BEVRoad: a Cross-Modal and Temporary-Recurrent 3D Object Detector for Infrastructure Perception

Xiaohai Li, Jieyao Zhang, Jiaming Gu, Xiaoyuan Lu and Liang Zhang

May 26, 2024

BEVRoad: A Cross-Modal and Temporary-Recurrent 3D Object Detector for Infrastructure Perception

Xiaohai Li[1][0000], Jieyao Zhang[2,3][1111], Jiaming Gu[3][2222], Xiaoyuan Lu[3][2222], and Liang Zhang[3][2222]

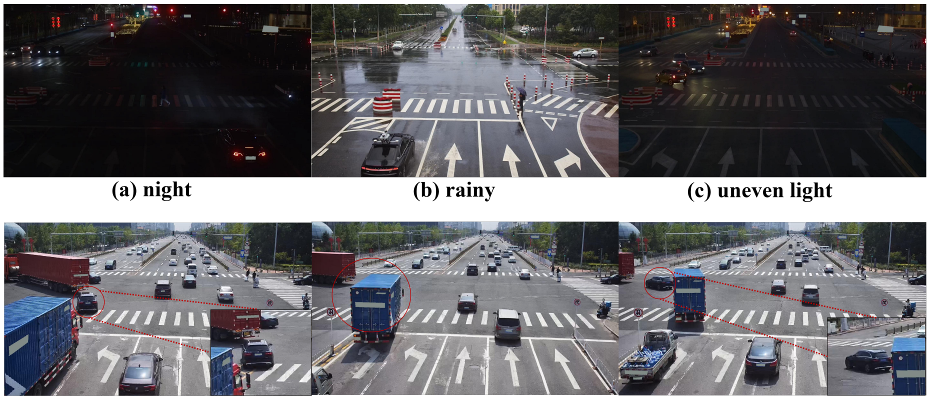
[1] Xidian University, Xian,China
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. BEV object detection has made significant strides in recent years. However, infrastructure perception focusing on roadside scenes is mainly underexplored, and most recent detectors tend to only rely on monocular camera, which hinder the perception capability under given scenarios such as rain, high humidity, and uneven light conditions. To address this problem, we propose an effective 3D object detection framework, dubbed BEVRoad. A lightweight spatial-channel adaptive fusion module (SCAFM) is designed for the impartial fusion of camera and LiDAR BEV features, greatly improving the representation capability of the model. Furthermore, to alleviate the blockage caused by the relative movement of objects under the road, we add a simple spatio-temporal network named TrajNet to perform temporal modeling on the BEV feature map and predict the target motion position, which achieves excellent performance improvements only with negligible computation cost compared to the single-frame baseline. Experimental results on DAIR-V2X demonstrate that BEVRoad achieves wonderful performance, including +11.09% for vehicle, +16.61% for pedestrian, and +6.64% for cyclist compared to BEVHeight.

Keywords: BEV Object Detection · Infrastructure Perception · Modal Fusion · Temporary Modeling.

1 Introduction

3D object detection is one of the core topics in the computer vision field, which plays a critical role in autonomous driving and intelligent transportation [17]. The application of robust 3D object detection in infrastructure perception helps improve traffic flow and intelligent transportation construction [31, 33]. Roadside 3D object detection aims to predict the locations, sizes, rotation, velocity, and classes of critical roadside objects, e.g., vehicles, pedestrians, and cyclists, generally taking camera, LiDAR, or 4D mmWave radar data as input.



(a) night

(b) rainy

(c) uneven light

(d) unoccluded previous frame

(e) occluded curent frame

(f) Unoccluded next frame

Fig. 1. Special Case. (a), (b), and (c) describe the conditions at night, rainy days, and uneven light. (d), (e), and (f) show the unoccluded and occluded states of the target during movement in the sequence frames.

Learning powerful representations in bird's-eye-view (BEV) for 3D perception is trending and drawing extensive attention, which provides a physics-interpretable way to capture rich semantic and geometric information for object detection or map segmentation. BEV perception represents features from different senors in a unified perspective to solve occlusion and scale problems, taking advantage of "God's View" adequately [9]. Due to the roadside camera mounted on poles a few meters above the ground, the infrastructure-centric detector can effectively cover the vehicle-centric perception blind spots. However, current camera-only infrastructure perception methods have two main challenges: 1) Detection from images is generally vulnerable to extreme weather and light conditions [18]. A unimodal detector based on monocular camera at night or on rainy days is much harder than detection on sunny days, as shown in Figure. 1 (a), (b), and (c). 2) At traffic intersections, a larger number of objects are expected to be observed in roadside view, as shown in Figure. 1 (d), (e), and (f). Thus, overlapping and occlusion between objects are very common phenomena, increasing the density and difficulty of a perceptual system [31].

