2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

GRIF Net: Gated Region of Interest Fusion Network for Robust 3D Object Detection from Radar Point Cloud and Monocular Image

Youngseok Kim[1], Jun Won Choi[2], and Dongsuk Kum[1*]

***Abstract*— Robust and accurate scene representation is essential for advanced driver assistance systems (ADAS) such as automated driving. The radar and camera are two widely used sensors for commercial vehicles due to their low-cost, high-reliability, and low-maintenance. Despite their strengths, radar and camera have very limited performance when used individually. In this paper, we propose a low-level sensor fusion 3D object detector that combines two Region of Interest (RoI) from radar and camera feature maps by a Gated RoI Fusion (GRIF) to perform robust vehicle detection. To take advantage of sensors and utilize a sparse radar point cloud, we design a GRIF that employs the explicit gating mechanism to adaptively select the appropriate data when one of the sensors is abnormal. Our experimental evaluations on nuScenes show that our fusion method GRIF not only has significant performance improvement over single radar and image method but achieves comparable performance to the LiDAR detection method. We also observe that the proposed GRIF achieve higher recall than mean or concatenation fusion operation when points are sparse.**

I. Introduction

The intelligent vehicle technology, such as advanced driver assistant systems (ADAS), plays an important role in the safety of the driver. Accurate and robust 3D object detection on roads with various traffic participants is essential for intelligent vehicles. The LiDAR is a popular choice for highly automated vehicles (e.g., Level 4 and 5) owing to its high accuracy, but LiDAR is not suitable for mass-production vehicles yet due to its high-cost, high-maintenance, and low-reliability. The radar and camera are the only available sensors that suit for commercial vehicles.

In spite of the fact that radar and camera have advantages in mass-production and been used for ADAS over a decade, each sensor has clear advantages and disadvantages as Table I. The camera provides RGB pixels with a dense angular resolution and rich visual cues that can distinguish between types of objects, but does not provide the range information and easily affected by weather and lighting conditions. The camera-based distance measuring relies on a geometrical relationship using camera calibration [1], but the performance can be degraded by pitch motion or the slope of the road. Meanwhile, the radar operates robustly in harsh weather conditions and measures the range to the long distance accurately. However, the angular resolution and accuracy of the radar are low due to the nature of wavelengths and its operating mechanism. The radar-based object detection using traditional signal processing methods is applied to commercial ADAS in highway environments, but cannot be guaranteed to work properly in complex environments. The sensor fusion is required to complement properties of radar and camera and improve the performance.

	Classifi-cation	Radial Accuracy	Angular Accuracy	Weather Condition	Lighting Condition	Detection Range	Cost	Mainte-nance	Reliability
Camera	○	△	○	×	×	×	○	○	○
Radar	△	○	△	○	○	○	○	○	○
LiDAR	△	○	○	△	△	△	×	×	×

○: Good, △: Normal, ×: Bad

TABLE I: Characteristics of sensors used in automotive

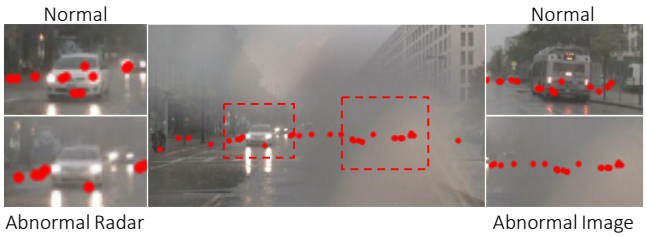


Fig. 1: Example of partially abnormal data in challenging nuScenes dataset. *Left:* missing radar point on the object due to low resolution. *Right*: blurred image by a raindrop. Red dots represent radar points projected on the image.

Several radar-camera sensor fusion studies have been conducted to complement the limitations of a single sensor. However, existing researches are mainly focused on utilizing object-level detection results of each sensor to remove false positives through cross-validation [2] or reduce computation cost by reducing the region of interest (RoI) [3]. Such an object-level late fusion scheme is difficult to expect high performance gain because the typical signal processing step loses out the amount of information from the low-level data and cannot overcome the drawbacks of each sensor. In order to fully utilize the advantages and complement disadvantages of each sensor, low-level early fusion is necessary.

The learning-based low-level radar and camera fusion have not yet been thoroughly investigated since there was no dataset containing low-level radar data before the nuScenes [4] was released. Meanwhile, a number of LiDAR and camera fusion studies have been conducted on the KITTI [5], mainly propose architectures for using different types of sensors. These architectures combine feature representations from different sensors, but there is not enough consideration on *how to combine* them. Most previous works assume that both sensor data are useful and use mean or concatenation operation to combine them. However, unlike the LiDAR

This research was supported by the Technology Innovation Program (No. 10083646) funded By the Ministry of Trade, Industry & Energy, Korea.

* corresponding author
[1]Youngseok Kim and Dongsuk Kum are with the Graduate School for Green Transportation, KAIST, Daejeon, Republic of Korea. {youngseok.kim, dskum}@kaist.ac.kr
[2]Jun Won Choi is with the Dept. of Electrical Engineering, Hanyang University, Seoul, Republic of Korea. junwchoi@hanyang.ac.kr

