RADIANT: Radar-Image Association Network for 3D Object Detection

Yunfei Long[1], Abhinav Kumar[1], Daniel Morris[1], Xiaoming Liu[1], Marcos Castro[2], Punarjay Chakravarty[2]

[1] Michigan State University
[2] Ford Motor Company
{longyunf, kumarab6, dmorris, liuxm}@msu.edu, {mgerard8, pchakra5}@ford.com

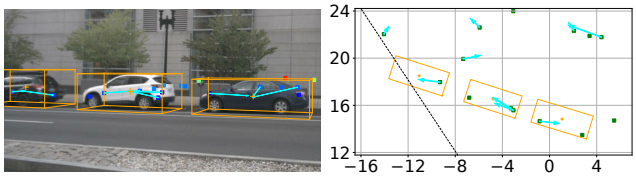
Abstract

As a direct depth sensor, radar holds promise as a tool to improve monocular 3D object detection, which suffers from depth errors, due in part to the depth-scale ambiguity. On the other hand, leveraging radar depths is hampered by difficulties in precisely associating radar returns with 3D estimates from monocular methods, effectively erasing its benefits. This paper proposes a fusion network that addresses this radar-camera association challenge. We train our network to predict the 3D offsets between radar returns and object centers, enabling radar depths to enhance the accuracy of 3D monocular detection. By using parallel radar and camera backbones, our network fuses information at both the feature level and detection level, while at the same time leveraging a state-of-the-art monocular detection technique without retraining it. Experimental results show significant improvement in mean average precision and translation error on the nuScenes dataset over monocular counterparts. Our source code is available at https://github.com/longyunf/radiant.

Introduction

Three-dimensional object detection is a core vision problem where the task is to infer 3D position, orientation, and classification of objects in a scene. Applications that rely on this include robotics (Saxena, Driemeyer, and Ng 2008), gaming (Rematas et al. 2018), and automotive safety (Simonelli et al. 2020). In the latter application, Advanced Driver Assistance Systems can move beyond automated braking and use precise location and orientation of nearby vehicles and other objects to perform collision avoidance maneuvers. However, a key limiting factor is the relatively poor accuracy of 3D object detection, both in current systems and in state-of-the-art (SOTA) methods that rely on widely used sensors, namely cameras (Park et al. 2021; Lu et al. 2021) and radars (Yang et al. 2020). Thus, there is a significant need for improved 3D object detection, which is the goal of this paper.

Image-based object detection achieves high 2D accuracy (Brazil and Liu 2019b); for instance, DD3D (Park et al. 2021) achieves 94% 2D mean average precision (mAP) for cars on KITTI (Geiger, Lenz, and Urtasun 2012). However, the performance of these same SOTA methods drops precipitously on 3D object detection with DD3D (Park et al. 2021) only achieving 16.9% 3D mAP of cars on KITTI. The lower 3D performance of monocular detection comes largely from poor depth estimation (Ma et al. 2021; Kumar et al. 2022). This is expected as image projection removes depth, and recovering object depth from an image suffers from the depth scale ambiguity (Tang, Dorn, and Savani 2020), and is therefore error-prone. Indeed, when (Xu et al. 2021) uses precise depth from LiDAR, mAP of cars increases to over 90%. Typically LiDAR is a specialized sensor that is relatively expensive (Kumar, Brazil, and Liu 2021) and only available on a small fraction of vehicles. On the other hand, radar is small (Lien et al. 2016), inexpensive, and widely available on existing vehicles so we choose to use radar for the broader impact. Thus, this paper explores the fusion of direct depth measurements available from the radar with a 3D monocular detector.



(a) Pixel Offset (b) BEV Offset

Figure 1: Predicted radar-offsets (cyan arrows) from radar points (dots) to object centers (orange plus) in (a) pixel space and (b) bird's-eye view (BEV) in meters. RADIANT trains a network to predict these offsets and improves monocular 3D detection. Orange boxes represent the ground truth (GT) bounding boxes, and dashed lines denote borders of the camera field of view.

