2024 7th International Symposium on Autonomous Systems (ISAS) | 979-8-3503-6317-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS61044.2024.10552581

Fusing LiDAR and Radar with Pillars Attention for 3D Object Detection

1st Hanchen Tai
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
22232108@zju.edu.cn

2nd Yijie Qian
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
22332148@zju.edu.cn

3rd Xiao Kang
China North Vehicle Research Institute
Beijing, China
kangxiaotop1@126.com

4th Liang Liu*
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
leonliuz@zju.edu.cn

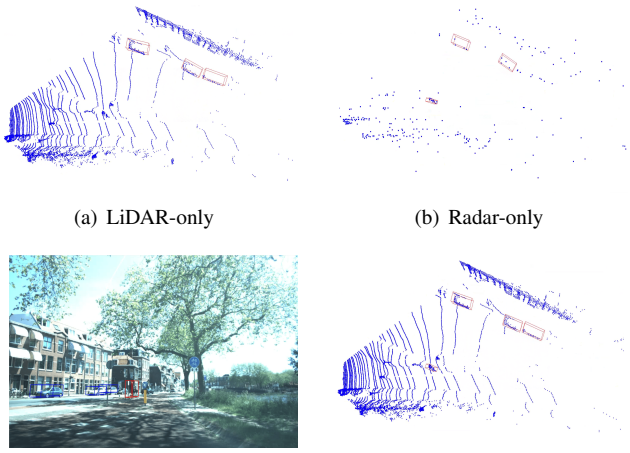
5th Yong Liu*
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
yongliu@iipc.zju.edu.cn

***Abstract*—In recent years, LiDAR has emerged as one of the primary sensors for mobile robots, enabling accurate detection of 3D objects. On the other hand, 4D millimeter-wave Radar presents several advantages which can be a complementary for LiDAR, including an extended detection range, enhanced sensitivity to moving objects, and the ability to operate seamlessly in various weather conditions, making it a highly promising technology. To leverage the strengths of both sensors, this paper proposes a novel fusion method that combines LiDAR and 4D millimeter-wave Radar for 3D object detection. The proposed approach begins with an efficient multi-modal feature extraction technique utilizing a pillar representation. This method captures comprehensive information from both LiDAR and millimeter-wave Radar data, facilitating a holistic understanding of the environment. Furthermore, a Pillar Attention Fusion (PAF) module is employed to merge the extracted features, enabling seamless integration and fusion of information from both sensors. This fusion process results in lightweight detection headers capable of accurately predicting object boxes. To evaluate the effectiveness of our proposed approach, extensive experiments were conducted on the VoD dataset. The experimental results demonstrate the superiority of our fusion method, showcasing improved performance in terms of detection accuracy and robustness across different environmental conditions. The fusion of LiDAR and 4D millimeter-wave Radar holds significant potential for enhancing the capabilities of mobile robots in real-world scenarios. The proposed method, with its efficient multi-modal feature extraction and attention-based fusion, provides a reliable and effective solution for 3D object detection.**

***Index Terms*—3D Object Detection, Multiple Sensors Fusion**

I. Introduction

While there has been remarkable progress in 2D object detection [1, 16, 17] leading to significant advancements in robotics perception, when it comes to tasks involving 3D environmental perception, the use of point clouds which provide rich geometric features but less appearance cues has proven to be more effective in supporting semantic annotation and scene understanding [9, 18, 24]. Further research indicates that when relying solely on LiDAR for perception, the performance will be limited by the lack of appearance information from a single sensor and the sparse point cloud description.



(a) LiDAR-only (b) Radar-only

(c) LiDAR&Radar

Fig. 1. Example 3D detection result from the VoD validation set. (a): LiDAR detection results based on PointPillars [9], missing the detection of a cyclist which is a small object.(b):Radar detection results based on PointPillars, missing the detection of a car which is stationary. (c):LiDAR&Radar detection results based on our fusion model, which maintain detection for large objects while enhance sensitivity to moving objects.

