for monocular 3D object detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[9] X. Ye *et al.*, "Rope3D: The roadside perception dataset for autonomous driving and monocular 3D object detection task," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 309–21 318.

[10] L. Yang *et al.*, "BEVHeight: A robust framework for vision-based roadside 3D object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 611–21 620.

[11] L. Xie, G. Xu, D. Cai, and X. He, "X-View: Non-egocentric multi-view 3D object detector," *IEEE Transactions on Image Processing*, vol. 32, pp. 1488–1497, 2023.

[12] J. U. Kim, H.-I. Kim, and Y. M. Ro, "Stereoscopic vision recalling memory for monocular 3D object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2749–2760, 2023.

[13] H. Liu, H. Liu, Y. Wang, F. Sun, and W. Huang, "Fine-grained multilevel fusion for anti-occlusion monocular 3D object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 4050–4061, 2022.

[14] K. Peng *et al.*, "MASS: Multi-attentional semantic segmentation of LiDAR data for dense top-view understanding," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 824–15 840, 2022.

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

[16] S. Li *et al.*, "Bi-Mapper: Holistic BEV semantic mapping for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7034–7041, 2023.

[17] C. Chen, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Trans4Map: Revisiting holistic bird's-eye-view mapping from egocentric images to allocentric semantics with vision transformers," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 4013–4022.

[18] Z. Teng *et al.*, "360BEV: Panoramic semantic mapping for indoor bird's-eye view," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[19] Y. Li *et al.*, "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[20] K. Yang, X. Hu, and R. Stiefelhagen, "Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 1866–1881, 2021.

[21] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6568–6577.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, vol. 12346, 2020, pp. 213–229.

[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[25] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 913–922.

[26] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Conference on Robot Learning (CoRL)*, 2022, pp. 180–191.

[27] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *European Conference on Computer Vision (ECCV)*, vol. 13687, 2022, pp. 531–548.

[28] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3D object detection," in *British Machine Vision Conference (BMVC)*, 2018, p. 285.

[29] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1265–1274.

[30] C.-Y. Tseng, Y.-R. Chen, H.-Y. Lee, T.-H. Wu, W.-C. Chen, and W. Hsu, "CrossDTR: Cross-view and depth-guided transformers for 3D object detection," *arXiv preprint arXiv:2209.13507*, 2022.

[31] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conference on Robot Learning (CoRL)*, 2017, pp. 1–16.

[32] S. Fan, Z. Wang, X. Huo, Y. Wang, and J. Liu, "Calibration-free BEV representation for infrastructure perception," *arXiv preprint arXiv:2303.03583*, 2023.

[33] N. Hendy *et al.*, "FISHING Net: Future inference of semantic heatmaps in grids," *arXiv preprint arXiv:2006.09917*, 2020.

[34] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural attention fields for end-to-end autonomous driving," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 773–15 783.

[35] W. Yang *et al.*, "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 536–15 545.

[36] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 750–13 759.

[37] C. Yang *et al.*, "BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 830–17 839.

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

[39] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8445–8453.

[40] V. Cartillier, Z. Ren, N. Jain, S. Lee, I. Essa, and D. Batra, "Semantic MapNet: Building allocentric semantic maps and representations from egocentric views," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 2, 2021, pp. 964–972.

[41] X. Chi *et al.*, "BEV-SAN: Accurate BEV 3D object detection via slice attention networks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 461–17 470.

[42] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 697–12 705.

[43] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, "UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5116–5125.

[44] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[45] Z. Chong *et al.*, "MonoDistill: Learning spatial features for monocular 3D object detection," in *International Conference on Learning Representations (ICLR)*, 2022.

[46] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3D object detection," in *European Conference on Computer Vision (ECCV)*, vol. 13670, 2022, pp. 87–104.

[47] X. Li *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 21 002–21 012.

[48] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-Net: Multimodal VoxelNet for 3D object detection," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 7276–7282.

[49] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9287–9296.

[50] G. Brazil, G. Pons-Moll, X. Liu, and B. Schiele, "Kinematic 3D object detection in monocular video," in *European Conference on Computer Vision (ECCV)*, vol. 12368, 2020, pp. 135–152.

[51] X. Ma *et al.*, "Delving into localization errors for monocular 3D object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4721–4730.

[52] K. Yu *et al.*, "Benchmarking the robustness of LiDAR-camera fusion for 3d object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3188–3198.

[53] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, vol. 11211, 2018, pp. 3–19.

[54] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3559–3568.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[56] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *International Conference on Learning Representations (ICLR)*, 2015.

[57] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection," *arXiv preprint arXiv:2211.09386*, 2022.