To address the practical challenges of infrastructure-side perception, we hence tailor a cross-modal and temporary-recurrent robust 3D object detection method to the roadside application, dubbed BEVRoad. For effective modal fusion, cameras and LiDAR should be treated equally. No matter when a random sensor fails, the corresponding branch of the other modality can work independently [12]. To this end, we propose the spatial-channel adaptive fusion module (SCAFM), a lightweight network module that enhances the accuracy of the model while incurring minimal computational cost. In the timing module TrajNet, the optical flow motion of the target pixel is learned from the historical frame and the current frame in an iterative loop.

To demonstrate the advantages of our proposed BEVRoad, we have conducted extensive experiments on the two public roadside perception benchmarks, DAIR-V2X-I and DAIR-V2X-Seq [33]. Our model performs better and achieves great improvement with negligible additional parameters and calculations, including +11.09% for vehicle, +16.61% for pedestrian, and +6.64% for cyclist compared to BEVHeight [29]. BEVRoad runs at 11 FPS with a latency of 90.9 ms on the NVIDIA GeForce RTX 3090 GPU, sacrificing a little bit of speed.

In summary, our contributions are summarized as follows:

(1) To the best of our knowledge, this is the first 3D object detection model to introduce multi-modality in infrastructure perception, which can effectively deal with the interference of harsh environments.

(2) We propose a novel temporary-recurrent module that accurately predicts the speed and position information of the target, effectively alleviating the target occlusion problem compared to the single-frame baseline.

2 Related Work

Vehicle-side 3D Object Detection. Vehicle-side 3D detection refers to the process of analyzing multi-view images captured by cameras mounted on a vehicle to predict the 3D locations, dimensions, and orientations of the interest targets [17]. Since most vehicle-side sensors are mounted atop cars with an almost near-zero pitch angle, the optical axis of the sensors is parallel to the horizontal plane [7, 31, 33]. Popular methods can be divided into transformer-based and depth-based schemas according to the difference in view transformation. Transformer-based detectors generally design BEV grid queries or a set of object queries to perform the view transformation. According to BEV grid coordinates, BEVFormer [11] defines some learnable BEV queries of the spatial local cross-attention mechanism to interact with image features in the regions of interest. PETR [14] encodes the position information of 3D coordinates into image features based on camera-frustm, producing the 3D position-aware features. LSS [19] is a pioneering work on BEV perception that explicitly predicts the depth distribution per grid on 2D feature, then lifts the 2D feature per grid via the corresponding depth to voxel space based on pseudo point cloud frustum. BEVDepth [10] leverages depth supervision derived from point clouds to guide depth learning and encodes camera intrinsics and extrinsics into a depth learning module, which improves the quality of depth estimation and makes feature projection more accurate. However, the results of these methods are not particularly ideal when applied directly to roadside perception, which illustrates the differences between 3D object detection on the roadside and vehicle-side and makes them difficult to generalize.

Infrastructure-side 3D Object Detection. Roadside sensors are installed on the poles with distinct pitch angles of the viewpoint, mounting heights as well as various roadside environments, which raises questions about the algorithm's capacity for generalization [7, 31]. Recently, several roadside detection methods have been proposed since the release of datasets like DAIR-V2X-I [33] and Rope3D [31], which significantly promote the development of 3D perception in infrastructure-side scenarios. CBR [2] is a calibration-free BEV representation network that achieves BEV detection based on multi-layer perceptron (MLPs) without calibration parameters and additional depth supervision, but the perceptual accuracy is limited. BEVHeight [29] is a pioneering work initially focusing on roadside detection that predicts the relative height distribution of the scene image features and projects the features more accurately into 3D space through a height-based projection module, avoiding the problem of poor depth estimation quality. Unanimously, subsequent CoBEV [21] and BEVHeight++ [27] chose the same idea to improve BEV representations by seamlessly incorporating complementary geometry-centric depth and semantic-centric height cues, which achieved excellent performances. To solve the over-fitting problem of roadside backgrounds and camera internal and external parameters, SGV3D [30] proposes a background suppression module (BSM) to attenuate background features and introduce instances with new scenes and new poses. MonoUNI [7] proposes a normalized depth optimization goal through the derivation of theoretical formulas and develops 3D normalized cube depth for obstacles by geometric relationships, which unify vehicle-side and road-side detection and achieve **SOTA** on DAIR-V2X.

3 Method

3.1 Problem Definition

In this work, we aim to build a robust roadside 3D detector to detect a three-dimensional bounding box of given foreground objects of interest. The origin of the world coordinate system is the projection point of the LiDAR sensor center point on the ground; the x-axis is parallel to the ground and positive forwards, the y-axis is positive to the left, and the z-axis is positive upwards, conforming to the rules of the right-handed coordinate system. It is assumed that all point cloud coordinates have been converted to the world coordinate system for ease of expression.