The choice to use automotive radar rather than LiDAR presents several challenges. At first glance, one might consider using a similar 3D object-detector on radar points as for LiDAR (Wang et al. 2021f). However, that does not work because the radar data has completely different characteristics from the LiDAR point cloud (Wang et al. 2021f). First, each radar-point sweep is much sparser than a typical LiDAR measurement in azimuth (Yang et al. 2020) and has a single row in elevation, leaving radar coordinates without height information. Thus, differing from LiDAR, radar cannot acquire accurate shapes with dense point clouds. Additionally, the radar point positions have significant azimuth errors (Yang

et al. 2020) and are much less accurate (Wang et al. 2021f) as measurements of object surfaces. Next, for each radar scan there is often only a single radar return at longer ranges and sometimes no radar returns for small objects. This sparsity makes 3D object detection solely based on radar points difficult. Thus, rather than detection-level fusion, our approach uses feature-level fusion to augment radar points and these augmented radar points to refine monocular object depths. Finally, since the widely-used KITTI dataset (Geiger, Lenz, and Urtasun 2012) does not include radar data, we use the nuScenes dataset (Caesar et al. 2020) for experiments.

The association and fusion of image and radar modalities is possible at input-level, feature-level, or even at detection-level. However, any such association must address the missing height, imprecise angular location of radar returns and the observation that occluded portions of objects also return radar points (Long et al. 2021b). Rather than using inaccurate radar projections on image for association, this paper uses a neural network to explicitly predict point-wise 3D object centers (see Fig. 1). Predicting these point-wise object centers from a trained network allows us to use radar to correct depth errors of monocular detection, which results in a new SOTA for radar-camera 3D object detection.

This paper presents a radar-camera fusion method named RADar-Image Association NeTwork (RADIANT) for 3D object detection, including the following contributions:

- Our method enhances radar returns to obtain 3D object center detections from each radar return.
- We achieve camera-radar association at the detection level using the enhanced radar locations.
- Our architecture can leverage multiple different pre-trained SOTA monocular methods.
- We improve monocular object depth estimates by fusing enhanced radar depths and achieve new improved SOTA performance on nuScenes.

Related Work

Monocular Detection Differing from LiDAR-based detection (Shi, Wang, and Li 2019), monocular 3D detection is widely applied for its low cost and simple configuration (Brazil and Liu 2019a). Researchers have been improving detection performance via upgrading detection frameworks (Wang et al. 2021d; Liu et al. 2019), losses (Simonelli et al. 2020) and non-maximum suppression (Kumar, Brazil, and Liu 2021) as well as performing joint detection and 3D reconstruction (Liu and Liu 2021). To reduce 2D to 3D ambiguities, various strategies have been developed, *e.g.*, Pseudo-LiDAR (Wang et al. 2019; Ma et al. 2019, 2020; Simonelli et al. 2021; Park et al. 2021), novel convolutions (Ding et al. 2020) and backbones (Kumar et al. 2022), considering camera geometry (Zhou et al. 2021), using shape models (Liu et al. 2021; Chabot et al. 2017) and leveraging videos (Brazil et al. 2020).

Radar-Camera Fusion Radar has been fused with LiDAR and camera. Radar-LiDAR fusion has been utilized in 3D object detection (Yang et al. 2020) and object tracking (Shah et al. 2020) as radar complements LiDAR in long range and motion (*i.e.*, Doppler velocity) measurements. Nevertheless, most works combine radar with camera for advantages listed in the Introduction section. There has been a surge in research of radar-camera fusion recently since the release of new autonomous driving datasets (Ouaknine et al. 2021; Dong et al. 2020; Barnes et al. 2020; Wang et al. 2021e; Shuai et al. 2021; Caesar et al. 2020) with images and radar data collected. The radar data are in raw data formats (Ouaknine et al. 2021; Dong et al. 2020; Barnes et al. 2020; Wang et al. 2021e) (such as Range-Azimuth-Doppler format (Major et al. 2019)) or processed formats (*i.e.*, radar point clouds (Shuai et al. 2021; Caesar et al. 2020)). Raw radar formats contain denser measurements while point clouds are sparser but have less noise. Algorithms are designed according to specific radar formats used. In this paper, we use the radar point cloud format from the nuScenes dataset (Caesar et al. 2020). The goals for radar-camera fusion include depth completion (Long et al. 2021b; Lee, Jovanov, and Philips 2021), full-velocity estimation (Long et al. 2021a), object tracking (Nabati, Harris, and Qi 2021) and object detection (Nabati and Qi 2019, 2021).

A focus of radar-camera fusion is 2D object detection on images (Li and Xie 2020; Shuai et al. 2021; Nabati and Qi 2019; Yadav, Vierling, and Berns 2020), where projected radar points are used as extra features or candidates for potential objects. For example, method (Nabati and Qi 2019) maps radar detections to the image coordinate system and generates anchor boxes for each mapped radar detection point. Here radar plays an important role when the image is not clear because of darkness or long distances. However, the research on 3D object detection via radar-camera fusion is still at an early stage with few publications. One top-performing work is CenterFusion (Nabati and Qi 2021), which directly combines monocular detections with raw radar points in the neighborhood followed by a regression head to refine the depth estimate. In light of the significant gap between radar points and object centers, we believe this apples and oranges combination limits the utility of radar depths in CenterFusion. Our method, on the other hand, estimates the 3D object center for each radar return. This geometric correction performed by the radar branch then enables our fusion module to effectively combine multiple estimates of 3D center points from radar and camera for better 3D detection.

