Low-level Sensor Fusion for 3D Vehicle Detection using Radar Range-Azimuth Heatmap and Monocular Image

Jinhyeong Kim*†, Youngseok Kim*, and Dongsuk Kum

Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{1994kjhg, youngseok.kim, dskum}@kaist.ac.kr

Abstract. Robust and accurate object detection on roads with various objects is essential for automated driving. The radar has been employed in commercial advanced driver assistance systems (ADAS) for a decade due to its low-cost and high-reliability advantages. However, the radar has been used only in limited driving conditions such as highways to detect a few forwarding vehicles because of the limited performance of radar due to low resolution or poor classification. We propose a learning-based detection network using radar range-azimuth heatmap and monocular image in order to fully exploit the radar in complex road environments. We show that radar-image fusion can overcome the inherent weakness of the radar by leveraging camera information. Our proposed network has a two-stage architecture that combines radar and image feature representations rather than fusing each sensor's prediction results to improve detection performance over a single sensor. To demonstrate the effectiveness of the proposed method, we collected radar, camera, and LiDAR data in various driving environments in terms of vehicle speed, lighting conditions, and traffic volume. Experimental results show that the proposed fusion method outperforms the radar-only and the image-only method.

1 Introduction

A frequency-modulated continuous-wave (FMCW) radar and RGB camera have been widely used in advanced driver assistant systems (ADAS) thanks to their many advantages for mass production. Commercial radars and cameras have advantages of low-maintenance, high-reliability, and low-cost due to their stable design and mature market. Despite the many advantages of radar, the radar used in ADAS is limited to detecting a few forwarding vehicles as the radar data is processed using traditional signal processing algorithms. Learning-based methods are expected to show better performance when replacing existing rule-based algorithms. However, 3D object detection utilizing low-level radar data in deep learning frameworks has not yet been thoroughly investigated.

*Contributed equally to this work.
†This work was done when Jinhyeong Kim was at KAIST, prior to joining SOCAR.

The automotive FMCW radar can measure distances to distant objects and can operate robustly even in harsh weather conditions due to the nature of fundamental design and long wavelength. However, the long wavelength of radar also restricts its performance. The radar suffers from a low angular resolution and accuracy that makes it challenging to separate adjacent vehicles. Contrarily, the camera has a high angular resolution due to the dense pixels and dense RGB pixels can provide visual cues to classify the category of objects. As shown in Table 1, the camera and radar have very complementary properties. Therefore, the camera-radar sensor fusion is promising to complement the shortcomings of each sensor and improve the detection performance.
This paper aims to detect 3D vehicles by sensor fusion network using the radar range-azimuth heatmap and image data, as illustrated in Fig. 1. To demonstrate the effectiveness of the proposed fusion method on the various driving environment, we constructed a dataset because none of the public datasets contains the low-level radar with 3D annotations.