Formally, we are given the image $I \in R^{H \times W \times 3}$ and point cloud $P \in R^{N \times 4}$ from roadside sensors, whose transformation matrix $E \in R^{3 \times 4}$ and roadside camera's intrinsic matrix $K \in R^{3 \times 3}$ via senor calibration and synchronization. $E \in R^{3 \times 4}$ represents extrinsic parameters from the camera coordinate system to the world coordinate system. The detector is supposed to output a set of predicted 3D bounding boxes $\hat{B}_{world} = \{\hat{B}_1, \hat{B}_2, ..., \hat{B}_n\}$, and the corresponding set of GT boxes is $B_{world} = \{B_1, B_2, ..., B_n\}$. Each 3D bounding box $\hat{B}_i$ is formulated as a vector with 7 degrees of freedom:

$$\hat{B}_i = (x, y, z, l, h, w, \theta) \tag{1}$$

where (x, y, z) is the center location of each 3D bounding box in the world coordinate system. (l, h, w) denotes the box's length, height, and width, respectively. θ

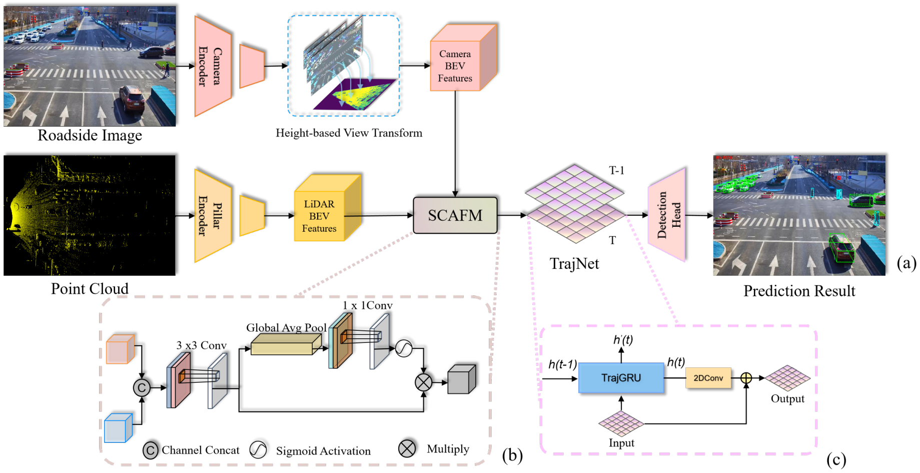


Fig. 2. Overview of the BEVRoad framework. The camera encoder first extracts high-dimensional features from the roadside image. The Height-based View Transform module takes camera features as input and transforms them into BEV features. The point cloud is converted to pillars through the pillar encoder to form LiDAR BEV features represented in the form of pseudo-images. **SCAFM** integrates the BEV feature from two modalities. Then, the BEV feature in the current frame is fused with previous ones through **TrajNet**. Finally, a CenterPoint [32] detection head generates object heatmaps and attributes.

represents the yaw angle of each instance with respect to the z-axis. Specifically, a Camera-LiDAR 3D object detector $FDet$ can be defined as follows:

$$\hat{B}_{world} = F_{Det}(I, P, E, K|\omega) \tag{2}$$

where ω is the learned weights of the detector.

3.2 Overview

Overall Architecture of BEVRoad. As shown in Fig. 2, given a monocular roadside image $I \in R^{H \times W \times 3}$ and point cloud $P \in R^{N \times 4}$ with corresponding extrinsic matrix $E \in R^{3 \times 4}$ and intrinsic matrix $K \in R^{3 \times 3}$, H and W represent the input image's height and width, and N is the number of points. Formally, a 2D convolutional camera backbone with FPN [13] neck aims to map I to the 2D high-dimensional mutil-scale image features $F^{2d} \in^{C_F \times \frac{H}{16} \times \frac{W}{16}}$, where C_F denotes the channel number. Then $\{E, K\}$ are fed into the Height-based View Transform module to lift the monocular image features F^{2d} from the 2D coordinate system to the 3D by calculating the height distribution $H^{pred} \in R^{C_H \times \frac{H}{16} \times \frac{W}{16}}$ and context feature $F^{context} \in R^{C_C \times \frac{H}{16} \times \frac{W}{16}}$, where C_H stands for the height bins and C_C denotes the channel of the context feature. The wedge-shaped 3D features $F^{3D} \in R^{X \times Y \times Z \times C_c}$ are obtained by taking the outer product of H^{pred} and

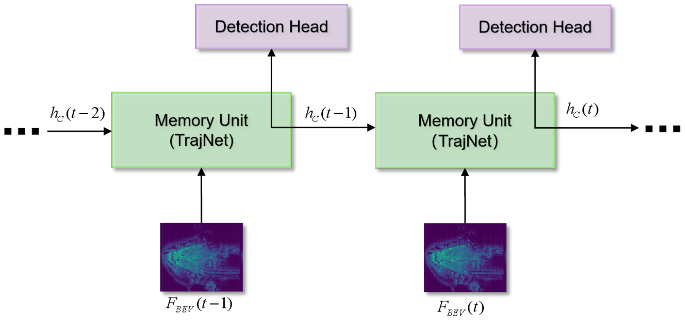


Fig. 3. BEVRoad's temporary-recurrent pipeline. It shows the process of temporary modeling at the BEV feature level, enabling the capture of long-term dependencies in sequential data.