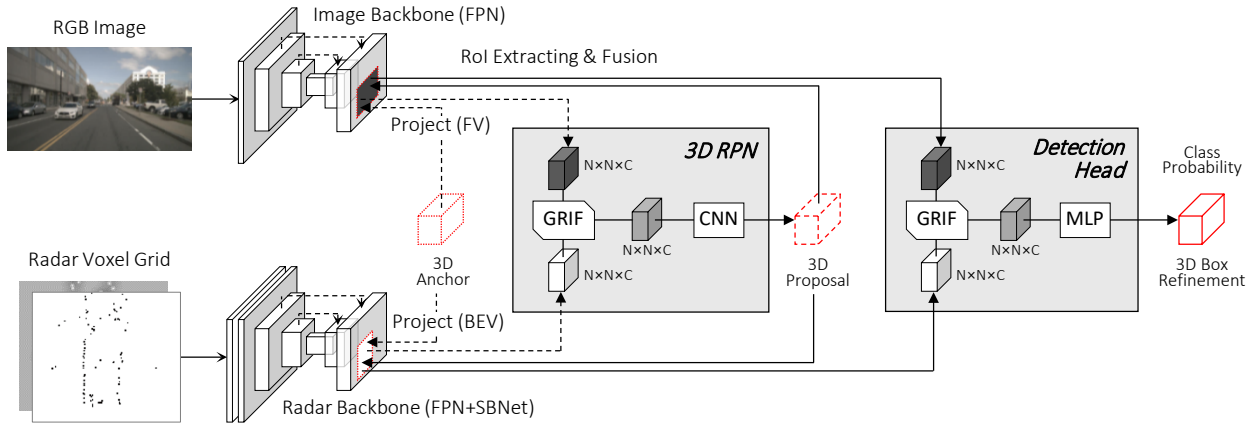


Fig. 2: The overall architecture of proposed GRIF Net. 3D anchors and 3D proposals in 3D space are projected into each modalities' feature map and cropped into regular size RoI. GRIF adaptively combines RoI features to robustly predict the class probability and 3D box. See §III for more details.

provides uniform, dense, and accurate points, the radar points are often missing on objects as illustrated in Fig. 1. In other cases, the image may be blurred while the radar points are reflected on the object normally. In these situations, the method of combining feature representations can significantly affect the performance since abnormal data cannot contribute to improving the performance by data fusion.

In this paper, we focus on fusion-based 3D vehicle detection in challenging environments, taking into account the characteristics of radar and camera, which are only suitable for mass-production. We summarize the challenges of using radar point cloud and our contributions as follows:

- We propose the Gated Region of Interest Fusion (GRIF) to solve the challenging situation where the radar points are very sparse on the object. The GRIF presents much robust recall performance to missing points than mean or concatenation fusion operation.
- We utilize feature pyramid network (FPN) and sparse block network (SBNet) [6] for a radar backbone network to achieve high performance and low computational cost. Note that the radar point cloud has the same data form as the LiDAR point cloud, but the point density of radar is much sparse.
- We use a multi-layer 3D anchor to detect objects located at various heights. Unlike the LiDAR point cloud provides height information, the radar point cloud only provides in-plane distance, making it challenging to detect objects height in a 3D space.

We evaluate the proposed method on the nuScenes [4] object detection task, which is the only dataset that provides the radar point cloud with 3D annotations. We also verify the effectiveness of GRIF quantitatively and qualitatively. The proposed approach significantly outperforms both image and single radar detection methods and achieves comparable performance to the LiDAR detection method.

II. Related Work

In this section, we briefly review related works of learning-based object detection method using the single and multiple modalities.

Camera-based 3D Object Detection: Monocular or stereo images provide rich texture information but do not provide direct depth information. Some studies process additional work to extract depth from the image input. Pseudo-LiDAR [7] utilizes the sub-network to obtain the disparity from the image to detect 3D objects. MonoDIS [8] proposes a loss disentangling transformation to detect the 3D object from a monocular image using two-stage architecture. However, the distance measuring performance of image is naturally inferior compared to range sensors such as LiDAR and radar.

Radar Deep Learning for Vehicle Intelligence: Existing learning-based radar studies mainly focus on classification and 2D object detection task. Major *et al.* [9] use range-azimuth-Doppler tensors to represent the radar signal and detect vehicles in the bird's eye view (BEV) space but assuming planar highway driving scenarios. Brodeski *et al.* [10] aim to detect and localize objects using a two-stage detector on the range-Doppler map, and evaluate the performance in the anechoic chamber. Kim *et al.* [11] use the time-series radar signal of the range-velocity image and classify target objects by convolutional recurrent neural network. Schumann *et al.* [12] take radar data as point cloud and process for object segmentation by using PointNet [13]. To the best of our knowledge, deep learning-based 3D object detection studies using radar that takes into account the height and z-position of objects have not been conducted.

Camera-Range Sensor Fusion for Object Detection: Most sensor fusion works use camera and LiDAR, assuming all sensor data is useful. F-PointNet [14] and Du *et al.* [15] use the cascade approach that two successive stages detect objects on each modality. The first stage detects objects' 2D bounding box on the image, and the second stage projects the 2D box into the LiDAR point cloud to regress the 3D bounding box. The performance of a cascade approach is limited to the performance of the single sensor. Meanwhile, the parallel approach fuses feature representations from each modality. MV3D [16] generates 3D object proposals from the LiDAR BEV map, then projects them into other modalities to obtain RoIs from each modality and, fuses RoI features. AVOD [17] proposes 3D proposals by fusing RoI features at the region proposal stage to achieve high proposal recall. MMF [18] fuses feature maps instead of fusing RoI features by projecting the image feature map into BEV space.