To enhance the robustness of algorithms in coping with diverse driving environments, some works have focused on the multi-sensors for object detection, mostly fusing camera images with LiDAR point clouds [7, 10]. However, camera and LiDAR systems are susceptible to adverse conditions, such as dust storms or precipitation, posing significant risks for autonomous driving. Given the strong penetrative capabilities of 4D Radar, which enables long-range observations in harsh conditions, the multi-modal fusion of LiDAR and Radar becomes both necessary and meaningful.But research on the fusion of LiDAR and Radar for 3D object detection

is still scarce, primarily due to the rarity of publicly available datasets containing ample LiDAR and Radar data with high-quality annotations. Additionally, the disparity in the quantity of points between LiDAR and Radar data poses challenges for effective fusion.

Addressing these issues, the presented Pillar Attention Fusion (PAF) module of LiDAR and Radar algorithm in our work effectively fuses multi-model information by introducing an attention mechanism. This mechanism enhances features within a single modality using the channel attention module and fuses pseudo-image features of LiDAR and Radar through a spatial attention module. We conducted experiments on View-of-Delft (VoD)[13] which is a high-quality multi-sensor automotive dataset for multi-class 3D object detection. Referring to the official settings of the VoD dataset, we evaluated the model proposed in this paper for three classes of objects: cars, pedestrians, and cyclists under different difficulty levels. The results are significantly improved compared with the single-modal 3D object detection using PointPillars [9] as the baseline.

The contributions are summaried as follow:

- We propose a novel framework for multi-modal 3D object detection with LiDAR and 4D millimeter-wave Radar.
- Our proposed pillar attention fusion module effectively integrates LiDAR and Radar point cloud data, while maintaining the detection efficiency for pure LiDAR baselines, integrating Radar to improve the model's detection ability, especially for moving objects.
- Our experiments on the VoD dataset show that the detection results of our multi-modal fusion method are significantly better than single modality methods.

II. Related Work

A. Multi-modal 3D Object Detection

In recent years, 3D object detection algorithms based on multi-modal fusion have achieved fruitful results. Proposal-level fusion represented by MVOD [3] and AVOD [8] first generates respective 3D proposals on different modal data, and then combines these 3D proposals into deep-fusion modules, the final candidate area is generated to complete the subsequent objects classification and 3D-box regression tasks. Result-level fusion represented by F-PointNets [14] and F-ConvNet [22] first uses a SOTA model in 2D target detector to obtain the 2D detection results of the image, and then converts these 2D detection frames into 3D frustums through the projection matrix between multi-sensors.Finally detect the point cloud in these frustums through 3D detection technology to achieve 3D box regression. Point-level fusion, represented by PointPainting [20] and PointAugmenting [21], first establishes the pixel-by-pixel correspondence between the point cloud and the image, and then attaches the score of the detection or segmentation in image through the 2D perception module to the corresponding point of its pixel to obtain the painted or augmented point cloud data. After that, input the painted point cloud into any advanced single-modal 3D object detection network to achieve better results than before.

It is worth mentioning that our survey shows that there are very few papers related to 3D object detection that integrate the different modal data of LiDAR and Radar, but Radar can well complement LiDAR in detecting moving objects under strong confrontation conditions. Therefore, the contribution of our work is to propose an attention-based framework that fuses LiDAR and Radar, and proves that it can achieve better performance than single modality through experiments on the VoD dataset.

B. Feature Fusion with Attention

The attention mechanism is a method inspired by the human visual system, which has proven successful in various computer vision tasks such as image classification, object detection, semantic segmentation, face recognition, and medical image processing [5, 6, 23]. Specifically, this mechanism adaptively weights features in order to guide attention to key areas in an image while disregarding irrelevant parts. The derivatives of attention mechanism can be classified as four basic methods of channel attention, spatial attention, temporal attention and branch attention with the two mixed methods of channel&spatial attention and spatial&temporal attention according to the data domain.