$F^{context}$ and then filling the view frustum. The voxel pooling [20] module compresses the 3D features into the camera BEV features, $F_{Cam} \in R^{X\times Y\times Z\times C_{Cam}}$. As for the LiDAR stream, the point cloud is converted to pillars through the LiDAR encoder to form LiDAR BEV features $F_{LiDAR} \in R^{X\times\ Y\times\ C_{LiDAR}}$ [8]. Then, F_{Cam} and F_{LiDAR} are sent to the BEV encoder consisting of SCAFM and TrajNet. Finally, the 3D detection utilizes fusion BEV features to predict the 3D bounding boxes composed of position (x, y, z), dimension (l, w, h), and orientation θ.

3.3 Spatial-Channel Adaptive Fusion Module.

For effective and impartial modal fusion, our Spatial-Channel Adaptive Fusion Module (SCAFM) combines spatial and channel attention similar to SENet [5], emphasizing learning important features along both the channel and spatial dimensions, as shown in Fig. 2 (b). Given two BEV features $F_{Cam} \in R^{X\times Y\times Z\times C_{Cam}}$ and $F_{LiDAR} \in R^{X\times\ Y\times\ C_{LiDAR}}$ in the unified space, a simple way is to concatenate them and apply channel attention by average pooling and convolution. SCAFM can be formulated as:

$$F_{fusion} = W_{CA}(f^{3\times 3}([F_{Cam}, F_{LiDAR}])) \tag{3}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the channel dimension of two features. $f^{3\times 3}$ represents a convolution operation with a kernel size of 3×3 to reduce the channel dimension of concatenated features into C_{Cam}. W_{CA} selects channel feature information of F using average pooling and convolution to obtain attention map, which can be summarized as follows:

$$W_{CA}(F) = \sigma(f^{1\times 1}(AvgPool(F)) \tag{4}$$

where σ denotes the sigmoid function, $f^{1\times 1}$ represents a convolution operation with a kernel size of 1×1, $AvgPool$ denotes average pooling.

3.4 TrajNet.

Our BEVRoad recycles the system's memory as an RNN to perform temporal modeling, which expands the receptive field in BEV space [6, 16], as depicted in Fig. 3. For motion patterns like rotation and scaling, the local correlation structure, like [22] , of consecutive frames will be difficult. So, we propose the use of a TrajGRU [23] to actively learn the targets' location-variant structure for recurrent connections. The main formulas of TrajGRU [23] are given as follows:

$$\begin{aligned}
&\mathcal{U}_t, \mathcal{V}_t = \gamma(\mathcal{X}_t, \mathcal{H}_{t-1}),\\
&\quad \mathcal{Z}_t = \sigma(\mathcal{W}_{xz} * \mathcal{X}_t + \sum_{l=1}^{L} \mathcal{W}_{hz}^{l} * \mathrm{warp}(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l})),\\
&\quad \mathcal{R}_t = \sigma(\mathcal{W}_{xr} * \mathcal{X}_t + \sum_{l=1}^{L} \mathcal{W}_{hr}^{l} * \mathrm{warp}(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l})),\\
&\quad \mathcal{H}'_t = f(\mathcal{W}_{xh} * \mathcal{X}_t + \mathcal{R}_t \circ (\sum_{l=1}^{L} \mathcal{W}_{hh}^{l} * \mathrm{warp}(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}))),\\
&\quad \mathcal{H}_t = (1 - \mathcal{Z}_t) \circ \mathcal{H}'_t + \mathcal{Z}_t \circ \mathcal{H}_{t-1}.
\end{aligned} \tag{5}$$

where L is the total number of allowed links. $\mathcal{U}_t, \mathcal{V}_t \in R^{L\times H\times W}$ denotes the flow fields that dynamically determine the recurrent connections. γ is a function that generates connection relationships from present features $\mathcal{X}_t$ and saved memory $\mathcal{H}_{t-1}$. The wrap$(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l})$ function selects the positions pointed out by $\mathcal{U}_{t,l}, \mathcal{V}_{t,l}$ from $\mathcal{H}_{t-1}$ via the bilinear sampling [23].

Specifically, our TrajNet is composed of a 2D convolution block and three stacked TrajGRU [23], as Fig. 2 (c) illustrates. In each frame, the present input and the history memory are sent to TrajGRU [23] to obtain the candidate hidden state $\tilde{h}(t)$ and the new memory $h(t)$ at the current moment. A convolution operation is applied to the new memory to restore the dimension of the channel, and then it is added to the input as the final output.