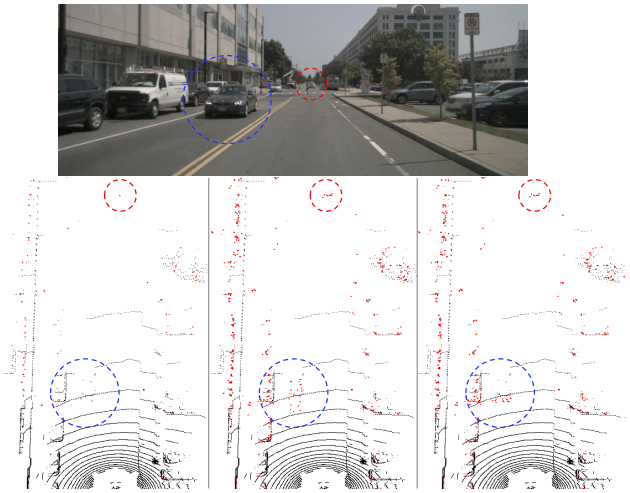


Fig. 3: Examples of radar point cloud accumulation. *Left:* single sweep, *Middle:* only ego motion compensated multiple sweeps, *Right:* both ego and moving object motion compensated multiple sweeps. Black and red points represent LiDAR points and radar points.

The following attempts propose a camera and radar sensor fusion network. Chadwick *et al.* [19] focus on detecting distant vehicles by fusing image and object-level radar data. Objects detected by the radar are projected on the front view image as an additional channel of image and detect 2D objects using SSD [20] architecture. Lim *et al.* [21] project image into the bird's eye view (BEV) by homography transform to match the coordinate systems of two sensors, assuming the planar road scene. Spatially aligned two feature maps from 2D range-azimuth heatmap and BEV image are concatenated to predict BEV bounding box using SSD [20] architecture. However, such fusion approaches are only evaluated on uncomplicated driving scenes and not thoroughly investigate how to combine two modalities.

Multiple Modalities Combining Methods: Some approaches utilize the Mixture of Expert (MoE) [22] in which the gating network explicitly assigns the weight of modalities. Mees *et al.* [23] detect people in various lighting conditions using an RGB-D camera. The separate detection networks first detect objects from RGB and depth input, and the gating network validates detected objects and predict the final classification score. AdapNet [24] produces weights of feature maps from image and depth by CNN and multi-layer perceptron (MLP), and fuses feature maps by element-wise sum for segmentation task. Kim *et al.* [25] project LiDAR point cloud into the image plane and combine LiDAR and image feature maps adaptively using trainable weight maps. Object detection is performed on the 2D image plane using SSD [20] architecture. Nonetheless, MoE fusion has not been used to detect 3D objects.

III. GRIF NET: GATED ROI FUSION NETWORK

As illustrated in Fig. 2, GRIF Net takes the monocular image and voxelized radar point cloud as input and predicts oriented 3D vehicles. It has two different feature extractors considering characteristics of modalities, and it generates two RoIs that corresponding to the coordinate system of inputs by projecting 3D anchors. Our method adaptively fuses two RoIs by GRIF on 3D region proposal network (RPN) and detection head to predict and regress 3D boxes.

Dataset	Type	Distance [m]					
		0-10	10-20	20-30	30-40	40-50	50-60
nuScenes [4]	ratio [%]	6	20	28	23	16	6
	mean [m]	1.51	1.57	1.64	1.69	1.72	1.59
	std. [m]	0.13	0.27	0.49	0.68	0.80	1.01
KITTI [5]	ratio [%]	10	23	23	18	12	12
	mean [m]	1.66	1.68	1.69	1.70	1.77	1.78
	std. [m]	0.11	0.19	0.29	0.44	0.55	0.61

TABLE II: The z-position analysis in nuScenes and KITTI.

A. Data Preprocessing

Radar Point Cloud Representation: The Frequency Modulated Continuous Wave (FMCW) radar used in nuScenes (Continental ARS 408-21) has a low distance measuring resolution and azimuth angle resolution, which are 0.39 m and 4.5° in the range of 70m at ±45°, respectively. Meanwhile, the LiDAR (e.g., HDL-64E) has a high vertical and horizontal resolutions of 0.4° and 0.08°. Furthermore, the LiDAR is usually mounted on top of the vehicle, while the radar is mounted on the bumper, making it easy to be occluded by the vehicle ahead. For these reasons, the average number of radar points within the field of view (FoV) of the front camera is 107, while the LiDAR is 14795, which is 138 times denser than radar.

As allowed in nuScenes [4] submission rule, we accumulate 6 radar sweeps collected during approximately 0.5 seconds to use the denser point cloud. While accumulating multiple radar sweeps, the displacement caused by the ego vehicle movement is compensated using accurate INS information provided by the dataset. Moreover, the displacement of each point caused by moving objects is compensated using the velocity of each point obtained by the Doppler Effect. By doing so, we get a richer point cloud without losing accurate contour of objects, as illustrated in Fig. 3. After that, we voxelize the accumulated radar points into a 3-channel voxel grid map with a 0.2-meter resolution. The voxel grid map is encoded with the absolute velocity, Radar Cross-Section (RCS) of point, and the voxel occupancy. The voxel occupancy is 1 if the grid contains the point, 0 otherwise. The velocity and RCS channels are normalized to have zero-mean. For each voxel, we select the point with the highest value if multiple points exist in a voxel.

Multi-layer 3D Anchor: We adopt RPN strategy to generate RoI proposals from image and radar by projecting 3D anchors into both modalities' feature maps. It is important that the anchors are positioned at the appropriate height in order to obtain an accurate ROI from the image. However, as shown in Table II, the standard deviation of the objects' z-position in nuScenes [4] is 1.5 times larger than KITTI [5]. Thus, anchors of varying height improve the 3D detection performance, especially at long distances.

The 3D anchor is parameterized by the center location x, y, z, size l, w, h, and orientation θ relative to the position of the ego vehicle. The size of 3D anchors is calculated into two clusters by k-means clustering on the *training set*. 3D anchor has an interval of 0.5 m on the x-y plane with two orientations, 0 and 90 degrees, and anchors are piled into three layers along the z-axis with 1 m interval.

