JUSTC

Article

http://justc.ustc.edu.cn

Received: January 15, 2023; Accepted: April 03, 2023

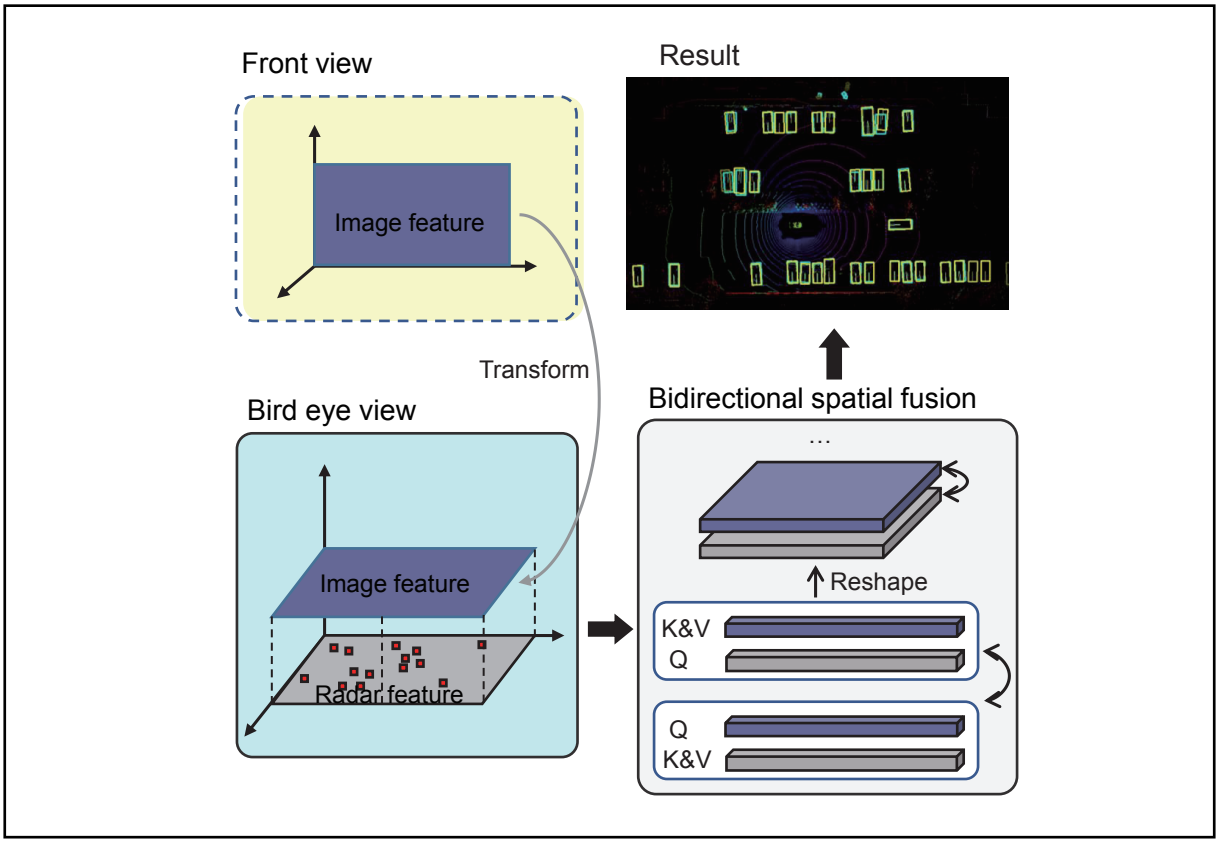
BEV-radar: bidirectional radar-camera fusion for 3D object detection

Yuan Zhao[1], Lu Zhang[2], Jiajun Deng[3], and Yanyong Zhang[1] ✉

[1]*School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China;*
[2]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China;*
[3]*Department of Electrical Engineering, University of Sydney, NSW 2006, Australia*

✉Correspondence: Yanyong Zhang, E-mail: yanyongz@ustc.edu.cn

Graphical abstract



BEV-radar simplifies 3D object detection by aligning camera and radar features in a bird-eye view (BEV) perspective, enhancing fusion through a bidirectional query-based transformer approach for complementary information exchange.

Public summary

■ BEV-radar easily align the multi-modality features adaptively, which is more suitable for fusion of radar and camera.

■ Bidirectional spatial fusion module make the features' representations from different domains towards unification.

■ BEV-radar performs effectively on velocity prediction and reduces 53% error compared to camera-only model.

Citation: Zhao Y, Zhang L, Deng J J, et al. BEV-radar: bidirectional radar-camera fusion for 3D object detection. *JUSTC*, **2024**, 54(1): 0101. DOI: 10.52396/JUSTC-2023-0006

JUSTC

Article

http://justc.ustc.edu.cn

Received: January 15, 2023; Accepted: April 03, 2023

BEV-radar: bidirectional radar-camera fusion for 3D object detection

Yuan Zhao[1], Lu Zhang[2], Jiajun Deng[3], and Yanyong Zhang[1] ✉

[1]*School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China;*
[2]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China;*
[3]*Department of Electrical Engineering, University of Sydney, NSW 2006, Australia*

✉Correspondence: Yanyong Zhang, E-mail: yanyongz@ustc.edu.cn

Cite This: *JUSTC*, **2024**, 54(1): 0101 (8pp)

Read Online

Abstract: Exploring millimeter wave radar data as complementary to RGB images for ameliorating 3D object detection has become an emerging trend for autonomous driving systems. However, existing radar-camera fusion methods are highly dependent on the prior camera detection results, rendering the overall performance unsatisfactory. In this paper, we propose a bidirectional fusion scheme in the bird-eye view (BEV-radar), which is independent of prior camera detection results. Leveraging features from both modalities, our method designs a bidirectional attention-based fusion strategy. Specifically, following BEV-based 3D detection methods, our method engages a bidirectional transformer to embed information from both modalities and enforces the local spatial relationship according to subsequent convolution blocks. After embedding the features, the BEV features are decoded in the 3D object prediction head. We evaluate our method on the nuScenes dataset, achieving 48.2 mAP and 57.6 NDS. The result shows considerable improvements compared to the camera-only baseline, especially in terms of velocity prediction. The code is available at https://github.com/Etah0409/BEV-Radar.