[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.
[59] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.
[60] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

APPENDIX A
NETWORK OVERVIEW

As shown in Fig. 5, given a monocular roadside image frame $X \in \mathbb{R}^{3\times H\times W}$ with its corresponding extrinsic matrix $E \in \mathbb{R}^{3\times 4}$ and intrinsic matrix $I \in \mathbb{R}^{3\times 3}$. Formally, a 2D convolutional image encoder with FPN neck maps X to the high-dimensional image feature $F \in \mathbb{R}^{C\times \frac{H}{16}\times \frac{W}{16}}$. Then, F and the camera parameters $\{E, I\}$ are fed into the *Camera-aware Hybrid Lifting* module. The purpose of this stage is to lift the monocular image features F from the 2D coordinate system to the 3D frame by calculating the depth distribution, feature context, and height distribution, respectively. The partial-pillar voxel pooling module then compresses the 3D features from the pillars on the BEV plane, which are voxels with height. Note that we perform height compression instead of completely flattening the features to a 2D plane in order to preserve a certain height axis. Therefore, the free flow of relevant visual cues in all dimensions can be achieved in the next novel *Complementary Feature Selection (CFS)* module, and then construct complementary BEV feature $F_{com} \in \mathbb{R}^{C_{com}\times \frac{H}{16}\times \frac{W}{16}}$. Finally, the complementary BEV feature F_{com} is fed into the detection head, in which the 3D bounding boxes $\mathbf{B_{ego}}$ composed of position (x, y, z), dimension (l, w, h), and orientation θ is output. Moreover, we design a fusion-to-camera *BEV Feature Distillation* framework by using a teacher that fuses the multi-modality information of LiDAR and camera to further enhance the accuracy of monocular 3D detection agnostic to the target size.

APPENDIX B
TRAINING OBJECTIVES

The final loss function can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{low} + \mathcal{L}_{high} + \mathcal{L}_{res}, \tag{B.1}$$

where the detection loss $\mathcal{L}_{det}$ inherits from [44].

APPENDIX C
IMPLEMENTATION DETAILS

Unless otherwise specified, the CoBEV detector employs ResNet101 [55] as its image encoder, while ResNet50 [55] is used for ablation experiments. The input image size is set to 864×1536. The width range of the Bird's Eye View (BEV) grid spans across $[-51.2m, 51.2m]$, with a length range of $[0m, 102.4m]$. The cell size of the BEV grid is $0.4m \times 0.4m$, resulting in a final resolution of 256×256. The depth range D of the target is defined as $[2m, 104.4m]$, with a unit size of $0.4m$, employing a Uniform Discretization (UD) that yields a total of 256 bins. The height range H of the target varies depending on the statistical distribution of the specific dataset. For DAIR-V2X-I [1] and Rope3D [9] datasets, we adhere to the official definitions, with height ranges of $[-2m, 0m]$ and $[-1.5m, 3m]$, respectively. In the case of the Supremind-Road dataset, the height range spans from $[-1m, 4m]$. The height cell size follows dynamic partitioning as outlined in Equ. 6 with α set to 1.5. CoBEV is implemented in PyTorch and incorporates random 2D rotation and scaling augmentation. Training is performed individually on all datasets, with a batch size of 16 for 150 epochs on each dataset. The learning rate is set to $2e-4$, and the optimizer is AdamW [58].

APPENDIX D
EVALUATION METRICS

Following the established detection metrics employed in various benchmark datasets, our evaluation differs depending on the dataset. For the DAIR-V2X-I dataset [54], we report the 40-point average precision ($AP_{3D|R40}$) of bounding boxes [59], which is further categorized into three modes: easy, middle, and hard, based on the box characteristics such as size, visibility, and truncation [60]. In the case of the Rope3D dataset [9], we employ the $AP_{3D|R40}$ and the $Rope_{score}$ [9] for assessment. The $Rope_{score}$ provides a comprehensive evaluation, taking into account factors such as bounding box center, orientation, area, and ground points. We present results for Rope3D under two conditions: easy and strict, characterized by bounding box Intersection-over-Union (IoU) thresholds of 0.5 and 0.7, respectively. For the Supremind-Road dataset, we utilize $AP_{3D|R40}$ as the evaluation metric, and categorize it into easy and hard modes, with the distinction based on the truncation of the boxes. According to different target sizes, the IoU threshold is 0.5 for vehicles and 0.25 for pedestrians, cyclists, and tricycles.