Background and Definitions

Radar-Positives. A key component of training a detection network involves the choice of candidates and the rule to construct positives. FCOS3D (Wang et al. 2021b) treats each pixel as an object candidate, and positive pixels for training detections are defined as small regions in the neighborhood of projected centers of 3D objects. To train the network to predict radar offsets, RADIANT follows the same strategy but now treats each radar return as an object candidate and considers the associated projected radar pixel as positive. Association of radar returns with an object is easy if the radar projection always falls on the object. However, radar returns returned by an object are sometimes outside the object bounding box due to measurement errors (Yang et al. 2020). Moreover, the nuScenes dataset also does not provide ground

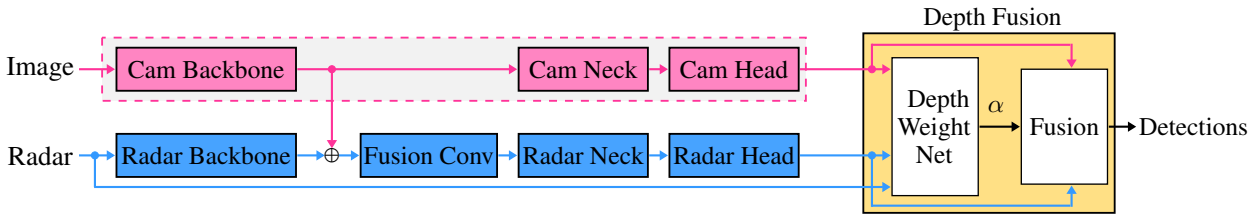


Figure 2: Overview of RADIANT. RADIANT architecture has two parallel input branches, an image branch and a radar branch that both operate in the image space. More details of these branches are shown in Fig. 3. The depth fusion module combines depth estimates from both the camera and radar heads to obtain a refined overall depth for each detection.

truth (GT) object labels for radar points. Thus, we carry out the radar-object association for training according to the position and velocity consistency. If the distance between an outside radar point and the object GT bounding box is smaller than a threshold and the projection of GT object velocity on radial direction is close to the Doppler velocity from the radar point, we associate the radar pixel with the object and consider that radar pixel as positive.

Radar Depth Offset. Radar depth offset $\Delta\hat{z}^r$ is the depth difference between the 3D center of its associated object z and a positive radar point z^r, and is thus given by

$$\Delta\hat{z}^r = z - z^r. \tag{1}$$

As mentioned, this residual depth is a result of radar measurement error and relative position of radar hit on object surfaces. We infer the residual depth from both radar and image information around.

RADIANT

Monocular 3D detection without depth as input suffers from inaccurate depth estimation (Ma et al. 2021; Wang et al. 2021a), especially for far objects. Our goal is to *upgrade* 3D camera detections with more accurate depth from radar with minimal changes to the image detection pipeline. To achieve this we address two difficulties. (1) While radar returns typically provide more precise depth than camera-based detections, which part of an object they measure can be difficult to determine as it may be the front surface or an internal or occluded point on the object. This unknown offset can add significant error to a radar depth estimate when combined with an image-based detection. (2) The radar return point can sometimes be outside the true object bounding box, and this complicates the association between radar points and objects.

Our solution is to train a radar-focused network to predict the unknown offsets between radar points and object centers. These offsets include both an image-plane offset to the projection of the 3D center, and a depth offset to the object center. Assuming these offsets are correctly estimated, the association between radar points and objects becomes much easier. Furthermore, since they predict the object center, we use the offsets to correct the object center depth.

Our architecture consists of two branches and a fusion block as shown in Fig. 2. The upper branch is a monocular 3D detection network, such as FCOS3D (Wang et al. 2021b), which remains unchanged, while the lower branch is the radar detection branch. The image branch predicts object centers and pixel offsets to these centers in the image plane. Now radar points are projected into the image plane, providing coarse alignment with image detections, and processed with the radar backbone network. Since radar points are sparse and lack contextual information, we bring features from the image backbone into the radar network providing image-plane spatial information to the radar pipeline. Then the radar neck and head portions perform the radar-based detection in the image space, but only at radar pixels, *i.e.*, the projection of radar points onto the image plane, at five resolutions to maintain consistency with FCOS3D. We then fuse these two sets of detections through a *depth fusion* module that updates the predictions' depths using a confidence score.