Different channels in feature maps usually represent different objects, so the channel attention which can adaptively recalibrate the weight of each channel is regarded as an instance selection process deciding to focus on which objects in a scene [6]. Similar to the channel attention, the spatial attention can also be considered as an adaptive spatial region selection mechanism choosing which locations in the space to focus on [26]. As a dynamic time selection mechanism, the temporal attention can decide which frames to concentre on in a temporal domain. Previous work often emphasize how to simultaneously capture short-term and long-term cross-frames feature dependencies [12]. A relatively novel one is the branch attention which focuses on specific branches within a multi-branch network like a dynamic branch selection. On the basis of the above primary methods, the channel&spatial attention combines the advantages of the channel attention and spatial attention, selecting critical objects and regions while emphasizing spatial and channel information features [23]. Spatial&temporal attention blends the advantages of spatial attention and temporal attention to adaptively select important regions and key frames. Some works [4] calculate temporal and spatial attention separately, while others [5] create joint spatial and temporal attention maps.

Compared to previous works that focused solely on attention within a single modality, our study aims to enhance the fusion effect across multiple modalities through the attention mechanisms. By integrating information from various sensors, we can get richer insights and more robust performance in complex tasks.

III. Fusion of LiDAR&Radar for Object Detection

In this work, we propose an object detection framework based on PointPillars that fuses LiDAR and Radar point

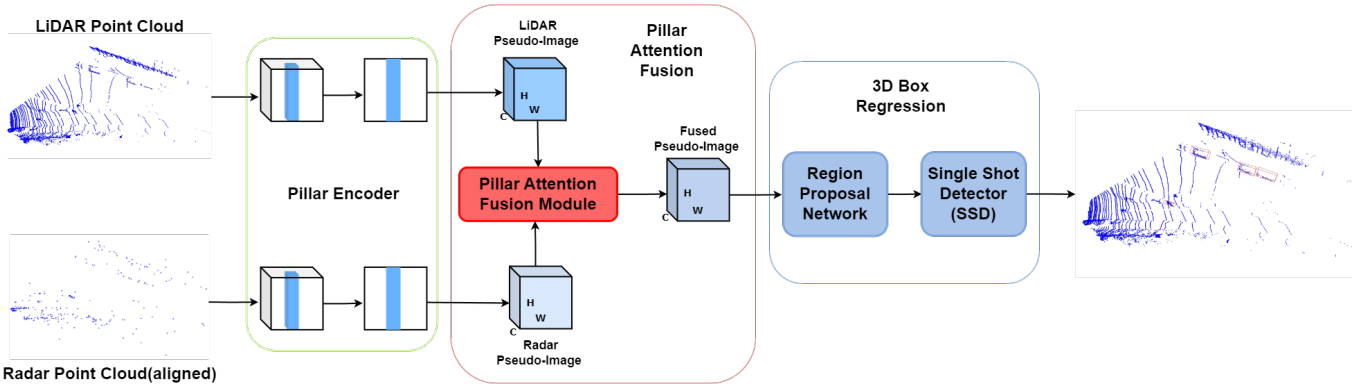


Fig. 2. The overall framework of our proposed LiDAR&Radar fusion model.(1)Pillar Encoder: Encode the input point clouds of LiDAR and Radar into pillars and generating corresponding pseudo-images. (2)Pillar Attention Fusion(PAF): The core module we proposed for fusing pseudo-images from LiDAR and Radar, which is detailed in Fig. 3. (3)3D Box Regression: Adopting the subsequent networks in PointPillars to achieve 3D detection through our fused feature map.

cloud inputs(Fig. 2). First, the transformation matrix between LiDAR and Radar is established through the extrinsics of the sensors and the Radar point clouds are projected into the LiDAR coordinate system. After aligning the multi-modal inputs, the LiDAR and Radar input point clouds are voxelized respectively and extracted into 2D feature maps or pseudo-images $F \in \mathbb{R}^{C\times H\times W}$ and enhance the attention of pseudo-images to specific objects by the channel attention module. Then concatenate the LiDAR and Radar channel-attentioned pseudo-images into a feature map $F^{'} \in \mathbb{R}^{2C\times H\times W}$ to obtain the weight of the feature map in space domain by a spatial attention module and reweight the channel-attentioned LiDAR and Radar feature maps with the weight. Finally fuse directly the two reweighted feature maps, and the fused feature map is obtained as input to the subsequent 2D-CNN and SSD modules to regress the 3D box.