APPENDIX E
FUTURE WORK

In the future, we are interested in exploring the performance of the CoBEV framework in vehicle-side multi-camera 3D detection tasks, aiming for a unified, high-precision, and robust detector for both vehicles and infrastructure. Moreover, we intend to leverage generative techniques, such as diffusion models, to simulate and augment training data, particularly in scenarios with limited real-world samples, like traffic accident scenarios. This approach will enhance the model's reliability under safety-critical conditions and help to address the long tail problem in training. We are also keen to explore the application of the CoBEV framework in addressing perception challenges related to vehicle-to-road collaboration and multi-modal fusion. Our goal is to enhance the perception range of intelligent vehicles and improve the responsiveness of the overall transportation system by constructing robust features within a unified BEV space.

APPENDIX F
QUALITATIVE RESULT VISUALIZATIONS

In this section, we present qualitative comparison results between BEVDepth [7], BEVHeight [10], and the proposed CoBEV across the DAIR-V2X-I [1], Rope3D [9], and

Fig. F.1: *Qualitatively Comparisons on the DAIR-V2X-I dataset.* BEVDepth and BEVHeight exhibit missed detections when encountering distant targets. In contrast, CoBEV demonstrates the capability to effectively address challenging long-distance targets, surpassing the performance of prior methods.

Supremind-Road datasets. Furthermore, in Fig. F.4, we provide a qualitative comparison of the robustness of the CoBEV framework in comparison to previous methods when subjected to camera parameter disturbances.