B. Backbone Network

Image Backbone Network: The backbone network for the image stream is modeled after the Feature Pyramid Network (FPN) [26]. The image backbone network has four convolutional blocks, and each block contains 2, 2, 3, 3 residual layers [27] followed by Batch Normalization and ReLU. The size of the input image decreases by the max pooling with the factor of 2, while the numbers of feature maps increase twice from 32 to 256 on every block. The last layer increases the size twice with 1×1 convolution and bilinear up-sampling, then element-wise adds to the third block and repeating to the second block. Therefore, the size of the last feature map is 2 times down-sampled with respect to the input.

Radar Backbone Network: The backbone network for the radar stream is modeled after FPN and Sparse Block Network (SBNet) [6]. The radar point cloud is very sparse and most of the radar voxel is empty. Therefore, traditional convolution is inefficient because it operates across all the feature map where data is zero. Unlike the traditional convolution, SBNet performs convolution operation only on masked areas.

We set the mask as the circle with a radius of 1 meter around the radar point, taking into account the receptive field and the accuracy of radar. The radar point can be treated as a potential object because the point is only reflected by the object or wall but not reflected by the ground, in contrast to the many LiDAR points are reflected by the ground plane. The mask occupies 26% of the voxel grid map on average, which theoretically 3.8× speedup the radar backbone network. The size and the number of layers of the radar backbone network are the same as the image backbone network, and the SBNet parameters have a block size of 13, 9, 7, and 5 at each convolutional block with a stride of 3, offset of 1 at all blocks.

C. Detection Network

3D Box Projection and RoI Extracting: For both 3D region proposal network (RPN) and detection head, we project 3D boxes (3D anchors and 3D proposals) into feature maps to obtain RoIs from each view. As shown in Fig. 4, the 3D box is projected onto the feature maps of the front view and the bird's eye view by transformation matrix T so that the projected RoIs of each view corresponds to the 3D box in the 3D space, formulated as:

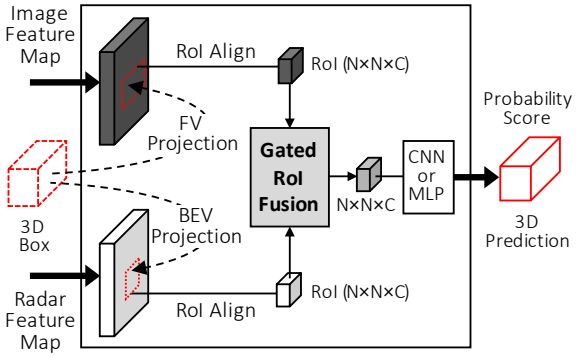


Fig. 4: 3D box projection and RoI extraction. The 3D box is projected into feature maps by each transformation matrix. Projected boxes are cropped into a regular grid RoIs by RoIAlign and combined by GRIF and used to predict the 3D box regression.

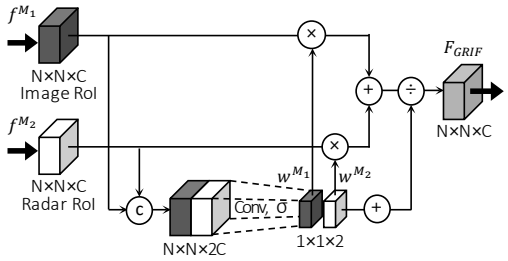


Fig. 5: Gated Region of Interest Fusion (GRIF). RoIs from image and radar are concatenated and output two scalar weights of each modality.
C: concatenation, ×, ÷, +: element-wise product, divide, add

$$\text{RoI}_v = T_{3D\to v}(\text{box}_{3D}), v \in \{FV, BEV\}$$

We employ RoIAlign [28] to minimize the quantization of the RoI boundaries since the 3D box has a continuous length in 3D space, but feature maps are quantized in different resolutions from different modalities. RoIAlign uses the bilinear interpolation to compute the exact values of RoI on the feature map and aggregate it into a regular size grid using the average. We use the grid size of 5 for both stages. The 3D RPN takes 3D anchors and predicts the 3D proposals using two convolutional layers of size 256. The detection head takes 3D proposals and refines the proposals into the oriented 3D box by three MLP of size 2048.

Gated Region of Interest Fusion (GRIF): Prior sensor fusion methods combine two features by concatenation or element-wise mean operation. However, the output of the concatenation or mean operation is sensitive to changes of the input data. In contrast, we combine RoI features from image and radar feature maps by convolutional Mixture of Experts (MoE), so that MoE explicitly assign weights to features, as described in Fig. 5.

Given RoI features from two modalities f^{M_1}, f^{M_2}, gating network predicts weights w^{M_1}, w^{M_2} by the convolution and sigmoid layer. The convolution layer has a size of N×N×2 without padding. Afterward, the RoI features are multiplied by weights and element-wise added together. The element-wise added RoI feature is divided by the sum of weights to normalize and used to predict object. The fusion method can be expended to an arbitrary n number of modalities as

$$F_{GRIF} = \sum_{i=1}^{n} w^{M_i} \cdot f^{M_i}, \text{ with } \sum_{i=1}^{n} w^{M_i} = 1$$

D. Implementation Details

We use the image size of 896×1600 and discretize the radar point cloud in the range of [0, 70.4] × [-40, 40] into the size of 352×400. Two anchors have length, width, and height of ((4.52, 1.91, 1.67), (5.52, 2.18, 2.11)) meters with a z center of (2.6, 1.6, 0.6) meters. The 3D region proposal stage generates 1024 proposals during training, while using 300 proposals in inference, and applies Non-Maximum Suppression (NMS) on BEV space with 0.8 and 0.001 Intersection over Union (IoU) threshold. We assign a positive and negative label to anchors with the matching threshold using the Euclidean center distance instead of IoU. For the 3D RPN and detection network, the anchor closer than 1 and 0.6 meters is assigned to the positive, and the anchor farther