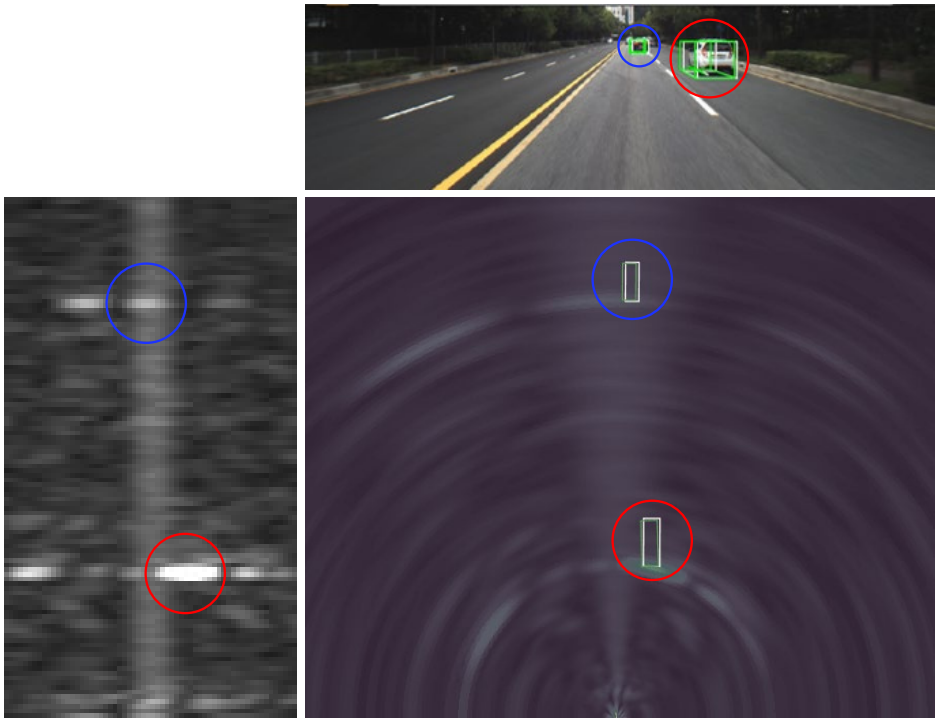

Fig. 1. Detection results of the proposed fusion method on the camera image (top), radar range-azimuth heatmap in the polar coordinate system (left), and radar in the Cartesian coordinate system (right). White and green boxes refer to ground truths and prediction results.

Table 1. Characteristics of sensors widely used in vehicle intelligence.

	Classifi -cation	*Radial Accuracy*	*Angular Accuracy*	*Weather Condition*	*Lighting Condition*	*Measuring Range*	*Cost*	*Mainte -nance*	*Reliability*
Camera	◯	△	◯	×	×	×	◯	◯	◯
Radar	△	◯	△	◯	◯	◯	◯	◯	◯
LiDAR	△	◯	◯	△	△	△	×	×	×

2 Background

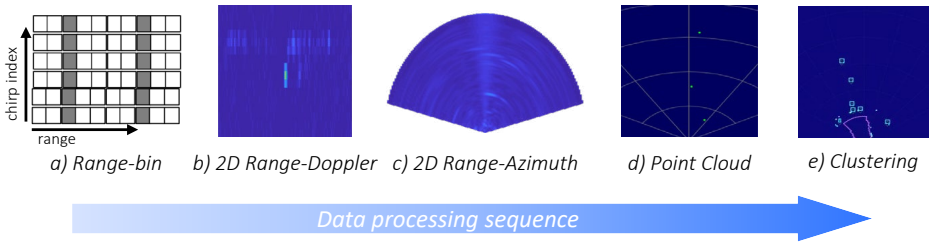

Fig. 2. Data processing sequence of FMCW radar. The proposed method uses 2D Range-Azimuth heatmap representation (c) instead of point cloud (d) or object-level representation (e).

We summarize the basic principle and data processing process of the FMCW radar in Fig. 2. The FMCW radar transmits a chirp signal that linearly increases frequency and receives the reflected signal. The frequency difference between transmitted and received signal obtained by Analog-to-Digital Converter (ADC) is calculated by Fast Fourier Transform (FFT) to calculate the distance (Fig. 2-a). The velocity and angle are estimated by measuring the phase difference across signals. The velocity can be calculated by two chirps measured at a successive time (Fig. 2-b), and the angle is calculated by the same chirp measured from multiple RX antennas. As a result, a 3D radar tensor with a range-azimuth-Doppler dimension is obtained as a result of FFTs (Fig. 2-c). In this paper, this radar representation is referred to as a radar heatmap. After that, conventional signal processing techniques such as Constant False-Alarm Rate (CFAR) [1] process low-level data to identify valid point targets among clutters (Fig. 2-d). Finally, objects are detected by a clustering algorithm and verified using filtering and tracking algorithms (Fig. 2-e). Conventional signal processing algorithms using hand-coded features (e.g., CFAR, MUSIC) works

robustly in simple scenarios, but their performance drops significantly in complex urban driving environments with many metal objects such as streetlights. To be able to use radar in complex environments, we design a learning-based method to fully exploit information on image-like radar range-azimuth heatmap representation (Fig. 2-c) rather than object-level representation (Fig. 2-e).

3 Related Work

Dataset for Autonomous Driving. A number of public datasets for autonomous driving have recently been published. KITTI [2] provides a monocular and stereo camera, 3D LiDAR for many computer vision tasks such as 3D object detection, tracking, and depth prediction. However, it is pointed out that the diversity of the dataset may not be sufficient because data is only collected during the daytime and on sunny days. Apolloscape [3] collected a total of 143,969 frames, which contains the largest labels among public datasets and it claims to have a higher diversity compared to KITTI. However, KITTI and Apolloscape do not provide radar data. NuScenes [4] is a multimodal dataset for 3D object detection and tracking tasks and contains radar data, but radar data is processed as a point cloud representation. We argue that a lot of valuable information can be lost during the signal processing. Oxford RobotCar [5] provides camera, 3D LiDAR, and radar data as range-azimuth heatmap representation, which is the same representation used in this paper. However, RobotCar [5] does not provide 3D object labels because the dataset is aimed at the odometry task. The synthetic datasets such as Virtual KITTI [6] are used as alternatives to address the data limitation issue. However, it is known to be challenging to generate synthetic radar data since the radar beam is difficult to simulate due to the nature of electromagnetic waves.