Radar Branch

One of the goals in designing RADIANT is to make minimal changes to the existing monocular architectures. Therefore, RADIANT builds seamlessly on top of an existing SOTA monocular network, such as FCOS3D (Wang et al. 2021b), which can be separately trained. While it would be natural to simply augment color images with additional radar channels and retrain the image network, we found this ineffective, with the resulting network unable to benefit from the radar data. Instead, we use a separate backbone, ResNet-18 (He et al. 2016), for the radar processing and freeze the image branch while training the radar branch (see Fig. 3).

The inputs to the radar branch are in image coordinates with values on radar projections, consisting of radar depth, radar bird's-eye view (BEV) coordinates, Doppler velocity, and a mask for radar pixels. Except for the backbone and losses, the majority of the radar branch is similar to the image branch, with a radar backbone processing inputs and generating radar features, which are concatenated with image features at three resolution levels, then go through an independent neck consisting of a Feature Pyramid Network (FPN) (Lin et al. 2017) and the radar heads. The radar branch outputs data in the same space, *i.e.*, five levels of image resolutions, and uses the same classification and regression losses. This makes the radar branch outputs compatible with the image branch outputs. RADIANT performs image and radar fusion at two stages: feature-level fusion from the backbone and detection-level fusion after the heads.

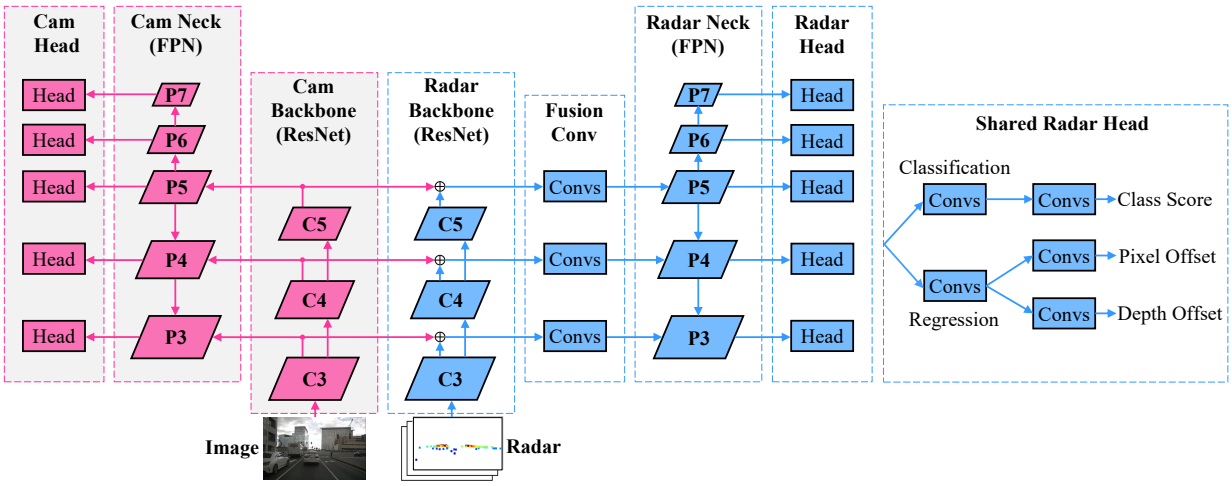


Figure 3: RADIANT Architecture Details. RADIANT architecture includes two parallel branches. On the left (in magenta) is the unchanged monocular detection pipeline. On the right (in blue), the radar network processes image-projected radar points and borrows features from the monocular network to predict offsets to the radar pixels and to their depths in the shared radar head. C3 to C5 denote feature maps from level 3 to 5 in the backbone, and P3 to P7 represent feature maps from level 3 to 7.

Radar Heads

Radar heads take in fused radar and image features of five resolutions and predict class scores as well as relative positions to the object center, namely the depth offset and the pixel offset. The radar head ignores prediction of object sizes as radar points are too sparse to reveal shape information. We build radar heads independently from monocular camera heads because of the differences in what they predict as well as the definition and positions of positive pixels: (1) camera heads estimate the full depth of objects while radar heads estimate offset depth with respect to radar measured depth; (2) positive camera pixels are only a 3×3 region at the object center while positive radar pixels may be further away from the center. Nevertheless, we use the same strategy of FCOS3D (Wang et al. 2021b) to assign positive pixels to different resolution levels so that larger objects on images are detected at lower resolutions. In summary, the radar head is complementary to the camera head with more accurate position estimation for those objects with radar hits on them.