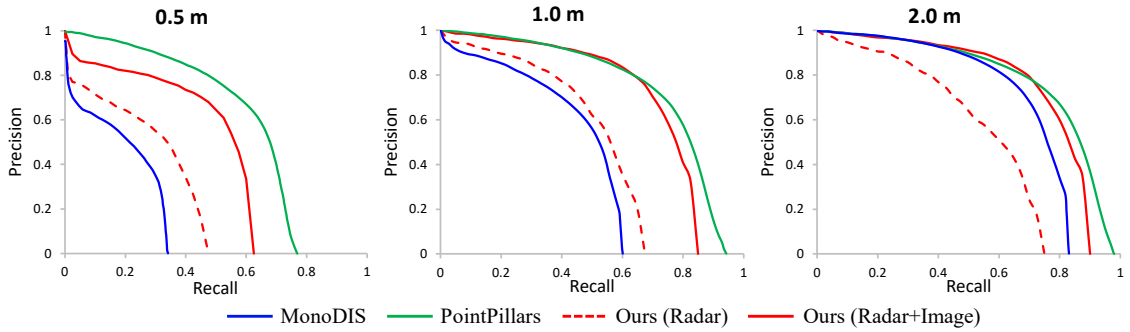


Fig. 6: Precision-recall curves at different thresholds.

than 1.25 and 0.8 meters to the negative. During training, we randomly drop out the one of input with a 30% chance to avoid overfitting.

We apply multi-task loss for classification, offset, and orientation in an end-to-end fashion. We use binary cross-entropy for the classification loss L_{cls}. For the regression loss L_{reg} and orientation loss L_{dir}, we compute the smooth L_1 loss on each dimension of the object $(x, y, z, \log(l), \log(w), \log(h))$ and orientation θ as [17]. The total loss is defined as:

$$L = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} + \lambda_{dir} L_{dir}$$

where weights are $\lambda_{cls} = 3$, $\lambda_{reg} = 5$, and $\lambda_{dir} = 1$.

We initialize the model from random initialization and use mini-batch containing 1 frame with 800 and 1600 RoI samples for the 3D RPN and detection network. We train the network using Adam optimizer with an initial learning rate of 0.0001. The learning rate is decayed by a factor of 0.1 at 150k iterations, and training ends after 170k iterations.

IV. Experiments

In this section, we evaluate the proposed method on the nuScenes [4] and show the effectiveness of the GRIF method quantitatively. We also conduct the number of ablation studies to verify the benefits of the components since apples-to-apples comparisons between radar and LiDAR methods are not available. Finally, we present qualitative results in challenging situations with some failure cases.

A. Dataset and Metric

The proposed method is trained and evaluated on the challenging nuScenes dataset. The nuScenes is collected with 6 cameras, 1 LiDAR, and 5 radars that cover 360°. In this work, we use one front camera and 3 front radars. The data is taken from different scene locations with various weather and lighting condition, which is more diverse and challenging than KITTI. The nuScenes contains 19.4% and 11.6% of data collected in the rain and night. The dataset provides 1000 selected scenes of 20 seconds duration each at 2Hz. Of the 1000 scenes, 700, 150, and 150 are used for training, validation, and testing, respectively. We filter out vehicles farther than 50 meters or bounding boxes that do not contain any radar or LIDAR points, according to the official nuScenes evaluation rules.

We evaluate predicted 3D objects using the average precision (AP) metric with a match threshold of 2D center distance $d \in \mathbb{D} = \{0.5, 1, 2, 4\}$ meters instead of IoU as introduced in nuScenes. In addition to AP, we evaluate a maximum recall performance at 10% precision using 300 proposals with the distance match threshold.

Method	Modality	AP [%]			
		0.5m	1.0m	2.0m	4.0m
MonoDIS [8]	Image	10.5	36.0	64.8	80.0
PointPillars [29]	LiDAR	53.0	69.6	74.1	76.9
Ours	Radar	25.5	44.0	51.7	54.2
	Radar+Img	44.1	66.5	71.9	74.9

TABLE III: Evaluation results on the nuScenes.

B. Quantitative Evaluation

Comparisons with alternative approaches on nuScenes: We evaluated 3D object detection results using AP for the car class on the *validation set*. Our single radar approach has radar stream only while the other architecture configurations (e.g., input representation, backbone, number of layers) are kept identical, and uses radar RoI as it without fusion. We compare the proposed method with previously published works in Table III and Fig. 6. Note that the results of other methods are from the nuScenes leaderboard.

Our single radar method yields a low recall because many objects do not have any radar point on it due to the low resolution and occlusion, and it cannot be overcome by using only radar. We also hypothesize that the lack of contextual information makes it difficult to distinguish vehicles with metallic objects and lead to low precision. However, the performance or radar has been significantly improved by fusing with the image. The proposed radar and image fusion approach outperforms the single radar method and MonoDIS [8] by 22.5% and 30.5% AP at 1 m threshold. Also, we achieve a comparable result, which is 3.1% AP lower than the LIDAR-based PointPillars [29] at 1 m threshold despite using a very cheap and sparse radar point cloud. The performance differences at strict thresholds (0.5 and 1 m) are substantial because the characteristics of each modality vary considerably. At the same time, the performances at easy thresholds (2 and 4 m) are similar in all methods.