A. Multi-modal Pointcloud Encoder

For multi-modal fusion, a crucial step is the alignment of multi-modal data, and the degree of data alignment has a significant impact on the final detection results. In this work, we draw inspiration from the methods employed in PointPainting [20], which project the Radar point cloud into LiDAR space with the extrinsics of multi-sensors. The raw Radar point cloud contains seven aspects of information, we filter and select five types of features: 3D location (x, y, z), reflectivity (RCS), and absolute radial velocity (v_{rc}) based on the official settings of VoD and the test results.

After aligning LiDAR and Radar data in the LiDAR coordinate system and filtering out unnecessary features from the raw Radar data, we convert these two modal point clouds to pseudo-images separately following PointPillars. The point cloud is firstly discretized into grids in the x-y plane and all points falling into the same grid are considered to constitute a pillar. Given the sparsity of the point cloud, the majority of the pillars are empty and non-empty pillars will in general have few points in them. Further processing involves filtering out non-empty P pillars, subsampling or supplementing the points in within each pillar to achieve a total of N points and obtain the corresponding index vectors. Following the encoder process, pillar features of dimensions $\mathbb{R}^{D\times P\times N}$ can be constructed from the original point cloud data, where D represents the feature dimension of the raw input (4 for LiDAR and 5 for Radar). Then utilize a simplified PointNet [15] to generate point-wise features $\mathbb{R}^{C\times P\times N}$ for the points in pillars and perform a max pooling layer to produce pillar-wise features of $\mathbb{R}^{C\times P}$. The sparse pseudo-image $F \in \mathbb{R}^{C\times H\times W}$ can be generated by scattering the dense features to the x-y plane based on the grid positions corresponding to each pillar.

B. Pillar Attention Fusion Module

After converting the point clouds outputted by LiDAR and Radar into pseudo-images that are aligned as much as possible, the key step lies in how to fuse them. The fusion process needs to preserve the detection capabilities of LiDAR while leveraging the advantages of Radar in detecting moving objects. In this work, we propose a Pillar Attention Fusion(PAF) module as shown in Fig.3, which incorporates attention mechanisms to enable the network to learn the fusion of LiDAR and Radar on its own. To enhance the representational capacity of features, the pseudo-images of LiDAR and Radar are individually subjected to channel-wise attention, focusing more on channels containing crucial information, retaining valuable features, and discarding less relevant ones. The separate process of channel attention is as follows:

$$F^{'} = M_c(F) \otimes F \tag{1}$$

in which $F^{'}$ represents the feature maps for each modality after applying channel attention. M_c is the channel attention module composed of two parallel max pooling and average pooling layers followed by an MLP with two hidden layers and finally a sigmoid activation to generate the channel attention map $M_c(F) \in \mathbb{R}^{C\times 1\times 1}$.

Subsequently, the pseudo-images enhanced with channel-wise attention from different modalities are concatenated in the channel domain, resulting in a single pseudo-image $F^{''} \in \mathbb{R}^{2C\times H\times W}$. With LiDAR and Radar modalities aligned, this pseudo-image effectively combines features from LiDAR and

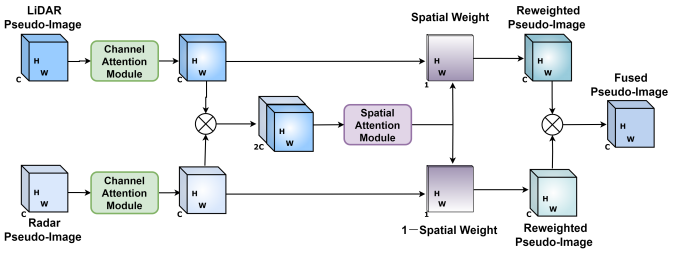


Fig. 3. Pillar Attention Fusion(PAF) Module. The channel attention module directs the model to focus more on specific feature channels that are more meaningful for detecting targets. Fusion through the spatial attention model allows the model to learn which modality's features to emphasize at different spatial locations.