Learning-based Object Detection using Low-level Radar data. Only a few studies have been conducted using low-level radar data for object detection. We assume this is because of the absence of the public dataset containing low-level radar data and the ground truth label. He et al. [7] and Kwon et al. [8] use a time-serial micro-Doppler map to classify human activities using a convolutional neural network (CNN) and multi-layer perceptron (MLP), but their works do not consider detecting the position of the object. Brodeski et al. [9] and Zhang et al. [10] utilize U-Net [11] like architecture to detect objects on the range-Doppler map. These studies, however, are demonstrated in restricted environments such as a chamber and a vacant lot. Major et al. [12] collect radar range-azimuth-Doppler data on highway driving scenario and detect vehicles on a bird's eye view (BEV). They employ a one-stage detection network SSD [13] and it provides good detection performance in highway environment, but it is not be guaranteed to work well in complex urban situation.

Sensor Fusion-based Object Detection. The number 3D object detection studies have been conducted using multiple sensors, mainly LiDAR and camera.

MV3D [14] generates 3D proposals from BEV LiDAR feature map and projects proposals into a front view LiDAR and image feature map to fuse a projected region of interest (RoI). Similarly, AVOD [15] projects 3D anchors to LiDAR and image feature maps, respectively. RoIs from different feature maps are fused on region proposal network (RPN) stage and generate high-recall object proposals. Few methods exploit radar and camera sensors but not fully investigated on detecting 3D objects. Chadwick et al. [16] focus on detecting distant vehicles by using object-level radar data. It projects radar data into the image plane and detects a 2D bounding box using SSD [13] in the image pixel coordinate system rather than vehicle coordinate system. Meyer and Kuschk [17] exploit radar point cloud and camera to detect 3D vehicle using AVOD [15] architecture. Lim et al. [18] utilize low-level radar and camera. Images are projected into the BEV plane using Inverse Perspective Mapping (IPM) to match the coordinate system with radar, on the dataset collected in [12]. However, their experiments are conducted in the highway driving scenario and assume a planar road scene to use IPM, but IPM approach is difficult to be adopted to the road with slope.

4 Dataset

4.1 Sensor Configuration

We use the Hyundai Ioniq vehicle platform equipped with a camera, radar, and LiDAR to collect data. Sensor specifications and placements are described in Table 2 and Fig. 3. We mount radar and LiDAR on the front bumper parallel to the ground, while the camera is mounted on the top of the vehicle.

Table 2. Sensor specification.

Sensor	Specification
Camera	**1×FLIR Blackfly**, 10Hz capture frequency, RGB, 1/1.8" CMOS, 1920×704 resolution, auto exposure, JPEG compressed
Radar	**1×TI AWR1642**, 10Hz capture frequency, 77 to 81-GHz FMCW, 4Rx and 2Tx antennas, 120° horizontal FoV, ≤55m range
LiDAR	**3×IBEO LUX**, 25Hz capture frequency, 4 beams, 85° horizontal FoV, ≤80m range

We carefully calibrated intrinsic parameters and extrinsic parameters to obtain a reliable ground truth and to transform the coordinate system between sensors. First, we calibrated the camera to be undistorted and rectified by intrinsic parameters, then calibrated extrinsic parameters between camera and LiDAR using the approach proposed in [19]. After that, 6-DOF rigid transformation parameters between LiDAR and radar are obtained using a laser scanner while keeping the two sensors are mounted in parallel.

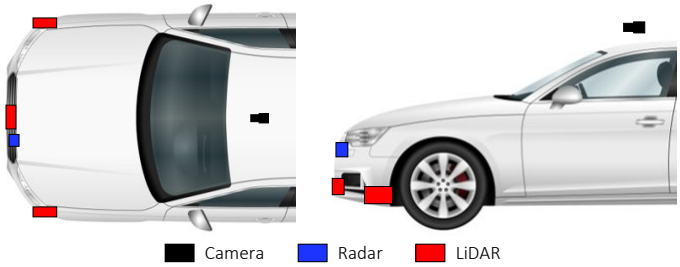

Fig. 3. Sensor placements of vehicle platform.