Keywords: 3D object detection; sensor fusion; millimeter wave radar

CLC number: TP399 **Document code:** A

1 Introduction

The perception system in autonomous driving is usually equipped with different types of sensors. Complementary multi-modal sensors avoid unexpected risks but take on new challenges while sensor fusion. Recent works have focused on visual sensors[1], typically providing dense and redundant information. However, visual sensors are usually not stable enough for adverse weather conditions (i.e., rain, snow, and fog). In addition to the high cost, the fusion of visual sensors cannot fully sustain the perception system in variable autonomous scenarios, which requires robustness.

Aside from LiDAR and cameras, radar has aslo been widely used in autonomous scenes for speed measurement and auxiliary location prediction, but rarely in visual tasks due to its physical nature. While stability and penetration benefit from their physical properties, sparse results, noisy features, and lack of vertical information are crucial problems brought by frequently-used automotive radar. Randomly scattered signals among vehicles, buildings, and obstacles obtain high specular reflectivity and multi-path effects. While the complementary characteristics of camera and radar are effective, the fusion strategy faces several challenges. First, the results of the mm-wave radar projected on the image view only have direction and range, which does not provide vertical information and leads to some bias when projected on the camera view. Moreover, the image cannot rely merely on the projected radar depth, as multi-path effectivity produces inaccurate results for radar detection.

Compared to the richer and more accurate information provided by visual sensors, the alignment of features between the camera and the radar is a challenging problem. Without vertical information, some methods[2, 3] rectify the vertical direction in the front view after projecting radar points to image planes. Higher performance leverages on first-stage proposals from the camera and then constructs a soft association between objects and features according to the extrinsic matrix, as shown in Fig. 1. Instead of association methods, transforming both features to bird-eye views (BEV) can extremely relieve the problem, concerning two key points: a more compatible decoupled fusion strategy for radar data and a better promotion for both modalities.

Inspired by BEV fusion methods[4, 5], we implement BEV-radar, an end-to-end fusion approach for radar and cameras, which can be conveniently used for other BEVs for camera baselines. Before fusion, radar encoders are used for pillar extraction and tensor compaction. BEV-radar focuses on inserting dense radar tensors into the BEV image features generated by the camera baseline. Bidirectionally, radar features and image features are promoted to their respective decoders according to cross-attention. Despite the simplicity of the basic idea, the evaluation on the nuScenes dataset performs outstanding results in the 3D object detection benchmarks. It achieves an improvement over the camera-only baselines and performs well even compared to other radar-camera fusion

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

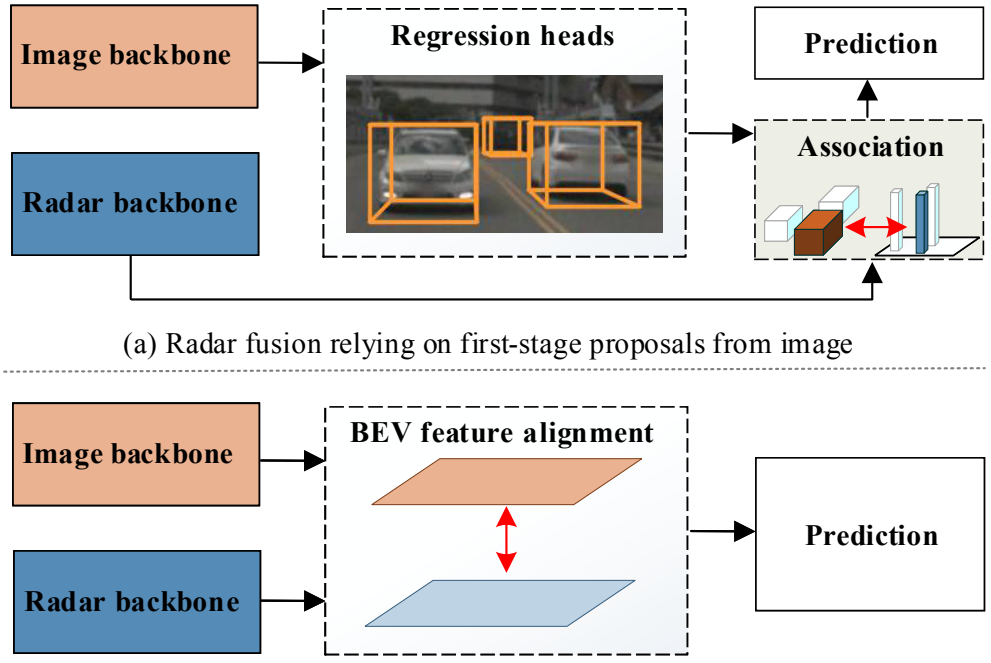


Fig. 1. Comparison between the two alignment methods. (a) Radar fusion methods relying on the first stage proposals: after generating the initial proposals, association methods to their corresponding radar regions is necessary, leading to ignoring of objects which are not detected in the first stage. (b) Our adaptive radar fusion view: instead of aligning proposals from the first stage, features are directly aligned in BEV, thus prediction is guided by multi-modality features.

studies. Besides, for the original intention of the experiment, radar fusion behaves stably with +10% mAP and +15% NDS boost in adverse scenes.