Radar points at the same 3D positions. A spatial attention mechanism is then employed to enable the network to learn the importance of LiDAR and Radar features at each spatial position. After calculating the spatial attention module generates weights $W_{spatial}$ which represent the importance of each position, the pseudo-image before concatenation is reweighted and summed to obtain the final fused feature map $F^{''}$. The fusion process of spatial attention is as follows:

$$W_{spatial} = M_s(F^{'}_{LiDAR} \oplus F^{'}_{Radar}) \tag{2}$$

$$F^{''} = W_{spatial} \otimes F^{'}_{LiDAR} + (1 - W_{spatial}) \otimes F^{'}_{Radar} \tag{3}$$

where $F^{'}_{LiDAR}$ and $F^{'}_{Radar}$ represents the individual channel attention feature maps. M_s is the spacial attention modulereducing channel dimensionality through global max pooling and mean pooling seperately, and concatenating the results with applying a convolutional layer and a sigmoid activation to generate the spatial attention map $W_{spatial} \in \mathbb{R}^{1\times H\times W}$.

C. 3D Box Regression

Up to this step, we have achieved the multi-modal fusion of LiDAR and Radar. We choose to adopt the framework from PointPillars for the subsequent 3D bounding box regression and perform top-down subsampling on the fused feature map through 2D CNN. Each subsampled feature is combined through upsampling and concatenation, resulting in a concatenation of features from different strides. Then input the final global feature into the Single Shot Detector (SSD) [11], where 2D Intersection over Union (IoU) is utilized to match prior boxes with ground truth. Subsequently, perform regression for object height and elevation given 2D match.

IV. Experiments

A. Datasets and Metrics

In this study, we verify the effectiveness of the proposed model on the View-of-Delft (VoD) dataset [13]. The VoD dataset [13] is a comprehensive dataset that provides synchronized data of images, LiDAR point cloud, and 4D Radar point cloud. It consists of 8,600 scans and includes 3D bounding box annotations for over 26,000 pedestrians, 10,000 cyclists, and 26,000 cars. Notably, the VoD dataset offers three types of 4D Radar point cloud: single-scan, three-scan, and five-scan. For our evaluation, we utilize the original features extracted from the Radar point cloud, which consist of seven dimensions [13]. The feature vector is represented as:

$$[x, y, z, \text{RCS}, v_r, v_{rc}, \tau] \tag{4}$$

in which (x, y, z) denote the coordinates of the Radar points, RCS represents the Radar signal reflection-intensity, v_r is the radial Doppler velocity relative to the ego vehicle, v_{rc} is the absolute Doppler velocity, and τ indicates the time ID indicating which scan the point belongs to.

To preserve the originality of radar data, we utilize the single-scan Radar point cloud data. Our evaluation focuses on three distinct object categories: cars, pedestrians, and cyclists [13] and divides the detection difficulty into simple, moderate, and difficult according to the difference in object size and occlusion degree. We employ the commonly used 3D average precision (AP3D) values as the evaluation metric for each object category. Additionally, we calculate the mean 3D AP (mAP-3D) and mean bird's-eye view AP (mAP-BEV) values to provide a comprehensive assessment of our algorithm's performance [13]. Following the official settings of the VoD dataset, the IoU thresholds used for calculating the performance metrics are set to 0.5 for cars, 0.25 for pedestrians and cyclists [13].