Also, we argue that the performance of radar can be improved depending on the mounting position of the sensor. The LiDAR is mounted on the top of the vehicle, whereas the radar is installed on the front bumper so that the field of view of radar can be occluded by surrounding vehicles. Missing radar points by occlusion degrades the performance of the single radar method. The qualitative examples are introduced in §IV-D.

Fusion Method	# of points	Recall [%] 0.5m	1.0m	2.0m
mean	0≤pts<5	17.6	34.3	47.5
	5≤pts<10	50.6	77.2	85.2
	10≤pts	51.8	80.2	85.7
concat	0≤pts<5	21.8	37.5	51.6
	5≤pts<10	51.4	78.4	88.6
	10≤pts	62.9	86.7	**93.0** (+0.1)
GRIF	0≤pts<5	**24.6** (+2.8)	**45.6** (+8.1)	**53.1** (+1.5)
	5≤pts<10	**52.1** (+0.7)	**80.9** (+2.5)	**89.5** (+0.9)
	10≤pts	**64.3** (+1.4)	**88.9** (+2.2)	92.9

TABLE IV: Effects of GRIF at the different number of radar points in the object bounding box in terms of maximum recall. Round brackets denote the gap from the second-best result.

Modality	Dist. [m]	AP [%] 0.5m	1.0m	2.0m	4.0m
LiDAR + Image	0-10	**76.9**	**77.2**	**84.4**	**84.8**
	10-20	**76.6**	**83.6**	**85.0**	**85.0**
	20-30	**66.8**	**75.1**	**75.9**	**80.3**
	30-40	**50.3**	**60.8**	**63.3**	**69.3**
	40-50	**33.2**	46.4	49.9	55.5
	50-60	**30.2**	38.1	43.6	46.0
Radar + Image (Ours)	0-10	45.7	61.3	63.1	69.5
	10-20	51.6	67.3	73.0	74.2
	20-30	46.7	69.7	73.3	77.9
	30-40	34.7	58.5	62.8	67.9
	40-50	23.6	**47.1**	**54.1**	**56.6**
	50-60	20.1	**38.7**	**45.4**	**47.6**

TABLE V: Comparison between radar and LiDAR by the distance of object in terms of average precision (AP).

Effect of GRIF: In Table IV, we compare the recall performance of the proposed RoI fusion method GRIF with the element-wise mean and concatenation operation. We group the objects according to the number of radar points in the bounding box to evaluate the effectiveness of GRIF when the object has few radar points. Models use the same fusion method on both 3D RPN and detection network. As hypothesized, GRIF performs better than others in the case where the number of points is fewer. In particular, the recall of GRIF is 8.1% higher than the second-best concatenation method, where the number of points is less than 5 at 1 m threshold. However, we observe that GRIF does not provide much gain at other thresholds. We conclude that the 0.5 m threshold is too severe, and 2 m is too generous considering the accuracy of radar, so the improvement is not significant.

Advantage in detection range: Table V shows the AP comparisons between radar and LiDAR by the distance of objects. The LiDAR-image fusion method takes a LiDAR voxel representation as [16]. We train both models in the same manner but filter out ground truths up to 60 m rather than 50 m to verify the long-range performance. At range closer than 30 m and the 0.5 m threshold, the LiDAR performs better due to the precise and dense point cloud of LiDAR. On the other hand, at range farther than 30 m and over the 1 m threshold, the performance of radar is comparable and even exceed the LiDAR with the advantage of a long detection range. We hypothesize that the radar-based method is less sensitive to the distance because the LiDAR points become sparse as the distance increases, while the radar points are uniform compared to the LiDAR.

C. *Ablation Study*

We conduct ablation studies to analyze the effects of each component of our method in Table VI. We train models using a full *train set* and evaluate on a *minival set* obtained by extracting one frame out of every 10 frames from the *val set*.

Multi-layer Anchor: In Table VI-a, we compare the multi-layer anchor to the single-layer anchor with the z-center of 1.6 meters. As we hypothesized in §III-A and Table II, the multi-layer anchor improves AP performance by 4.1 points and maximum recall by 3.7 points at 0.5 m threshold by detecting vehicles located at various z-location. The result shows that the object that cannot be detected on a single-layer anchor can be detected using anchors on other layers.

Backbone Network: Table VI-b shows the comparison of two backbone networks. The baseline ResNet-10 has the same numbers of residual blocks without SBNet [6], and FPN [26] is replaced by the 4× bilinear upsampling to make the last feature map the same size as ours. As expected, a deep and advanced backbone brings a significant AP gains in all thresholds.

# of Layer	AP [%] 0.5m	1.0m	Recall [%] 0.5m	1.0m
single-layer	39.2	65.8	58.8	82.2
multi-layer	**44.0**	**69.9**	**62.5**	**85.3**
	+4.8	*+4.1*	*+3.7*	*+3.1*

(a) **Multi-layer Anchor**

Backbone	AP [%] 0.5m	1.0m	2.0m	4.0m
ResNet-10	41.2	64.4	69.9	72.4
FPN+SBNet	**44.0**	**69.9**	**71.9**	**76.5**
	+2.8	*+5.5*	*+2.0*	*+4.1*

(b) **Backbone Network**

Method	RoI Size	AP [%] 0.5m	1.0m	2.0m	4.0m
RoIPool	3×3	38.8	63.5	69.9	75.9
	5×5	42.4	68.6	71.3	76.2
RoIAlign	3×3	40.6	66.2	69.8	76.2
	5×5	**44.0**	**69.9**	**71.9**	**76.5**