We make the following contributions:

(I) We construct an end-to-end BEV framework for radar and camera fusion. Instead of relying on the first-stage detection results provided by the camera, this integral network instructs a portable and robust type that does not depend strongly on the camera.

(II) We propose a novel bidirectional fusion strategy compared to vanilla cross attention, which is suitable for multi-modal features with spatial relationships. It performs effectively despite the huge diversity of radar and cameras.

(III) We achieve a comparative camera-radar 3D detection performance on the nuScenes dataset. Compared to a single modality, we solve the difficult problem of velocity prediction, which is non-trivial in autonomous.

2 Materials and methods

2.1 Related work

Camera-only 3D detection. Monocular 3D detection requires the estimation of 3D bounding boxes while using a monocular camera. The key question is how to regress the depth information on the 2D view. Earlier works relied on 2D detection networks with additional sub-networks for 3D projection[6, 7]. Several works have attempted to convert RGB information into 3D representations, such as pseudo-LiDAR[8, 9] and orthographic feature transform[10]. Several studies[11] introduced key point detection for centers and used 2D object detection prediction as regression auxiliary. In recent works, camera-only methods directly predicted results on 3D spaces or BEV features[5, 12, 13]. They operated directly on the BEV features transformed from the front view according to calibration.

Camera-fusion 3D detection. The key point of the association modality fusion methods is to find the interrelated spatial relationships among multi-modal sensors. In recent years, fusion approaches have mainly focused on LiDAR and cameras. Some earlier works[14, 15] mapped the data from multi-views into unified types like image or BEV. Pointpainting[1] creatively proposes the segmentation of information from images onto point cloud. Due to the sensitivity to adverse weather conditions, MVDNet[16] first designed a fused area-wise network for radar and LiDAR in a foggy simulated environment. Motivated by the cost of LiDAR, Ref. [17] researched the improvement of fusion on tiny objects with camera and radar, and Ref. [18] introduced the transformer for feature-level fusion. However, the 2D convolution of the projected radar points comes with useless computations and does not take into account the sparsity of the radar. Restricted by the front view, spatial relationships between different modalities rely on the results predicted during the first stage. By transforming features from their respective views to a unified BEV, BEVFusion[4] predicted the depth probabilities for image features and projected the pseudo-3D features to the BEV based on their extrinsic parameters. Transfusion[19] compressed camera features along the vertical axis to initialize the guiding query and the align results of the first stage back to image planes.

2.2 Approach

In this work, we present BEV-radar, a radar-camera fusion

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

framework based on camera-only 3D object detection. As shown in Fig. 2, given a set of multi-view images and sparse radar points as inputs, we extract respective BEV features separately and then decode the features using bidirectional attention modules as inserted fusion decoders called BSF (bidirectional spatial fusion). Instead of simple cross-attention, BSF performs better fusion for both modality features and aligns features from different domains effectively. In the following subsections, we first review the preliminaries for related tasks and then elaborate on the implementation details of the BSF.

2.2.1 Preliminary

Generation of BEV features. Traditional sensor fusion operates on separate views so that the perspective front view and BEV are aligned on the actual pixel-to-pixel spatial relation. However, even with a high-precision extrinsic calibration, projected radar points deviate from the true positions due to the absence of vertical information. Moreover, this pixel-to-pixel spatial alignment is not tight enough due to the geometric distortion of the camera and the sparse attributes of the 3D points. Therefore, a unified BEV representation instead of a geometric association is crucial for sensor fusion.

Transformers for 2D detection. Vision in transformer (ViT)[20] proposed patched images with positional encoding instead of 2D convolution, which makes progress on image feature based on original natural language processing (NLP)[21]. The original attention mechanism is formulated as follows. Given a query embedding f^{q}, key embedding f^{k}, value embedding f^{v}, and dimension of key embedding d^{k}, these inputs will be computed in a single-head attention layer as:

$$\mathrm{Attn}(f^{\mathrm{q}}, f^{\mathrm{k}}, f^{\mathrm{v}}) = \mathrm{softmax}\left(\frac{f^{\mathrm{q}} f^{\mathrm{k}}}{\sqrt{d^{\mathrm{k}}}} \cdot f^{\mathrm{v}}\right). \tag{1}$$

As for the prediction decoder, the promoted detection transformer (DETR)[22] and transformers[13, 23, 24] are widely used for detection tasks based on reforming a set of matched bounding boxes. Thus the usual 3D regression problem is transformed into a bipartite matching problem and the non-maximum suppression (NMS) algorithm is no longer needed.

2.2.2 BEV unified representation

In this part, we state the details of the representation of the two sensors. Transforming raw features extracted from their original data type to BEV is nontrivial for alignment.

To camera. Following BEVDet[5], the BEV camera baseline predicts the depth of multi-view image semantic features from the backbone and feature pyramid network and then transforms all features into a unified BEV grid space relying on the associated extrinsic matrix. Thus, the baseline forms a BEV camera feature map $F_{\mathrm{L}} \in \mathbb{R}^{C\times H\times W}$, which is downsampled from the origin size by 8×, and H, W describe the size of the BEV map. BEV image features provide a global representation for multi-view transformations.