For the radar, built-in subsystems such as analog-to-digital converter (ADC), digital signal processing (DSP), and ARM-based processer are integrated with the radar sensor and process the signal as Fig. 2. In this paper, we use low-level range-azimuth data instead of a point cloud or object-level data to fully exploit the potential of radar. We modify the C++ implemented radar firmware of chip to access 2D range-azimuth data from the radar.
In order to reduce the data misalignment between multiple sensors, we synchronize data using CPU time. As a result, radar and LiDAR data captured closest to the camera are used, and data is sampled at 2Hz.

4.2 Data Acquisition and Annotation

Driving data has been recorded while driving around campus, urban areas, and motorways in Daejeon, Korea. After recording raw data, we select interesting scenes considering the diversity of data with respect to the speed of ego vehicle, the volume of traffic, and lighting conditions. 'Stop' means that the ego vehicle slows down and stopped while surrounding vehicles are moving (e.g., stopping at intersection or red traffic light), 'Low' means that the ego vehicle drives below 40kph (e.g., campus), and 'Normal' is a general road driving environment that drives above 40kph. The scenes according to the light condition consist of 'Sunny,' 'Cloudy,' and 'Night.'
This paper focuses on demonstrating the effectiveness of the proposed method on detecting car class. We carefully annotated the 3D position, size, and orientation of the car. The ground truths label are annotated using LiDAR point clouds and transformed into the radar coordinate system. Note that we annotated vehicles with more than half of the vehicle is inside the image frame, and vehicles located within 50 meters.

4.3 Dataset Analysis

We carefully split collected data into a train set and test set while making sure that training and testing data does not come from the same scene. We analyze the distribution of the dataset in terms of the ego vehicle speed, lighting conditions, distance to annotation, and the number of annotation in frame in Table 3

and Fig. 4. The distribution analysis shows that our dataset can represent the complex urban driving scenarios.

Table 3. Statistics of collected dataset.

Split	Size (Hr)	Number of frames	Number of annotations
Training set	0.78	5512	9115
Test set	0.32	2232	5853
Total	1.10	7744	14968

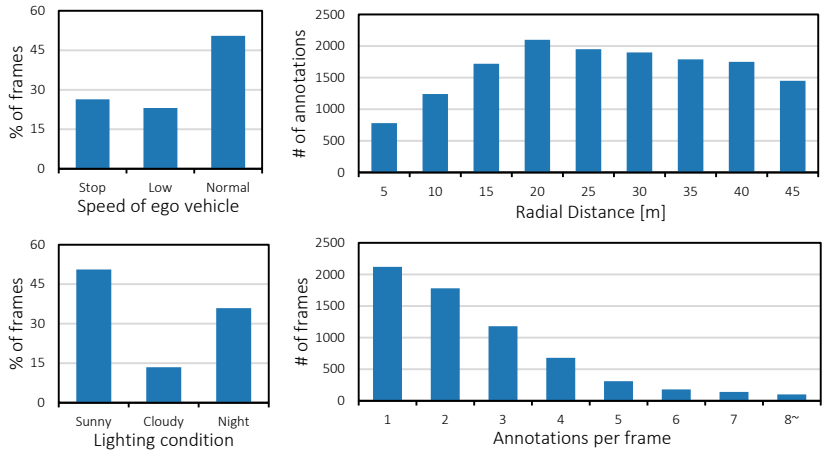

Fig. 4. Distributions of collected dataset consists of ratio of driving conditions, distance of annotation from the ego vehicle, and the number of annotations in each frame.

5 Methodology

The performance of the sensor fusion-based object detection network can vary depending on the stage in which the two modalities are combined. Sensor fusion methods can be broadly categorized into early, middle, and late fusion, and each method has its strengths and weaknesses. The conventional sensor fusion method for ADAS is the late fusion method that combines object-level outputs processed by each modality, which has advantages of high flexibility and modularity. However, it discards the benefit of rich information of intermediate features. While, the early fusion method combines the two raw sensor data and feeds it into the network. It can utilize the information of the raw data, however, two modalities

have to be on the same coordinate system so that two modalities can be aligned. Meanwhile, middle fusion is a compromise between early and late fusion because it combines feature representations from two modalities at intermediate layers. It can fully exploit both input data by using appropriate feature extractor for each modality, and network design is advantageous because the coordinate systems of the two sensors do not have to be the same.
As illustrated in Fig. 5, our proposed method has a middle fusion method that combines the region of interest (RoI) features from two modalities based on a two-stage object detection architecture with region proposal network (RPN) and detection head following AVOD [15] architecture.