4 Experiments

4.1 Datasets

DAIR-V2X-I. DAIR-V2X [33] is a large-scale, multi-modality dataset for vehicle-infrastructure perception. As the original dataset contains images from vehicles and roadside units, this benchmark consists of three tracks to simulate different scenarios. Here, we focus on the roadside subsets, DAIR-V2X-I and DAIR-V2X-Seq. DAIR-V2X-I contains images and corresponding LiDAR point clouds, with 5042 frames in the training set and 2016 frames in the validation set. Additionally, it is worth nothing that testing examples are not yet publicly disclosed.

Table 1. Comparison with the state-of-the-art on the DAIR-V2X-I validation set. We highlight the best results in red and the second ones in **bold**.

Method	Modality	Veh.(IoU=0.5)			Ped.(IoU=0.25)			Cyc.(IoU=0.25)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PointPillars [8]	L	63.07	54.00	54.01	38.53	37.20	37.28	38.46	22.60	22.49
SECOND [26]	L	71.47	53.99	54.00	55.16	52.49	52.52	54.68	31.05	31.19
MVXNet [24]	C&L	71.04	53.71	53.76	**55.83**	**54.45**	**54.40**	54.05	30.79	31.06
BEVDepth [10]	C	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
MonoGAE [28]	C	**84.61**	75.93	74.17	25.65	24.28	24.44	44.04	47.62	46.75
BEVHeght [29]	C	77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
BEVHeght++ [27]	C	79.31	68.62	68.68	42.87	40.88	41.06	60.76	60.52	60.01
CoBEV [21]	C	81.20	68.86	68.99	44.23	42.31	42.55	61.28	61.00	61.61
SGV3D [30]	C	83.44	72.52	72.81	46.12	44.81	44.92	**65.84**	**65.11**	**65.04**
BEVRoad(Ours)	C&L	81.67	**76.86**	**76.92**	**56.63**	**55.90**	**55.88**	**62.94**	**66.72**	**67.02**
BEVRoad*(Ours)	C&L	**86.73**	**88.50**	**88.52**	40.61	39.86	40.12	57.82	57.67	58.40

-Note: * denotes the voxel size of the BEV grid is [0.8m, 0.8m], and the BEV feature map size is reduced to 176 x 256.

DAIR-V2X-Seq. DAIR-V2X-Seq is the sequential V2X dataset, which includes data frames, trajectories, vector maps, and traffic lights captured from natural scenery. V2X-Seq comprises two parts: the sequential perception dataset, which includes more than 15,000 frames captured from 95 scenarios, and the trajectory forecasting dataset, which contains about 80,000 infrastructure-view scenarios, 80,000 vehicle-view scenarios, and 50,000 cooperative-view scenarios captured from 28 intersections' areas. Here, we only focus on the roadside subset with temporary information and partition DAIR-V2X-Seq into a training set (60%), a validation set (40%) to study temporal modeling in the detection task.

4.2 Experiments Details

Our proposed BEVRoad is trained on a RTX 3090 with the AdamW [15] optimizer and an initial learning rate of 2e-4 for 30 epochs. The image input resolution is set to 864 x 1536. The BEV grid spans [-70.4m, 70.4m] in width and [0.0m, 204.8m] in length, while the voxel size of the BEV grid is [0.4m, 0.4m], resulting in a final resolution of 352 x 512. Unless otherwise specified, the BEVRoad utilizes ResNet50 [4] as its image encoder. In the case of the DAIR-V2X-Seq, the BEV grid spans across [-70.4m, 70.4m], with a length range of [0m, 140.8m]. The BEV map resolution is 176 x 176. For each sequence in the DAIR-V2X-Seq, the first 60% of each sequence is used for training, the next 20% is used for verification, and the last 20% is used for testing. It should be emphasized that we performed comparative experiments with other methods on DAIR-V2X-I [33] and performed ablation studies on DAIR-V2X-Seq.

Table 2. Robustness analysis on the DAIR-V2X-I validation set, including three disturbed factors of roadside cameras (i.e., the focal length, the roll angle, and the pitch angle). We highlight the best results in **bold**.

Method	focal	roll	pitch	Veh.(IoU=0.5)			Ped.(IoU=0.5)			Cyc.(IoU=.5)		
				Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
BEVHeight [29]				77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
	✓			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
	✓	✓	✓	71.71	59.92	59.92	27.81	26.43	26.36	47.42	51.19	51.26
CoBEV [21]				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
	✓	✓	✓	**75.53**	63.46	63.55	30.75	30.08	29.17	51.42	54.78	54.97
BEVRoad(Ours)				81.67	76.86	76.92	56.63	55.90	55.88	62.94	66.72	67.02
	✓			75.33	70.42	70.44	51.33	50.77	50.57	56.51	62.02	62.02
		✓		78.30	74.80	72.76	57.33	56.70	56.65	60.89	66.04	66.37
			✓	78.18	75.25	75.30	54.85	55.52	55.59	60.53	65.10	65.48
	✓	✓	✓	75.12	**70.21**	**70.20**	**50.78**	**50.03**	**49.93**	**55.60**	**60.78**	**59.24**
	w.r.t CoBEV			-0.41	6.75	6.65	20.03	19.95	20.76	4.18	6.00	4.27