To radar. The radar data format has a completely different style compared to the camera, similar to LiDAR but sparser, with about 300 points per 6 frames. To avoid overly sparse inputs, a sequence of points $R \in \mathbb{R}^{N\times d\times X\times Y}$ is accumulated, where X and Y denote spatial coordinates, d denotes attributes including velocity, and N is the size of the point set. In the absence of vertical information, pillars[25] as feature extraction considerably alleviates the computation of sparse radar data to traverse the BEV plane. Naturally, the unified BEV radar features $F_R \in \mathbb{R}^{C\times H\times W}$ are formed after a linear transformation.

2.2.3 Bidirectional BEV alignment

Traditional sensor fusion first concatenates individual features directly and then uses attention or convolution blocks to extract features from different modalities and align them according to their spatial relationships. However, for BEV radar and image features, sparsity makes it non-trivial to align both modalities spatially only, so we need to generalize each sparse feature. In this section, we instruct a module

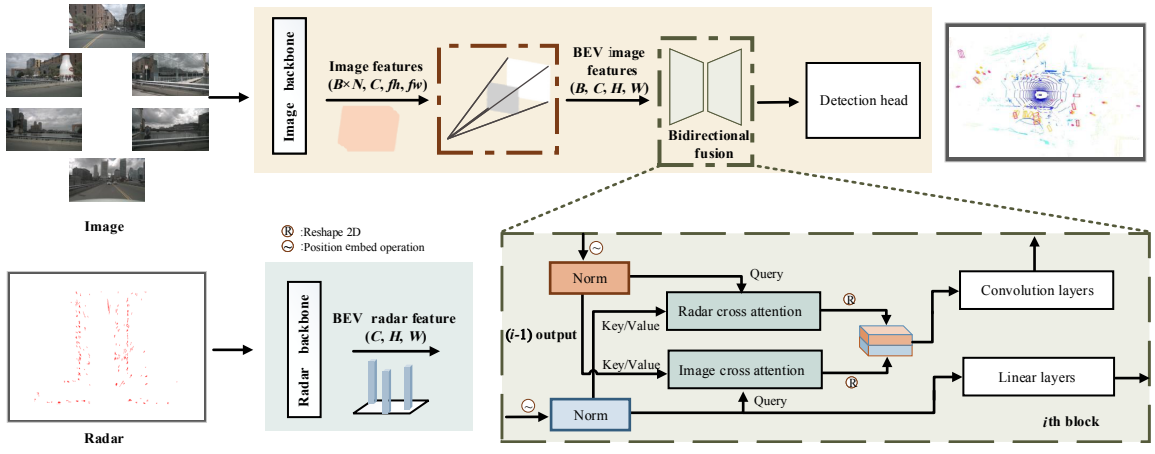


Fig. 2. Overall architecture of framework. Our model is constructed on separate backbones to extract the image BEV features and the radar BEV features. Our BSF (bidirectional spatial fusion) blocks consist of several blocks sequentially: First, a shared bidirectional cross-attention for communicating both modalities. Spatial alignment is followed to localize the radar and camera BEV features. After all blocks, both outputs will be sent in a deconvolution module to descend the channel.

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

consisting of cross-attention and convolution blocks to progressively embed the duplex features in each other, which results in better alignment.

Specifically, a block consists of two parts: an interaction module to communicate each feature, and a convolution-based fusion operation. As shown in Fig. 2, the fusion part can be divided into N equal blocks, and the positional embedding operation is applied before fusion. For the camera branch in the ith block, given a C dimensional camera BEV feature map $F \in \mathbb{R}^{C\times H\times W}$ as query, the radar BEV feature map $F \in \mathbb{R}^{C\times H\times W}$ is used as key and value, and vice versa for the radar branch. We use deformable cross-attention[24] to remedy the computational cost caused by the sparsity of BEV features, which can be formulated as follows:

$$F_C^i = H(\mathrm{Attn}(\mathrm{norm}(F_C^{i-1}),\ \mathrm{norm}(F_R^{i-1}))\oplus \mathrm{Attn}(\mathrm{norm}(F_R^{i-1}),\ \mathrm{norm}(F_C^{i-1}))). \tag{2}$$

$$F_R^i = G(\mathrm{Attn}(\mathrm{norm}(F_R^{i-1}),\ \mathrm{norm}(F_C^{i-1}))). \tag{3}$$

Different from the NLP vanilla transformer, spatial information obtaining objects' location is vital for detection tasks F_{out}. Designed for 2D structures, convolution kernels are better at extracting local spatial correlations than 1D attention. F_{out} is transformed to image style again and sent to convolution blocks, which are then patched again before the next $(i+1)$th block. At the same time, a transform block for F_R remains synchronized with F_{out} for F_C, and they are returned separately as the next inputs. In this way, multi-blocks increase the fitness of F_C and F_R, while bidirectional design updates obtain the alignment of feature domains. In each block, convolution layers are required to extract the local spatial relations, see Section 3.2.2 for a related verification.

2.2.4 Prediction heads and losses

The BEV fusion features are applied to 3D object detection prediction heads. Referring to transfusion[19], we simply use the class embedding heatmap transformed from the fusion features as query initialization to predict the centers for all objects in each scene. A vanilla transformer is used as the decoder for DETR[22] prediction parts through the Hungarian algorithm[26], and we set the regularized matching loss function by a weighted sum of the classification, regression, and IoU calculation:

$$\mathcal{L}_{\mathrm{tot}} = \lambda_1\mathcal{L}_{\mathrm{cls}} + \lambda_2\mathcal{L}_{\mathrm{reg}} + \lambda_3\mathcal{L}_{\mathrm{IoU}}, \tag{4}$$

where λ_1, λ_2, and λ_3 represent each coefficient parameters, and $\mathcal{L}_{\mathrm{cls}}$, $\mathcal{L}_{\mathrm{reg}}$, and $\mathcal{L}_{\mathrm{IoU}}$ are individual loss function for above.