(c) **RoI Extracting Method**

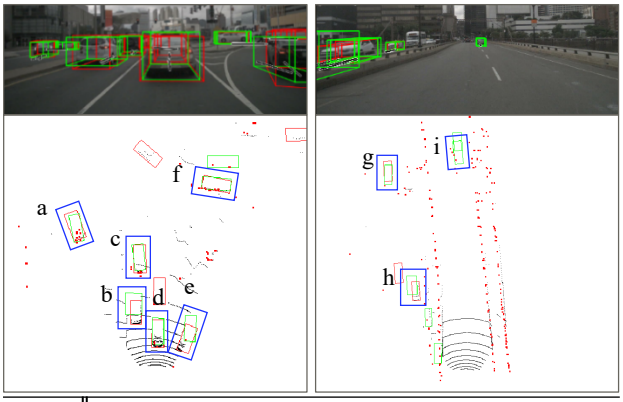
3D RPN	Detection Network	AP [%] 0.5m	1.0m	Recall [%] 0.5m	1.0m
mean	mean	38.9	63.8	55.3	79.1
concat	concat	41.1	65.4	58.7	82.2
GRIF	concat	40.8	68.0	59.8	82.4
concat	GRIF	42.4	68.4	58.9	84.0
GRIF	GRIF	**44.0**	**69.9**	**62.5**	**85.3**

(d) **RoI Fusion Method**

TABLE VI: Ablation study on the nuScenes *minival set*.

RoI Extracting Method: RoI extracting methods are compared according to the RoI feature size and the extracting method, as shown in Table VI-c. The result shows that the RoI size plays a more important role in improving the performance than the RoI extracting method. At the 0.5 m threshold, RoIAlign [28] shows a gain of 1.6 points over RoIPool, and a 5×5 size grid improves AP about 3.4 points over a 3×3 size grid. As the one voxel of the last radar feature map occupies 0.4×0.4 meters, the misaligned voxel can lead to a significant performance difference at the strict threshold.

RoI Fusion Method: Table VI-d shows the effects of fusion methods on the 3D RPN and detection network. We observe that the effect of the fusion method on the 3D RPN is less significant than the detection network. We hypothesize that this could be because the RPN suggests hundreds of proposals so that most objects are detected on the RPN regardless of the fusion method. On the other hand, GRIF on the detection network improves both AP and recall over the concatenation, and the element-wise mean yields the worst performance.



	a	b	c	d	e	f	g	h	i
# of pts	14	0	5	4	1	9	1	0	2
image	0.58	0.83	0.69	0.65	0.70	0.63	0.68	0.75	0.49
radar	0.42	0.17	0.31	0.35	0.30	0.37	0.32	0.25	0.51

Fig. 7: Examples of weights assigned to image and radar feature at the detection head. # of pts is the number of radar points inside bounding box.

D. Qualitative Results[1] *and Discussion*

We qualitatively analyze the weights assigned to image and radar RoI by GRIF in Fig. 7. On RoIs with many radar points and sufficient visual cues (c, d, f), GRIF assigns more weight to the image, whereas on RoIs with radar points but weak visual cues (a, i), GRIF assigns more weight to the radar. RoIs without any radar point (b, h) use almost image features. Interestingly, RoIs with one or two radar points (e, g) have similar weight distributions with the case (c, d, f).

We visualize the qualitative detection results in Fig. 8 (a~e). The proposed method produces accurate 3D bounding boxes in various environments. The network detects vehicles