4.3 Comparison with the State-of-the-Arts

In the DAIR-V2X-I setting, we compare our BEVRoad with other methods like MVXNet [24], BEVDepth [10], MonoAGE [28], BEVHeight [29], BEVHeight++ [27], CoBEV [21], and SGV3D [30]. Following the established detection metrics employed in previous benchmark datasets, like KiTTi [3], we evaluate the 40-point average precision (AP3D|R40) of predicted 3D boxes, which is further classified into three modes: easy, middle, and hard, depending on the box attributes [1]. The experimental results of each network on the DAIR-V2X-I validation set are listed in Table 1.

As can be seen from Table 1, BEVRoad outperforms all other methods across the board. Specifically, it achieves a performance boost of +4.3%(76.82% vs. 72.52%)% on the vehicle category, +11.71% (56.52% vs. 44.81%) on the pedestrain category, and +1.22% (66.33% vs. 65.11%) on the cyclist category, as compared to the previously best detector.

Notably, we found that voxel size has a significant impact on detection performance. The smaller resolution of the BEV map corresponding to the large voxel size strengthens the characteristics of big targets such as vehicles, making the BEVRoad's detection accuracy (88.50%) in the category of vehicles surpass MonoUNI [7](87.20%) and achieve state-of-the-art results. The smaller voxel size corresponds to a larger resolution of the BEV map, which greatly improves the model's perception of pedestrians and cyclists.

Table 3. Ablation Study on different components of our overall framework on DAIR-V2X-Seq val set.

Exp	Cam	Lidar	TrajNet	GFLOPs	Params(M)	Veh.(IoU=0.5)			Ped.(IoU=0.5)			Cyc.(IoU=.5)		
						Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
(a)	✓			320.0	71.10	73.66	62.14	60.15	19.10	16.20	16.33	34.95	35.40	33.91
(b)	✓	✓		341.0	75.40	85.89	83.89	83.90	25.45	27.60	27.79	53.45	58.84	58.90
(c)	✓		✓	612.0	80.50	75.06	65.50	65.48	18.39	16.54	16.61	38.98	40.20	38.82
(d)	✓	✓	✓	633.0	84.80	**87.40**	**85.14**	**85.15**	**26.43**	**29.83**	**28.83**	**57.83**	**62.96**	**63.04**

4.4 Results on Noisy Extrinsic Parameters

In real-world scenarios, camera parameters always change for various reasons, such as wind, vibrations, human adjustments, and other environmental conditions. Following the simulation approach described in [34], we simulate robustness scenarios where the external parameters of the camera change by introducing offset noise with a $N(0, 1.67)$ distribution to the roll and pitch angles, with the scaling coefficient following a $N(1, 0.2)$ distribution applied to the camera focal length. BEVRoad maintains the best accuracy in the medium difficulty category with noisy camera parameters, which reveals BEVRoad's excellent anti-interference ability, as detailed in Table 2.

4.5 Ablation Studies

In this section, we conduct a series of ablation experiments to investigate the effects of each component of BEVRoad on the DAIR-V2X-Seq dataset.

Effectiveness of Different Components. We evaluate the effectiveness of the SCAFM and the TrajNet through ablations on the DAIR-V2X-Seq val set. As shown in Table 3, the comparison between Exp (a) and Exp (b) clearly demonstrates the efficacy of SCAFM, which greatly enhances the detection ability of the model, making the indicators of vehicle, pedestrian, and cyclist increase significantly. The comparison between Exp (a) and experiment Exp (c) and the comparison between Exp (b) and Exp (d) fully reflect the advantages of TrajNet in temporary modeling, improving the accuracy of vehicles and cyclists observably.

Temporary Fusion Strategy. We also experimented with other traditional methods based on spatio-temporal sequence prediction. However, TrajNet performs the best. ConvGRU is second, and the impact of other techniques like self-attention and SwinLSTM [25] on temporary modeling is insignificant. The results are shown in Table 4.

Impact of Training Sequence Length. We conduct experiments with varying numbers of training frames and show results in Tab. 5. The performance of BEVRoad continues to grow when adding more training frames. Expanding to 12 frames brings limited performance improvement, so we train our models on 8 frames for experimental efficiency.

Table 4. Ablation Study of temporary fusion strategy on DAIR-V2X-Seq val set.