3 Results and discussion

3.1 Implementation details

Training. This end-to-end work is implemented on the open-sourced MMDetection3D[27] in PyTorch. For nuScenes, following M^{2}BEV[28], we use a pre-trained model on NuImage with swin-transformer as the camera backbone. BEVDet[5] is chosen as our image BEV baseline, and the settings keep the same. Considering the synchronization of the real systems, we accumulate a sequence of radar points around 6 frames to resolve its sparsity and eliminate the effect of ego-motion. Pillars[25] are generated from radar points and scattered onto the BEV grids as a pseudo image. The pillar size is set to (0.1 m, 0.1 m, and 8 m). We adopt the random flip, the rotate strategies, and CBGS (class-balanced grouping and sampling) dataset configuration on the data augmentation. Optimization is set to AdamW with official weight decay and the learning rate is 0.0001 for 4 NVIDIA RTX A6000.

Testing. As a result of the transformer detection head, we use all the outputs through a bipartite matching algorithm instead of traditional NMS. Evaluation metrics are formed as mean average precision (mAP) across 10 foreground classes and the nuScenes detection score (NDS) without test-time augmentation.

3.2 Experiments

In this section, we compare our work with other state-of-the-art methods and fusion strategies on nuScenes firstly. Then we evaluate the ablation studies for the modules designed. Moreover, we design extra variable weather conditions to compare our visual baseline to show its robustness against severe scenes. The visual baseline is BEVDet-Tiny for camera-only detection. By the way, as a separate fusion module, our framework can be easily extended to support more sensors.

NuScenes dataset. The nuScenes dataset is a large-scale outdoor autonomous-driving dataset for various 3D visual tasks. For our work, particularly, nuScenes is the unique dataset whose sensors include radar while with 3D ground truth. It consists of 700, 150, and 150 scenes for training, validation, and testing, respectively. Particularly, 6 calibrated multi-view cameras cover the surround horizontal field of view (FOV) with overlapping, while 5 calibrated radars are also distributed around ego on average. The model is evaluated according to mAP and NDS for 10 common classes. Instead of 3D IoU, AP for nuScenes is defined as the BEV center distance, concerning 0.5 m, 1 m, 2 m, and 4 m across 10 classes, and NDS is a weighted sum of mAP and other attribute metrics including velocity.

3.2.1 Main results

Our model is submitted to the nuScenes evaluation server and achieves competitive performance on its metrics. As shown in Table 1, our model outperforms camera-only baselines by 17% mAP with a 10.2 FPS inference speed. In contrast to other camera-radar fusion methods, BEV-radar achieves 7% mAP and 5% NDS boost in the *test* split. Moreover, the mean average velocity error (mAVE) metric is attractive due to the complementary fusion motivation. However, accurate velocity prediction is quite difficult for single-frame camera methods. Thus, radar fusion should remedy this burden for camera-only detection. Our results show a large improvement in velocity prediction of 14%–24% compared to other radar fusion models. In addition to evaluating the robustness, one of the complementary attributes provided by radar fusion, we design several experiments in different weather conditions, as shown in the Section 3.2.3.

Table 2 shows the per-class mAP comparison with methods in various modalities on the *val* set. With a similar

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

Table 1. State-of-the-art comparison on nuScenes *test* set. "L", "C", and "R" denote LiDAR, camera, and radar, respectively. † represents test time augmentation. Particularly, BEVDet-Tiny is our BEV camera-only baseline, and CenterNet is for CenterFusion and CRAFT. δ represents a SECOND[29] network as decoder compared to base version. The bold numbers represent the optimal values for the respective indicator.

Method	Modality	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS
CenterPoint[30]	L	67.3	60.3	0.262	0.239	0.361	0.288	0.136	61
PointPillars[25]	L	45.3	30.5	0.517	0.290	0.500	0.316	0.368	30
FCOS3D[11]	C	37.2	29.5	0.806	0.268	0.511	1.315	0.170	1.7
PGD[31]†	C	44.8	38.6	0.626	0.245	0.451	1.509	0.127	1.4
CenterNet[32]	C	40.0	33.8	0.658	0.255	0.629	1.629	0.142	–
BEVDet-Tiny[5]†	C	39.2	31.2	0.691	0.272	0.523	0.909	0.247	15.6
CenterFusion[33]†	C+R	44.9	32.6	0.631	**0.261**	0.516	0.614	0.115	–
CRAFT[34]†	C+R	52.3	41.1	0.467	0.268	0.456	0.519	**0.114**	4.1
Ours	C+R	54.3	44.3	0.501	**0.261**	0.492	0.408	0.129	**10.4**
Ours-δ	C+R	**57.6**	**48.2**	**0.444**	0.262	**0.452**	**0.371**	0.126	10.2

performance of BEVDet-Tiny and CenterNet for camera-only baselines, our fusion work achieves significant progress for radar fusion. Considering velocity attribute, for dynamic types (i.e., car, truck, bus, pedestrian, motorcycle, and bicycle) and static (i.e., barrier and traffic cone) classes divided separately, results also show a gap (14%–20%), since the valid radar points of dynamic objects are distinguished from background interference. More precisely, radar fusion performance is better for metallic and large objects (i.e., car, truck, and bus) than non-metallic and small objects (i.e., pedestrian and bicycle), caused by RCS (radar cross section). Accuracy for non-metallic and static objects is more camera dependent, and radar less helpful. In particular, though trailers and construction vehicles belong to the large, metallic, and dynamic classes, it is hard to achieve comparable satisfactory performance as they occur infrequently in the nuScenes dataset.