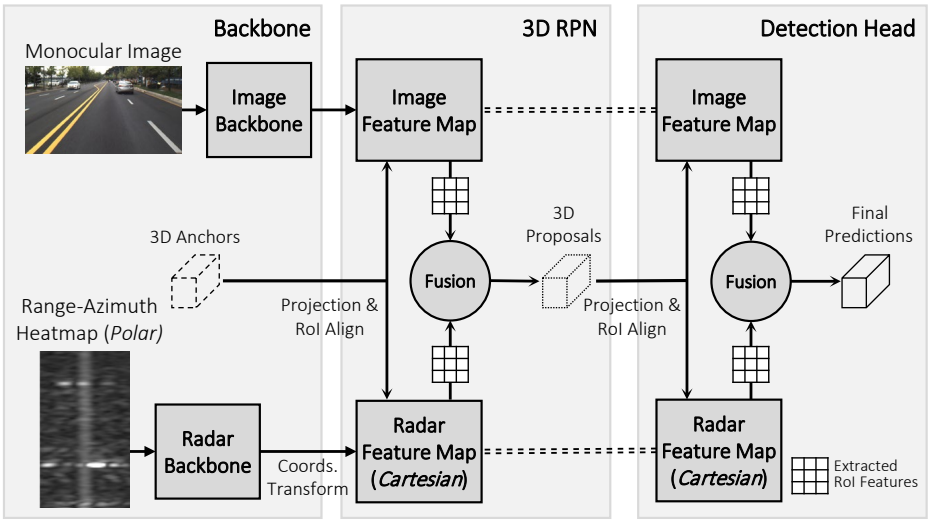

Fig. 5. Overall architecture of the proposed radar-image fusion method.

5.1 Radar Representation and Backbone

The 2D range-azimuth heatmap is naturally obtained in range-azimuth (polar) coordinate system, but the polar coordinate system has several shortcomings to detect objects on 3D space. In the polar coordinate system, the physical distance between two adjacent data in the polar space increases as the radial distance increases. As also claimed in [12], detecting objects in polar feature map has inferior performance than detecting objects in Cartesian space. Following [12], we extract features in polar space then explicitly transforms the extracted feature map into Cartesian space using bilinear interpolation.
The radar and image backbone has a modified VGG16 [20] model and a feature pyramid network (FPN) [21]. The width and height of input decrease by half using max pooling, while the number of feature increases from 32 to 256 every 2, 2, 3, and 3 layers. The network takes the image with a size of 1920×704 and

range-azimuth radar in polar space with a size of 128×64. The radar input has a resolution of 0.4392 m and 1.9 degree. The last radar feature map is transformed to Cartesian space with a size of 560×610 with 0.1 m resolution.

5.2 3D Region Proposal Network

In the 3D region proposal network (RPN) stage, a 3D anchor is used to generate proposals. Similar to 2D RPN, 3D RPN generates anchors in 3D space within a range of radar Cartesian space. Two Region of Interests (RoIs) are obtained by projecting given 3D anchor into two feature maps of each modality, and RoIs are cropped into 7×7 size feature by RoI Align [22]. Here, RoI Align is adopted to minimize the quantization effect occurring near the RoI boundary. Minimizing the quantization effect is especially important for the radar because the resolution of radar is low and a single pixel misalignment can lead to large localization error. Two extracted RoIs are fused by concatenation operation and fed to two branches of fully connected layers with two layers of size 256. One branch performs 3D box regression and the other outputs the objectness score. Smooth $L1$ loss is used for the 3D box regression and focal-loss [23] for objectness.
For training RPN, assigning a positive and negative (foreground and background) label to each anchor is required to compute the loss as introduced in Faster R-CNN [24]. Intersection-over-Union (IoU) is widely used for the matching metric in the object detection task, defined as:

$$IoU = |B \cap B^{gt}|/|B \cup B^{gt}| \tag{1}$$

where $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ and $B = (x, y, w, h)$ is the center position and size of ground truth and prediction results, respectively. However, we claim that IoU is not the best choice for the vehicle and radar application. Since the vehicle in BEV space has a long rectangular shape, small lateral displacement can greatly reduce the IoU, as an example shown in Fig. 6. The IoU can be especially fatal to the radar because the radar has low lateral accuracy. We claim that the IoU threshold is inconsistent and it can inhibit the network generalization. Therefore, we proposed the distance matching metric using L2 norm between center positions of ground truth and prediction as follows:

$$Distance\ matching = \|(x^{gt}, y^{gt}) - (x, y)\|_2 \tag{2}$$

In the RPN stage, anchor closer than 2 m are regarded as positive, and farther than 2.5 m are regarded as a background during training schemes. Predicted proposals are filtered by 2D non-maximum suppression (NMS) at an IoU threshold of 0.8 in BEV space.