Strategy	AP_{3D}			AP_{BEV}		
	Easy	Mod.	Hard	Easy	Mod.	Hard
No Fusion	23.58	23.62	23.67	32.81	33.80	33.42
Self-Attention	24.11	24.16	24.16	33.93	34.96	34.97
SwinLSTM [25]	24.31	24.80	24.34	34.43	35.14	35.17
ConvGRU	26.47	26.25	26.32	35.64	36.87	36.94
TrajNet	**27.12**	**27.14**	**27.23**	**36.51**	**37.78**	**37.86**

Table 5. Training frames for long-term fusion on DAIR-V2X-Seq val set

Training frames	AP_{3D}			AP_{BEV}		
	Easy	Mod.	Hard	Easy	Mod.	Hard
1	23.58	23.62	23.67	32.81	33.80	33.42
2	25.30	25.73	24.78	34.47	35.99	36.13
4	25.41	26.60	25.60	34.47	35.99	36.13
8	**27.12**	27.14	27.23	**36.51**	37.78	**37.86**
12	26.91	**27.15**	**27.25**	36.23	**38.20**	**37.86**

4.6 Visualization Results.

For the DAIR-V2X-Seq dataset, as shown in Fig. 4, we present the visual comparison results of BEVHeight [29] baseline and BEVRoad in the image view and BEV space. We utilize red boxes to denote true positives, and green for ground truth. Samples (1-3) represent three different newly roadside scenarios from DAIR-V2X-Seq. We use blue rectangles to highlight instances where our BEVRoad significantly outperforms BEVHeight [29]. From the samples in (1-3), we can observe that BEVHeight [29] has significant deviations from the ground truth of objects at medium and long distances. In contrast, BEVRoad performs much better, keeping objects in the correct position.

As depicted in Fig. 5, we further demonstrated the powerful detection capabilities of BEVRoad. The red solid rectangular box on the image represents the missing label in the DAIR-V2X-Seq GT boxes, and our BEVRoad can still accurately detect it, corresponding to the red circle and ellipse circled in the point cloud BEV perspective. The upper left corner of each sample is a visualization of BEV features, which blend point clouds and image modalities. From the partially enlarged BEV Map, we can clearly observe the target vehicle that is blocked in the distance, which confirms BEVRoad's excellent perception capabilities. However, for far-distance and severely truncated obstacles in sample 1 and

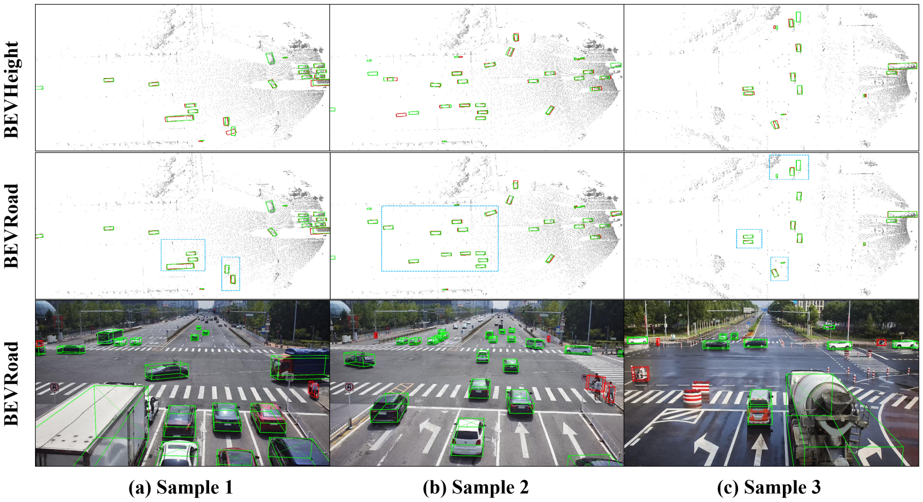


Fig. 4. Visualization Results of BEVHeight and our proposed BEVRoad on the DAIR-V2X-Seq. Samples (1-3) are sourced from the new scenario in the DAIR-V2X-Seq validation set. Blue rectangles are used to highlight instances where BEVRoad significantly outperforms BEVHeight.

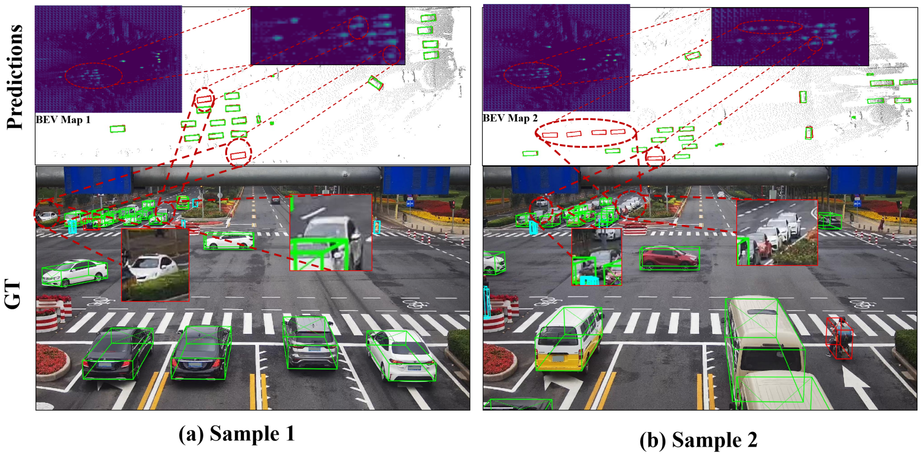


Fig. 5. Visualization examples on DAIR-V2X-Seq. Red solid rectangle is used to highlight instances of missing labels in the DAIR-V2X-Seq GT box. The upper left corner of each sample is a visualization of BEV features.