3.2.2 Ablation studies

We conduct several ablation studies on the nuScenes validation set to verify the effectiveness of the proposed components. Table 3 reports the improvement of bidirectional fusion (BF) or bidirectional spatial fusion (BSF) under the different settings of the model depth and training epochs. (ⅰ) The first row shows the result of the model which uses the concatenation followed by convolution layers instead of BF or BSF as the fusion module. (ⅱ, ⅲ) When the concatenation is removed and BF is added, mAP and NDS boost together as the number of layers increases. To compare BF and BSF, (ⅳ) shows the performance of the gap with respect to the spatial policy. Specifically, we set the number of fusion blocks to 3, showing that the model performs better in this setting. (ⅰ–ⅴ) show our BSF blocks work well for both modalities, though the features are not similar.

To support the idea that radar can enhance the accuracy of long-range object detection, Table 4 plots the improvement brought by radar fusion. The accuracy of the car gradually decreases as the distance increases on account of image resolutions. Our fusion method provides a performance boost for distant regions, where radar points are able to travel but hardly for the visual camera. Even for objects with a distance of 40 m, the AP of car brings 20% gain. As we discuss in Section 3.2.1, benefit from dynamic velocity and metallic large size, the classes in this table are promoted more in the range of 20–40 m (+14%–20%), where radar works unaffected by distance.

3.2.3 Robustness against weather conditions

We design two experiments to demonstrate the robustness of the proposed fusion model. Since only three submissions are allowed for the evaluation of the nuScenes dataset for the *test* dataset, the special experimental dataset is selected from

Table 2. Per-class AP comparison on nuScenes *val* set. Mod. represents modality. "C.V.", "M.C.", and "T.C." represent construction vehicle, motorcycle, and traffic cone evaluation, respectively. The bold numbers represent the optimal values for the respective indicator, and the numbers in parentheses represent the improvement that the method brings compared to its baseline.

Method	Mod.	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	Bicycle	T.C.	Barrier
PointPillars[25]	L	79.9	35.7	42.8	26.1	5.5	71.7	39.4	10.6	33.4	52.0
FCOS3D[11]	C	47.9	23.3	31.4	11.2	5.7	41.1	30.5	30.2	55.0	45.5
CenterNet[32]	C	48.4	23.1	34.0	13.1	3.5	37.7	24.9	23.4	55.0	45.6
BEVDet-Tiny[5]	C	51.2	22.3	31.3	16.0	7.2	34.5	27.3	22.5	50.0	49.8
CenterFusion[33]	C+R	$52.4_{(+4.0)}$	$26.5_{(+3.4)}$	$36.2_{(+2.2)}$	$15.4_{(+2.3)}$	$5.5_{(+2.0)}$	$38.9_{(+1.2)}$	$30.5_{(+5.6)}$	$22.9_{(-0.5)}$	$56.3_{(+1.3)}$	$47.0_{(+2.6)}$
CRAFT[34]	C+R	$69.6_{(+21.2)}$	$37.6_{(+14.5)}$	$47.3_{(+13.3)}$	$20.1_{(+7.0)}$	$10.7_{(+7.2)}$	$46.2_{(+8.5)}$	$39.5_{(+14.6)}$	$31.0_{(+7.6)}$	$57.1_{(+2.1)}$	$51.1_{(+5.5)}$
Ours	C+R	$\mathbf{72.1}_{(+20.9)}$	$\mathbf{43.0}_{(+20.7)}$	$\mathbf{49.0}_{(+17.7)}$	$21.6_{(+5.6)}$	$14.0_{(+6.8)}$	$44.5_{(+10.0)}$	$41.1_{(+13.8)}$	$\mathbf{42.4}_{(+19.9)}$	$57.4_{(+4.0)}$	$53.9_{(+4.1)}$

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

Table 3. Ablation of the bidirectional fusion module. BF: bidirectional fusion base module. BSF: bidirectional spatial fusion.

	Concat	BF	BSF	#Layers	#Epochs	mAP	NDS
i	√			3	12	38.4	48.7
ii		√		1	12	39.0	45.0
iii		√		3	12	41.3	50.0
iv			√	3	12	42.6	52.4
v			√	3	20	43.3	53.4

Table 4. The AP breakdown over BEV distance between object center and ego vehicle on the camera-only model, while decreases slowly on fusion model.

Class	Modality	<20 m	20–40 m	>40 m
Car	C	76.2	57.3	46.4
	C+R	87.4$_{(+11.2)}$	76.9$_{(+19.6)}$	66.5$_{(\mathbf{+20.1})}$
Truck	C	42.6	29.4	20.3
	C+R	61.1$_{(+18.5)}$	49.1$_{(\mathbf{+19.7})}$	37.2$_{(+16.9)}$
Bus	C	58.6	38.8	28.8
	C+R	64.2$_{(+5.6)}$	54.4$_{(\mathbf{+15.6})}$	42.9$_{(+14.1)}$
Bicycle	C	40.2	22.9	19.9
	C+R	49.2$_{(+9)}$	36.8$_{(\mathbf{+13.9})}$	33.3$_{(+13.4)}$

the validation dataset based on the description of each scenario. There are 5417, 602, and 1088 samples separately for the day, night, and raining times. All the mentioned parameters are kept the same as before except for dataset type. Table 5 shows the robustness to weather and sight conditions, with the performance for night (+10%) and raining (+12%) times providing an intuitive comparison.