B. Implementation details

We train on the official training set provided by the VoD dataset and validate and test on its official validation set. In the preprocessing stage of point clouds, both LiDAR and Radar discard points outside the set detection range. In the initialization stage of the model, the range of all point cloud inputs is limited to (0, 57.6), (-28.8, 28.8), (-3, 2), and the size of pixels is set to (0.16, 0.16, 5). Each sample can have the max pixels of 40000 in train, 16000 in val and there can be the max points of 10 within each pixel; The prior box sizes for bicycles, cyclists, and pedestrians are set to (0.6 0.8 1.73), (0.6 1.76 1.73), (1.6 3.9 1.56). In the training phase of the model, we adopt the AdamW optimizer and the OneCycleLR learning rate schedule with an initial learning rate of 2.5×10^{-4}, the max learning rate of 2.5×10^{-3} and the proportion of learning rate increase cycle is 0.4. The batch size and the maximal number of learning epochs are set to 8 and 100, respectively. All the experiments are conducted on a GTX 4090 GPU.

C. Quantitative Results

a) 3D Object Detection on the VoD dataset: According to the official settings of the VoD dataset, we also used PointPillars as the baseline to evaluate the 3D detection performance of LiDAR-only and Radar-only with the metrics in IV-A and compared them with the LiDAR&Radar fusion model presented in our study. As shown in Table I, our proposed PAF module utilizes multi-modal LiDAR&Radar point clouds for 3D object detection, achieving significantly better detection levels than any single modality at the same baseline: on the universal evaluation metric mAP-3D, our model is 6.83% higher than LiDAR-only and 22.35% higher than Radar-only, respectively.

TABLE I
QUANTITATIVE RESULTS ON THE VOD VALIDATION SET

Method	Modality		Car			Pedestrian			Cyclist			mAP-3D	mAP-BEV
	LiDAR	Radar	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Mod.	Mod.
PointPillars	✓		69.23	68.83	64.34	58.25	55.25	51.97	78.07	73.43	68.04	65.84	68.66
PointPillars		✓	32.90	40.57	33.91	41.90	38.00	34.59	77.68	72.40	65.28	50.32	57.07
Ours	✓	✓	**71.90**	**69.22**	**64.59**	**66.57**	**65.02**	**60.25**	**87.90**	**83.77**	**80.10**	**72.67**	**75.78**

b) Comparison with Other Methods on the VoD dataset: In addition to comparing with single-modal input data, we also compare our results with other algorithms [2, 19, 24, 25] on the VoD dataset. Due to slight variations in experimental settings, we use the overall mAP across three classes of objects at all difficulty levels for the comparison. From the results in Table II, it is evident that our model significantly outperforms existing algorithms on the VoD dataset. Of course, this outcome is largely attributed to the high detection capability of LiDAR in the fusion with Radar.

TABLE II
COMPARISON RESULTS WITH OTHER METHODS

Method	Modality			mAP-3D
	Radar	LiDAR	Image	
CenterPoint [24]	✓			45.42
SMIFormer [19]	✓			48.77
RCFusiont [25]	✓		✓	49.65
IA-SSD [2]		✓		62.82
Ours	✓	✓		**72.14**

c) Ablation Studies with PAF module: Ablation studies are conducted to validate the effectiveness of the proposed PAF module in fusing LiDAR and Radar, and the results are presented in Table III. Directly fusing LiDAR and Radar feature maps without using the PAF module yields the lowest detection accuracy. The introduction of channel attention alone improves detection accuracy by 2.21%, while spatial attention alone improves it by 7.48%. The use of the PAF module results in a final detection accuracy improvement of 8.27%. The above experimental results demonstrate that the PAF module we adopted makes the fusion of LiDAR and Radar more effective.

TABLE III
ABLATION STUDIES RESULTS

PAF Modules		mAP-3D
Channel Attention	Spatial Attention	Mod.
✗	✗	64.40
✓	✗	66.61
✗	✓	71.88
✓	✓	**72.67**

D. Qualitative Results

We provide qualitative results in Fig. 4. While our multi-modal training is jointly guided by LiDAR and Radar, the visualization of LiDAR&Radar fused detection results is conducted in the LiDAR space for a more intuitive emphasis on detection results. As shown in Fig. 4, large targets in LiDAR, which contain more points, exhibit better detection results. However, for small target classes with few points or targets heavily occluded, the detection performance is significantly compromised. In contrast, the number of Radar points contained within a target is not strongly correlated with the target's size and is more affected by the target's motion state, resulting in more points for moving targets due to its sensor characteristics. Therefore, LiDAR-based 3D detection may miss small targets, while Radar-based 3D detection may miss stationary objects. Our fused LiDAR and Radar 3D detection achieves good results for both of these targets.