5.3 Detection Head

In the detection head stage, the top 100 proposals from the RPN are projected onto each coordinate system again, and the RoI fusion proceeds as same in

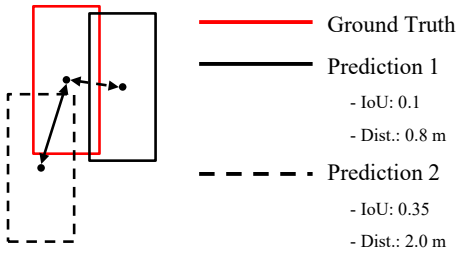

Fig. 6. Example of the IoU matching metric and the distance matching metric. The prediction 1 is closer to the ground truth than prediction 2 but it has a lower score when using IoU due to the shape of the vehicle.

the RPN. The extracted features are cropped and resized to 7×7 and fused through a concatenation operation used as an input to fully connected layers. The final detection head consists of one branch of a fully connected layer, which consists of three layers of size 2048 to output object probability, box regression, and orientation. Similar to RPN, proposals closer than 1 m are considered as positive, and farther than 1.25 m are negative. NMS threshold of 0.001 IoU is used to remove prediction results that are overlapped with each other.

5.4 Implementation Details

We apply multi-task loss for position and size regression, orientation, and classification in an end-to-end fashion same as [15].

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} + \lambda_{dir} L_{dir} \tag{3}$$

In (3), the regression and orientation terms use smooth $L1$ loss and weights are experimentally set to $\lambda_{cls} = 5$, $\lambda_{reg} = 3$, and $\lambda_{dir} = 5$. The classification loss for RPN has the focal loss [17] following:

$$FL\left(p_t\right) = -\alpha_t \left(1 - p_t\right)^{\gamma} \log(p_t) \tag{4}$$

We use $\alpha = 0.3$ and $\gamma = 2.25$ to enforce positive and negative samples to have a 1:3 ratio. The proposed network was trained using Adam optimizer, with an initial learning rate of 0.0001 with a decay of 0.8 for every 33k iterations until 220k iterations.

6 Experiments

We evaluate the performance of the proposed method using three different metrics and compare with a radar-only and an image-only method. The radar baseline method has the same architecture as the proposed two-stage method without

the image branch. For the image baseline, the state-of-the-art image-based 3D detection method M3D-RPN [25] is trained and evaluated on our dataset.
We use the average precision (AP) metric using bird's-eye-view (BEV) IoU threshold of 0.5 and 0.7 used in KITTI [2] and center distance threshold of 0.5, 1.0, and 2.0 meters used in NuScenes [4]. We also evaluate the localization performance using root-mean-square error (RMSE).

6.1 Quantitative Evaluation

Table 4 and 5 analyze AP performances using IoU and distance metrics. Results show that the image method has a better performance than the radar baseline method in most metrics and thresholds. We hypothesize that it is difficult for the radar alone to classify vehicles from metal obstacles due to the lack of contextual information, leading to many false positives, and resulting in poor precision. Moreover, radar alone is hard to separate between two adjacent objects due to the low angular resolution, and it leads to true negatives and lower recall. However, adding the image to the radar can boost the performance on all evaluation metrics. This verifies that fusing two modalities can complement each other and yield higher performance over the single modality.

Table 4. Average Precision (AP) using IoU matching.

Method	Modality	$AP_{BEV,IoU}$	
		IoU=0.5	IoU=0.7
M3D-RPN	Image	39.46	11.71
Radar baseline	Radar	26.88	12.91
Proposed	Radar+Img	**46.16**	**16.30**

Table 5. Average Precision (AP) using distance matching.

Method	Modality	$AP_{BEV,dist}$		
		0.5 m	1.0 m	2.0 m
M3D-RPN	Image	16.31	39.37	64.71
Radar baseline	Radar	15.36	32.98	44.34
Proposed	Radar+Img	**26.92**	**51.06**	**66.26**

The RMSE in a longitudinal and lateral direction is shown in Table 6. Note that only results detected in all three methods using 2.0 m distance threshold are used

to calculate the RMSE for the fair comparison. The radar baseline method has a low longitudinal error, and the image method has a low lateral error, which is reasonable given the characteristics of each sensor. As can be seen, fusing camera and radar sensors together contributes to reduce the localization errors in longitudinal and lateral directions, thus improve overall performance.