The camera-only model is severely affected by the sight condition, which is reduced by 20%. The sight condition does not mean that prediction error occurs only in dark environments but also in illuminated scenes due to the reflection of headlights, as shown in Fig. 3a. On rainy times, the HD waterproof camera is almost unaffected by the rain, except when the raindrops fall right on the camera lens, as shown in Fig. 3b. The examples above have one thing in common: the radar can still work well, therefore accurate distance prediction boosts the fusion. Compared to the camera-only model, our fusion model brings a better performance mainly when the camera is working as normal, which means that the fusion between the camera and the radar, the camera determines the lower bound of the accuracy of the fusion model. As a non-visual sensor type, the task of 3D object detection is a challenge for sparse radar sensors alone. However, the detection results ignored by the camera but provided by the radar fusion obtain an unsatisfactory confidence score, due to the nature of the radar that can not work alone but serve as a supplementary sensor.

Table 5. mAP breakdown while camera is influenced by weather or sight. This experiment dataset is separated from nuScenes *val* dataset according to descriptions for scenes' attributes.

Modality	Day	Night	Rainy
C	32.2	13.2	32.8
C+R	43.3$_{(+11.1)}$	23.2$_{(+10.0)}$	45.2$_{(+12.4)}$

3.2.4 Qualitative results

The visualization of compared results between camera-only and camera-radar fusion models are shown in Fig. 4. By contrast, the fusion method precisely refines the image proposals and completes the objects which are not correctly recognized by the camera-only model. The front perspective view of the

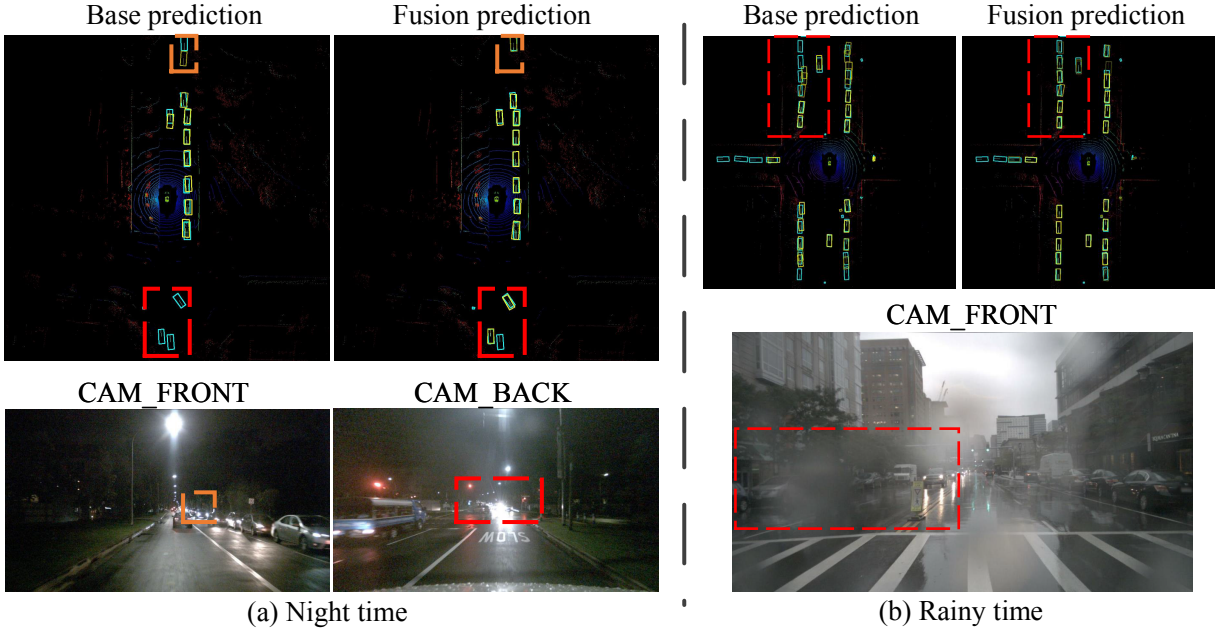


Fig. 3. In the first row, base prediction and fusion prediction separately represent the camera-only model and the radar-fusion model on BEV. The ground truth is plotted as blue boxes and the prediction boxes are plotted as yellow boxes, with lighter colors indicating higher confidence scores. The bottom row shows a visualization of the camera views for this frame, with the corresponding regions of interest marked by dashed boxes of the same color.

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

Fig. 4. Qualitative analysis of detection results. 3D bounding box predictions are projected onto images from six different views and BEV respectively. Boxes from different categories are marked with different colors and without ground truth. For BEV visualization, yellow means predicted boxes and blue ones are ground-truth, while LiDAR points are visualized as background.

camera may exhibit some errors, which are attributed to calibration inaccuracies and height discrepancies. The distinguishing characteristic of BEV-radar lies in its precise representation of position and dimensions in the BEV angle. Consequently, subsequent systems will also work on the BEV framework, making the projection error in the frontal view tolerable.

Most of the objects are detected by the camera-only model, however, there are errors in either the center or the size. In particular, the radar-fusion model is better at orientation prediction due to the accurate velocity auxiliary. As we discussed in Section 3.2.3, the radar-fusion method can remedy some errors made by the camera, but not all. Although some of the ignored objects are successfully detected by the radar, the fusion method gets lower confidence scores without camera judgments.