sample 2, our model suffers from missed detection objects, which is manifested in that our model does not detect all targets.

5 Conclusion

In this paper, we introduce BEVRoad, a cross-modal and temporary-recurrent 3D object detector for roadside perception. BEVRoad achieves state-of-the-art accuracy on the public roadside 3D detection datasets DAIR-V2X-I and DAIR-V2X-Seq [33]. Our framework comprises a camera stream based on BEVHeight [29] and a LiDAR stream that encode raw image and point cloud inputs into features in the unified BEV space, followed by a lightweight spatial-channel adaptive fusion module (SCAFM) to fuse these features across modals. In addition, TrajNet shows the adequacy of recurrent networks as a means to achieve temporary fusion, which turns out to be a more robust solution to overcome the problem of target occlusion. Extensive ablation experiments verify the effectiveness of BEVRoad's components. In the future, we believe that this approach can lead to broader research in infrastructure perception and deliver better results in practical applications.

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
2. Fan, S., Wang, Z., Huo, X., Wang, Y., Liu, J.: Calibration-free bev representation for infrastructure perception. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9008–9013. IEEE (2023)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
6. Huang, R., Zhang, W., Kundu, A., Pantofaru, C., Ross, D.A., Funkhouser, T., Fathi, A.: An lstm approach to temporal 3d object detection in lidar point clouds. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 266–282. Springer (2020)
7. Jinrang, J., Li, Z., Shi, Y.: Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. Advances in Neural Information Processing Systems **36** (2024)
8. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
9. Li, H., Sima, C., Dai, J., Wang, W., Lu, L., Wang, H., Zeng, J., Li, Z., Yang, J., Deng, H., et al.: Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
10. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1477–1485 (2023)

11. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)
12. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems **35**, 10421–10434 (2022)
13. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
14. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. pp. 531–548. Springer (2022)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Lozano Calvo, E., Taveira, B.: Timepillars: Temporally-recurrent 3d lidar object detection (2023)
17. Mao, J., Shi, S., Wang, X., Li, H.: 3d object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision **131**(8), 1909–1963 (2023)
18. Nobis, F., Geisslinger, M., Weber, M., Betz, J., Lienkamp, M.: A deep learning-based radar and camera sensor fusion architecture for object detection. In: 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF). pp. 1–7. IEEE (2019)
19. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 194–210. Springer (2020)
20. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
21. Shi, H., Pang, C., Zhang, J., Yang, K., Wu, Y., Ni, H., Lin, Y., Stiefelhagen, R., Wang, K.: Cobev: Elevating roadside 3d object detection with depth and height complementarity. arXiv preprint arXiv:2310.02815 (2023)
22. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
23. Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c.: Deep learning for precipitation nowcasting: A benchmark and a new model. Advances in neural information processing systems **30** (2017)
24. Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7276–7282. IEEE (2019)
25. Tang, S., Li, C., Zhang, P., Tang, R.: Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13470–13479 (2023)
26. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
27. Yang, L., Tang, T., Li, J., Chen, P., Yuan, K., Wang, L., Huang, Y., Zhang, X., Yu, K.: Bevheight++: Toward robust visual centric 3d object detection. arXiv preprint arXiv:2309.16179 (2023)

28. Yang, L., Yu, J., Zhang, X., Li, J., Wang, L., Huang, Y., Zhang, C., Wang, H., Li, Y.: Monogae: Roadside monocular 3d object detection with ground-aware embeddings. arXiv preprint arXiv:2310.00400 (2023)
29. Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., Chen, P.: Bevheight: A robust framework for vision-based roadside 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21611–21620 (2023)
30. Yang, L., Zhang, X., Li, J., Wang, L., Zhang, C., Ju, L., Li, Z., Shen, Y.: Towards scenario generalization for vision-based roadside 3d object detection. arXiv preprint arXiv:2401.16110 (2024)
31. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: 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. pp. 21341–21350 (2022)
32. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
33. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., 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. pp. 21361–21370 (2022)
34. Yu, K., Tao, T., Xie, H., Lin, Z., Liang, T., Wang, B., Chen, P., Hao, D., Wang, Y., Liang, X.: Benchmarking the robustness of lidar-camera fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3188–3198 (2023)