Table 6. Root-mean-square error (RMSE) on prediction results using 2 m distance threshold.

Method	Modality	RMSE [m]	
		Longitudinal	Lateral
M3D-RPN	Image	0.2486	**0.1529**
Radar baseline	Radar	0.2219	0.2080
Proposed	Radar+Img	**0.2210**	0.1828

6.2 Qualitative Results

We visualize qualitative results of radar, image, and fusion method in Fig. 7. Note that all figures are best viewed in color with zoom in. We observe that the radar baseline method often suffers from false positives and separating adjacent vehicles, and the image method typically fails to detect distant vehicles. The proposed method is able to detect and classify vehicles accurately.

Fig. 7. Qualitative comparison on test set using radar-only (top), image-only (middle), and proposed fusion method (bottom). The 3D bounding box is projected into the image space and BEV space for the visualization. White and red box denotes the ground truth and green and blue box denotes the prediction results.

Fig. 8 shows the advantage of fusion method compared to the radar baseline. As highlighted, radar alone suffers from detecting clutter signals as object (blue circle, false positive) and fails to separate adjacent vehicles (red circle, true negative). The proposed fusion method overcomes weaknesses of radar alone method by utilizing visual cues.

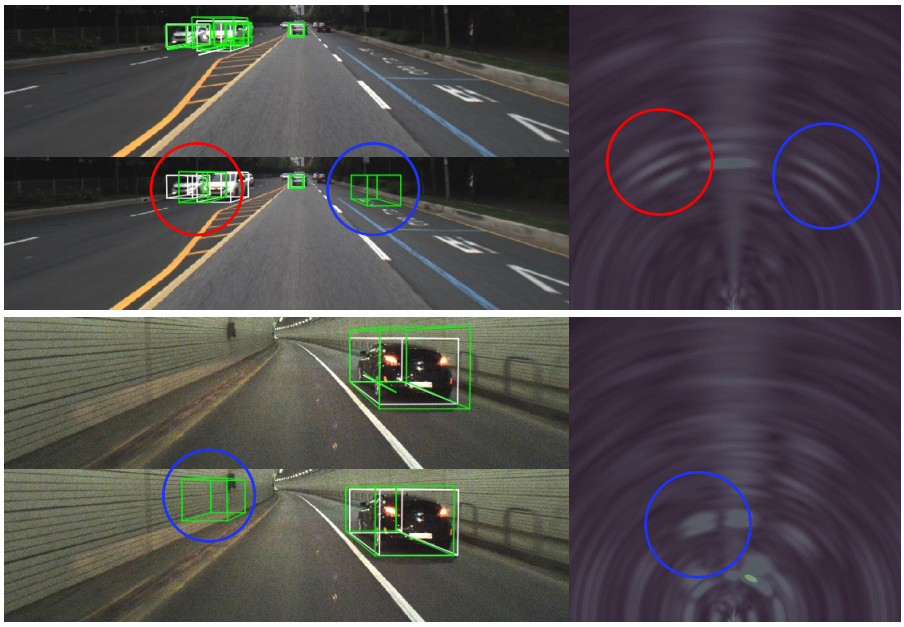

Fig. 8. Qualitative results in challenging scenarios. Predictions results using proposed fusion method (top), radar-only (bottom), and radar range-azimuth heatmap input (right).

6.3 Ablation Study

As we hypothesized in Section 5.2, the IoU matching is not suitable for the vehicle and radar application due to the shape of the vehicle and the characteristic of radar. To verify the benefit of the distance matching, we compare the proposed network with the network trained using the IoU metric. For the RPN stage, anchors with IoU less than 0.25 are considered as negative, while IoU greater than 0.3 are considered as positive. For the detection head stage, proposals with IoU less and greater than 0.35 and 0.4 are considered as negative and positive. For better understanding, we note that 0.4 IoU and 1.0 m distance thresholds are similar. Both networks are trained in the same manner as explained in Section 5.4.

As shown in Table 7 and Fig. 9, the number of positive samples on the RPN stage using distance matching is larger than the IoU matching even two networks use

a similar matching threshold. More positive samples during training can help to converge faster and lead to better performance. As a result, the distance matching shows better performance on both IoU and the distance evaluation metric by 4.44% and 4.53%.