4 Conclusions

In this paper, we propose an end-to-end robust camera-radar 3D detection framework with a bidirectional transformer-based fusion strategy to adaptively embed radar features while preserving their spatial relationship. Unlike other radar fusion methods, which typically require prior 3D object detection from camera baselines, our approach does not completely rely on visual results. However, as complementary BEV features enhance visual sensors, it is portable to other multi-sensor frameworks. Our work sets high-performance results on the nuScenes detection, and extensive experiments demonstrate the robustness of radar fusion. We also discuss the effectiveness of sensor fusion in different weather or sight conditions that realistic system concerns, and we hope that BEV-radar will inspire practical applications.

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yuan Zhao is currently an Algorithm Engineer in Horizon Robotics. She received her bachelor's degree from Sichuan University in 2020 and master's degree from the University of Science and Technology fo China in 2023. Her research interest mainly focuses on multi-modality sensor fusion, computer vision, and autonomous driving.

Yanyong Zhang is currently a Professor at the School of Computer Science and Technology in the University of Science and Technology of China (USTC). She received her B.S. degree from USTC in 1997 and Ph.D. degree from Pennsylvania State University in 2002. Her research mainly focuses on multi-modality sensor fusion and cyber physical systems.

References

[1] Vora S, Lang A H, Helou B, et al. PointPainting: Sequential fusion for 3D object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 4603–4611.

[2] Long Y, Morris D, Liu X, et al. Radar-camera pixel depth association for depth completion. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, **2021**: 12502–12511.

[3] Nobis F, Geisslinger M, Weber M, et al. A deep learning-based radar and camera sensor fusion architecture for object detection. In: 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF). Bonn, Germany: IEEE, **2019**: 1–7.

[4] Liu Z, Tang H, Amini A, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. arXiv: 2205.13542, **2022**.

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101

[5] Huang J, Huang G, Zhu Z, et al. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. arXiv: 2112.11790, **2021**.
[6] Xu B, Chen Z. Multi-level fusion based 3D object detection from monocular images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 2345–2353.
[7] Kundu A, Li Y, Rehg J M. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 3559–3568.
[8] You Y, Wang Y, Chao W L, et al. Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving. In: Eighth International Conference on Learning Representations, **2020**.
[9] Wang Y, Chao W L, Garg D, et al. 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). Long Beach, USA: IEEE, **2020**: 8437–8445.
[10] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3D object detection. arXiv: 1811.08188, **2018**.
[11] Wang T, Zhu X, Pang J, et al. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE, **2021**: 913–922.
[12] Li Z, Wang W, Li H, et al. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Avidan S, Brostow G, Cissé M, et al. Editors. Computer Vision–ECCV 2022. Cham: Springer, **2022**: 1–18.
[13] Wang Y, Guizilini V, Zhang T, et al. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In: 5th Conference on Robot Learning(CoRL 2021). London, UK: CoRL, **2021**: 1–12.
[14] Chen X, Ma H, Wan J, et al. Multi-view 3D object detection network for autonomous driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, **2017**: 6526–6534.
[15] Qi C R, Liu W, Wu C, et al. Frustum PointNets for 3D object detection from RGB-D data. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 918–927.
[16] Qian K, Zhu S, Zhang X, et al. Robust multimodal vehicle detection in foggy weather using complementary LiDAR and radar signals. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, **2021**: 444–453.
[17] Chadwick S, Maddern W, Newman P. Distant vehicle detection using radar and vision. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, **2019**: 8311–8317.
[18] Cheng Y, Xu H, Liu Y. Robust small object detection on the water surface through fusion of camera and millimeter wave radar. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, **2022**: 15243–15252.
[19] Bai X, Hu Z, Zhu X, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, **2022**: 1080–1089.
[20] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, **2021**: 9992–10002.
[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., **2017**: 6000–6010.
[22] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, et al. editors. Computer Vision–ECCV 2020. Cham: Springer, **2020**: 213–229.
[23] Zhang R, Qiu H, Wang T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection. arXiv: 2203.13310, **2022**.
[24] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv: 2010.04159, **2020**.
[25] Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, **2019**: 12689–12697.
[26] Kuhn H W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, **1955**, *2*: 83–97.
[27] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, **2020**. Accessed December 1, 2022.
[28] Xie E, Yu Z, Zhou D, et al. M^{2}BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation. arXiv: 2204.05088, **2022**.
[29] Yan Y, Mao Y, Li B. SECOND: Sparsely embedded convolutional detection. *Sensors*, **2018**, *18* (10): 3337.
[30] Yin T, Zhou X, Krähenbühl P. Center-based 3D object detection and tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, **2021**: 11779–11788.
[31] Wang T, Zhu X, Pang J, et al. Probabilistic and geometric depth: Detecting objects in perspective. In: Proceedings of the 5th Conference on Robot Learning. PMLR, **2022**, 164: 1475–1485.
[32] Duan K, Bai S, Xie L, et al. CenterNet: keypoint triplets for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, **2020**: 6568–6577.
[33] Nabati R, Qi H. CenterFusion: center-based radar and camera fusion for 3D object detection. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, **2021**: 1526–1535.
[34] Kim Y, Kim S, Choi J W, et al. CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer. arXiv: 2209.06535, **2022**.

DOI: 10.52396/JUSTC-2023-0006
JUSTC, **2024**, 54(1): 0101