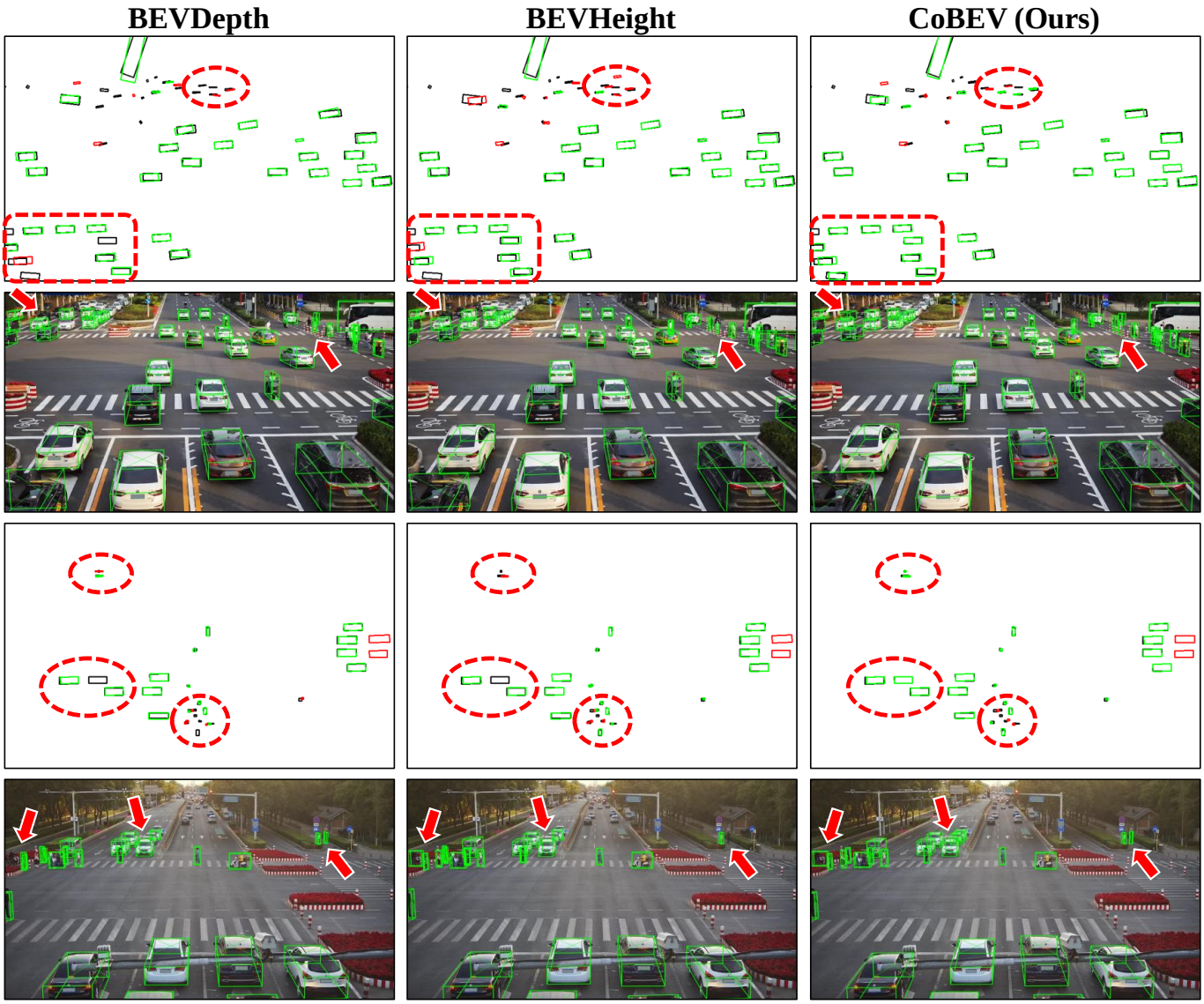


Fig. F.2: *Qualitatively Comparisons on the Rope3D dataset.* BEVDepth and BEVHeight prove less effective in handling small targets like pedestrians and cyclists. CoBEV, on the other hand, enhances the detection accuracy of such small targets by constructing complementary, fine-grained BEV representations.

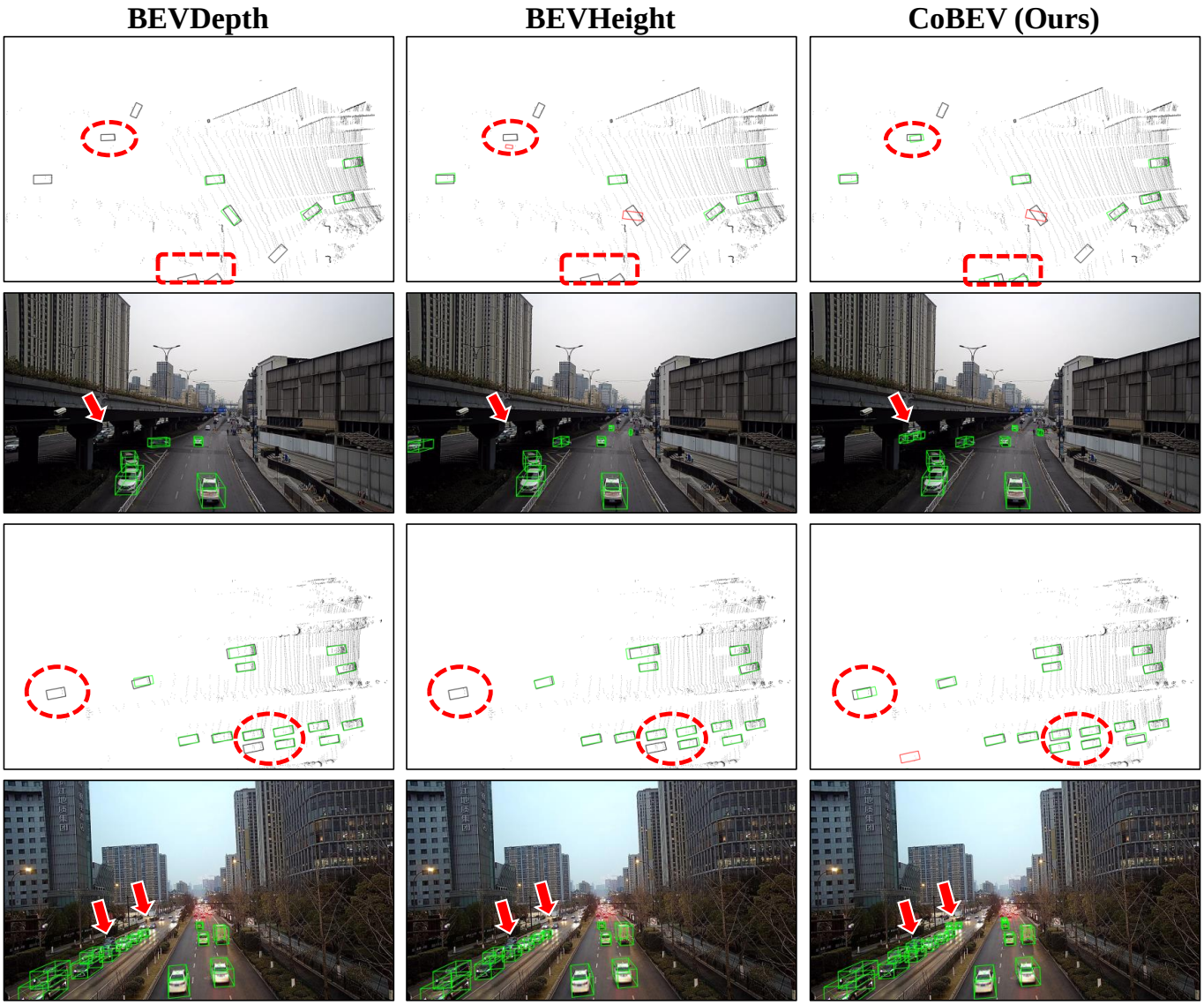


Fig. F.3: *Qualitatively Comparisons on the Supremind-Road dataset.* In comparison to BEVDepth and BEVHeight, the proposed CoBEV mitigates false alarms and missed detections, simultaneously enhancing the accuracy of detecting challenging long-distance targets.