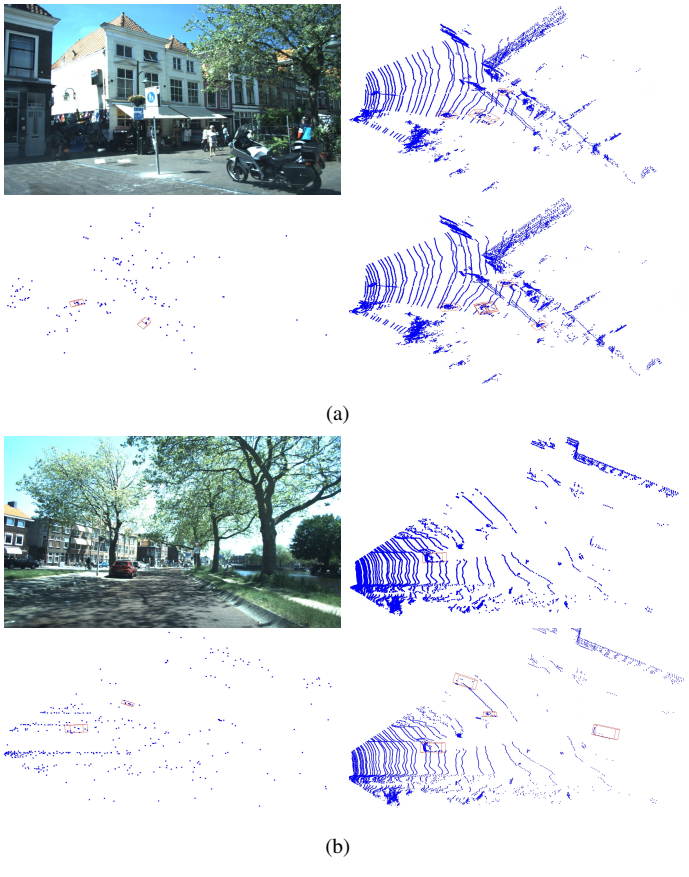


Fig. 4. Qualitative results on the VoD validation set. Two subfigures (a) and (b) are from different frames in the validation set of VoD. In each subfigure, the top-left shows the original camera image, the top-right shows the detection results from LiDAR-only, the bottom-left shows the results from Radar-only, and the bottom-right shows the results from our model fusing LiDAR&Radar.

V. Conclusion

This paper propose a novel fusion method combining LiDAR and 4D millimeter-wave Radar for 3D object detection demonstrates superior performance in terms of accuracy and robustness across different environmental conditions. The efficient multi-modal feature extraction and attention-based fusion enable seamless integration and comprehensive understanding of the environment. This approach holds significant potential for enhancing the capabilities of mobile robots in real-world scenarios, providing a reliable and effective solution for 3D object detection.

References

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

[2] King Wah Gabriel Chan. A study of attention-free and attentional methods for lidar and 4d radar object detection in self-driving applications. 2023.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[4] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 27(3):1347–1360, 2017.

[5] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8287–8294, 2019.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020.

[8] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. *IROS*, 2018.

[9] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[12] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13708–13718, 2021.

[13] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022.

[14] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[15] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[18] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[19] Weigang Shi, Ziming Zhu, Kezhi Zhang, Huanlei Chen, Zhuoping Yu, and Yu Zhu. Smiformer: Learning spatial feature representation for 3d object detection from 4d imaging radar via multi-view interactive transformers. *Sensors*, 23(23):9429, 2023.

[20] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[21] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[22] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1742–1749. IEEE, 2019.

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[24] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[25] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rcfusion: Fusing 4d radar and camera with bird's-eye view features for 3d object detection. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[26] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6688–6697, 2019.