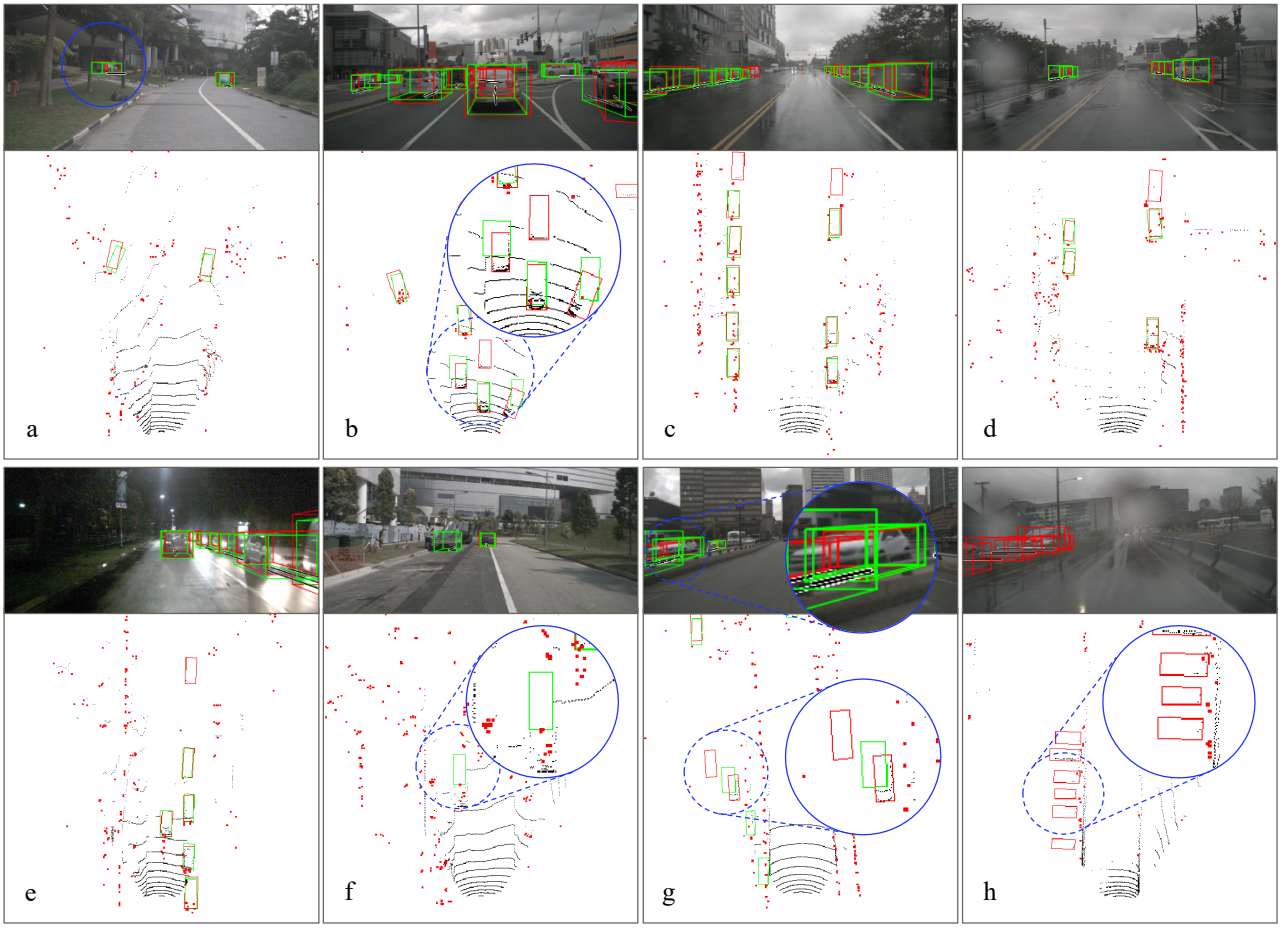


Fig. 8: Qualitative results of 3D object detection and failure cases on *val set*. Black and red dots represent the LiDAR and radar point. Red and green boxes denote ground truths and detection results. Note that the LiDAR point cloud is used only for visualization.

[1] More qualitative results are at `https://youtu.be/CyJrMpBhEGI`

located at various altitudes and far distances (Fig. 8a). The proposed method accurately estimates the orientation of vehicles and even detects vehicles without any radar point (Fig. 8b). It also operates robustly in the rainy road (Fig. 8c and d) and at night light condition (Fig. 8e).

We observe several failure cases as highlighted with the blue circle in Fig. 8 (f~h). Our network tends to classify the pick-up truck as a car due to its similar appearance (Fig. 8f). The most common failure cases are due to the missing radar points on objects, even the proposed method attempts to overcome the abnormal sensor input. The field of view (FoV) of radar is often occluded by other vehicles on the crowded road (Fig. 8b). Moreover, the radar does not provide the point beyond the guardrail and detects through the barbed-wire fence of a parking lot. In the absence of radar inputs, GRIF Net succeeds in detecting some visible vehicles (Fig. 8g) but fails to detect vehicles with low visual cues (Fig. 8h). We claim that the low mounting position of the radar may affect detection performance.

V. Conclusion

We have proposed the camera-radar sensor fusion-based robust 3D object detection network, named GRIF Net. We introduce the multi-modal fusion method GRIF to overcome the characteristics of radar that the points are very sparse on the object. The GRIF utilizes a gating mechanism to choose the appropriate modality. The experiments on nuScenes verify the robustness of GRIF on the vehicles with very sparse radar points and show the effectiveness of the radar sensor in detecting vehicles over long distances. Our approach achieves comparable performance with the LiDAR method despite using the low-cost radar and shows the potential of the radar sensor in autonomous driving.

References

[1] Y. Kim and D. Kum, "Deep Learning based Vehicle Position and Orientation Estimation via Inverse Perspective Mapping Image," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[2] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "On-Road Vehicle Detection and Tracking Using MMW Radar and Monovision Fusion," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* vol. 17, no. 7, 2016.

[3] R. Nabati and H. Qi, "RRPN: Radar Region Proposal Network for Object Detection in Autonomous Vehicles," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[4] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] M. Ren, A. Pokrovsky, B. Yang, and R. Urtasun, "SBNet: Sparse Blocks Network for Fast Inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[8] A. Simonelli, S. R. R. Bulò, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling Monocular 3D Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] B. Major, D. Fontijne, R. T. Sukhavasi, and M. Hamilton, "Vehicle Detection With Automotive Radar Using Deep Learning on Range-Azimuth-Doppler Tensors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop*, 2019.

[10] D. Brodeski, I. Bilik, and R. Giryes, "Deep Radar Detector," in *Proceedings of the IEEE Radar Conference (RadarConf)*, 2019.

[11] S. Kim, K. Lee, S. Doo, and B. Shim, "Moving target classification in automotive radar systems using convolutional recurrent neural networks," in *Proceedings of the 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018.

[12] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic Segmentation on Radar Point Clouds," in *Proceedings of the 21st International Conference on Information Fusion*, 2018.

[13] C. R. Qi, H. Su, K. Mo, and L. 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 (CVPR)*, 2017.

[14] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D Object Detection from RGB-D Data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] X. Du, M. H. Ang, S. Karaman, and D. Rus, "A General Pipeline for 3D Detection of Vehicles," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[16] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[18] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-Task Multi-Sensor Fusion for 3D Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] S. Chadwick, W. Maddern, and P. Newman, "Distant Vehicle Detection Using Radar and Vision," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[20] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[21] T.-y. Lim, B. Major, D. Fontijne, and M. Hamilton, "Radar and Camera Early Fusion for Vehicle Detection in Advanced Driver Assistance Systems," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) Workshop*, 2019.

[22] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation,* vol. 3, no. 1, 1991.

[23] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[24] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive semantic segmentation in adverse environmental conditions," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[25] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust Deep Multi-modal Learning Based on Gated Information Fusion Network," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018.

[26] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast Encoders for Object Detection from Point Clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.