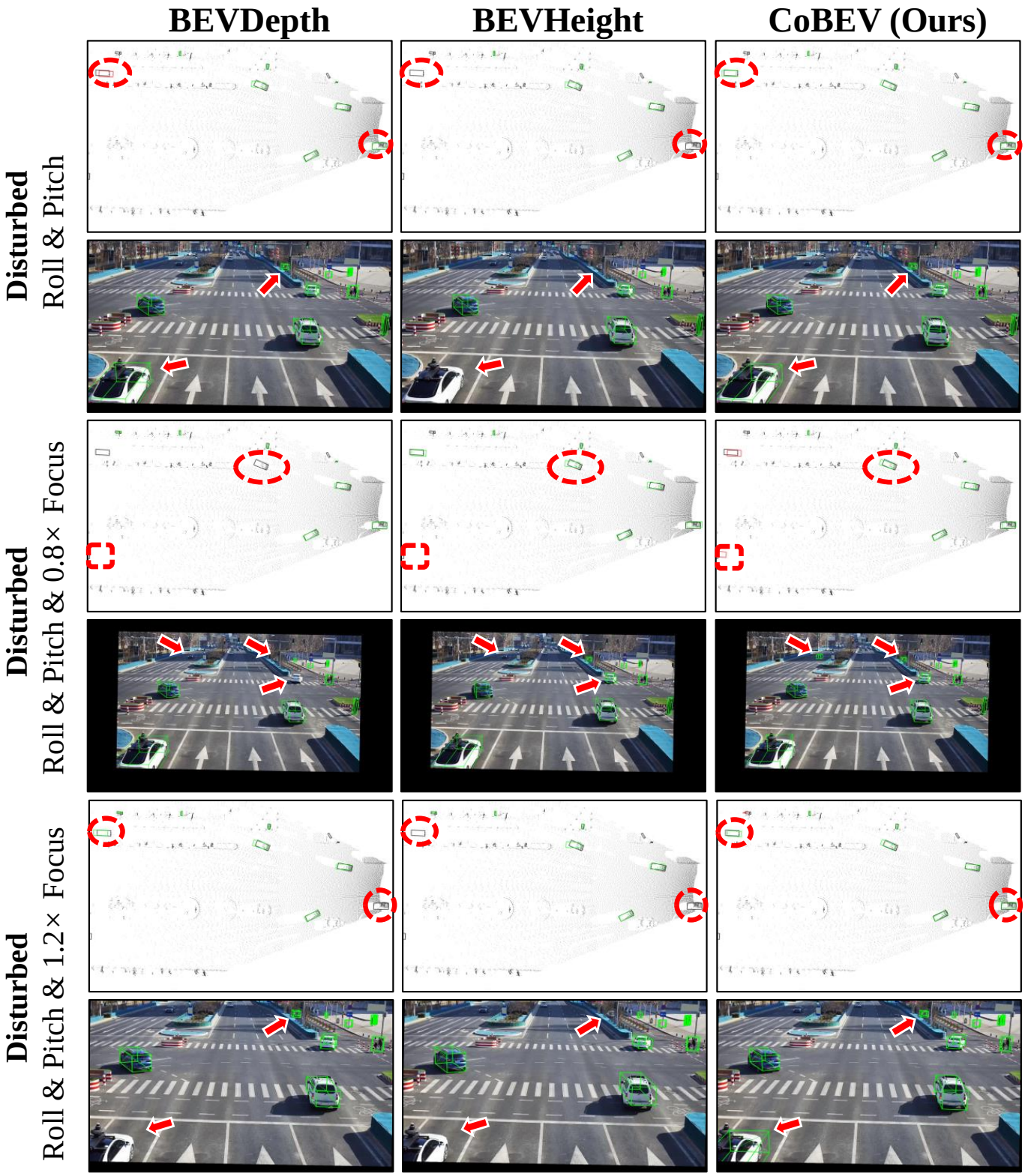


Fig. F.4: *Qualitatively Comparisons on the DAIR-V2X-I dataset under the camera parameters disturbance.* In the condition of various camera parameter disturbances, including focal length, roll, and pitch, CoBEV consistently demonstrates the most robust target detection capabilities. This applies to both targets truncated at close range and small targets situated at long distances.