Table 7. Comparison between IoU and distance matching method.

Training metric	# of positive anchors on RPN	Evaluation metric	AP [%]
IoU	4.3	IoU (0.5) Dist. (1.0 m)	41.72 46.53
Distance	20.4	IoU (0.5) Dist. (1.0 m)	**46.16** **51.06**

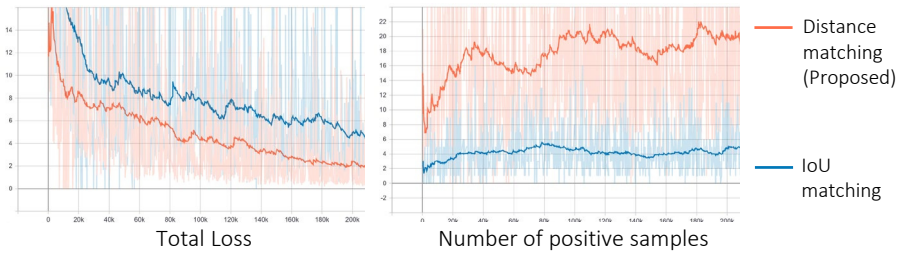

Fig. 9. Visualization of total loss and the number of samples by iteration.

7 Conclusion

In this paper, we introduced the sensor fusion-based 3D object detection method using radar range-azimuth heatmap and monocular image. We demonstrated the proposed low-level sensor fusion network on the collected dataset and showed the benefit of the proposed fusion method over radar alone method. In addition, we showed that the proposed distance matching method helps the network train stable and yields better performance compared to the IoU method in radar application. The proposed method has shown the potential to achieve high performance even with inexpensive radar and camera sensors.

8 Acknowledgement

This research was supported by the Technology Innovation Program (No. 10083646) funded By the Ministry of Trade, Industry & Energy, Korea and the KAIST-KU Joint Research Center, KAIST, Korea.

References

1. Rohling, H.: Radar cfar thresholding in clutter and multiple target situations. IEEE Transactions on Aerospace and Electronic Systems (1983)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR. (2012)
3. Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., Yang, R.: The apolloscape dataset for autonomous driving. In: CVPR Workshop. (2018)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. (2020)
5. Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I.: The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In: ICRA. (2020)
6. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtualworlds as proxy for multi-object tracking analysis. In: CVPR. (2016)
7. He, Y., Yang, Y., Lang, Y., Huang, D., Jing, X., Hou, C.: Deep learning based human activity classification in radar micro-doppler image. In: EuRAD. (2018)
8. Jihoon, K., Seungeui, L., Nojun, K.: Human detection by deep neural networks recognizing micro-doppler signals of radar. In: EuRAD. (2018)
9. Brodeski, D., Bilik, I., Giryes, R.: Deep radar detector. In: RadarConf. (2019)
10. Zhang, G., Li, H., Wenger, F.: Object detection and 3d estimation via an fmcw radar using a fully convolutional network. In: ICASSP. (2020)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. (2015)
12. Major, B., Fontijne, D., Sukhavasi, R.T., Hamilton, M.: Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In: ICCV Workshop. (2019)
13. Liu, W., Anguelov, D., Dumitru, E., Christian, S., Scott, R., Fu, C.Y., C. Berg, A.: Ssd: Single shot multibox detector. In: ECCV. (2016)
14. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: CVPR. (2017)
15. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.: Joint 3d proposal generation and object detection from view aggregation. In: IROS. (2018)
16. Chadwick, S., Maddern, W., Newman, P.: Distant vehicle detection using radar and vision. In: ICRA. (2019)
17. Meyer, M., Kuschk, G.: Deep learning based 3d object detection for automotive radar and camera. In: EuRAD. (2019)
18. Lim, T.y., Major, B., Fontijne, D., Hamilton, M.: Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In: NeurIPS Workshop. (2019)
19. Huang, J.K., Grizzle, J.W.: Improvements to target-based 3d lidar to camera calibration. IEEE Access (2020)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
21. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. (2017)
22. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV. (2017)
23. Lin, T.y., Girshick, R., Doll, P., He, K., Doll´ar, P.: Focal loss for dense object detection. In: ICCV. (2017)

24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. (2015)
25. Brazil, G., Liu, X.: M3d-rpn : Monocular 3d region proposal network for object detection. In: ICCV. (2019)