Depth Fusion Module

Detection heads generate detection candidates for pixels classified as positive. As the depths from radar and camera heads are predicted independently, to take advantage of both depths, they are fused in the Depth Fusion module as follows. First, radar pixels are associated with monocular detection candidates. This is straightforward as the radar head outputs depth and pixel offsets, and these can be matched to 3D centers of detection candidates. Second, a confidence-based weight is predicted for each radar pixel enabling its depth to be combined with the monocular-based depth. Note that we focus on using radar to enhance only depth prediction since the depth accuracy is a primary merit of radar compared to camera, while other aspects of detection such as object size and image position are unlikely to be improved by radar. This fusion process is described in detail as follows.

Radar-Camera Association. RADIANT outputs two sets of detections from each head: $\{\mathbf{b}_i^c\}_i$ from camera head and $\{\mathbf{b}_j^r\}_j$ from radar head, where i and j represent indices of detections from camera and radar head, respectively. Typical outputs from the camera head are box proposals for detection $\mathbf{b}_i^c$, consisting of projected center $(\hat{u}_i^c, \hat{v}_i^c)$, depth $\hat{z}_i^c$, classification index $\hat{\mathcal{Y}}_i^c$, detection score $\hat{\sigma}_i^c$, dimensions of the 3D box, orientation and deltas for 2D detections. The outputs from the radar head are box deltas for detection which contain the pixel offsets $(\Delta\hat{u}_j^r, \Delta\hat{v}_j^r)$ from the projected radar point (u_j^r, v_j^r), depth offset $\Delta\hat{z}_j^r$ of the object center from the radar depth z_j^r, radar classification index $\hat{\mathcal{Y}}_j^r$ and detection score $\hat{\sigma}_j^r$.

Since the radar head does not affect 2D detection, 3D dimensions and orientations, we omit these variables in the camera outputs $\mathbf{b}_i^c$ in the subsequent paragraphs for brevity. In other words, we only specify the relevant portion of $\mathbf{b}_i^c$ vector in the following text. We now write these two set of detections as

$$\begin{aligned} \mathbf{b}_i^c &= \left(\hat{u}_i^c, \hat{v}_i^c, \hat{z}_i^c, \hat{\mathcal{Y}}_i^c, \hat{\sigma}_i^c\right), \\ \mathbf{b}_j^r &= \left(\Delta\hat{u}_j^r, \Delta\hat{v}_j^r, \Delta\hat{z}_j^r, \hat{\mathcal{Y}}_j^r, \hat{\sigma}_j^r\right). \end{aligned} \tag{2}$$

We then filter the box proposals from both camera and radar as follows. The camera head outputs a maximum of 1,000 boxes with the highest scores on each level and with $\hat{\sigma}_i^c > T_c$ where T_c denotes the minimum threshold for the box to be valid. We employ a similar procedure for radar detection candidates and consider radar projections with $\hat{\sigma}_j^r > T_r$. After the filtering step, we have good box proposals from the two modalities.

We now have the radar pixel $\left(u_j^r, v_j^r\right)$ and corresponding depth/pixel offsets which we use for the radar-camera associ-

ation. We calculate the projected centers $\left(\hat{u}_j^r, \hat{v}_j^r\right)$ and depths $\hat{z}_j^r$ of the boxes as

$$\left(\hat{u}_j^r, \hat{v}_j^r\right) = \left(u_j^r, v_j^r\right) + \left(\Delta\hat{u}_j^r, \Delta\hat{v}_j^r\right), \\ \hat{z}_j^r = z_j^r + \Delta\hat{z}_j^r. \tag{3}$$

We consider a camera proposal is associated with the radar proposal if the predicted class labels of the two modalities match, and the projected centers and the depths are close to each other. In other words, we take a camera proposal $\mathbf{b}_i^c$ and iterate through the radar proposals $\mathbf{b}_j^r$ and a match is found if the following conditions are satisfied:

$$\hat{\mathcal{Y}}_i^c = \hat{\mathcal{Y}}_j^r \tag{4}$$

$$|| \left(\hat{u}_i^c, \hat{v}_i^c\right) - \left(\hat{u}_j^r, \hat{v}_j^r\right) ||_2 < T_p \tag{5}$$

$$|\hat{z}_i^c - \hat{z}_j^r| < T_d, \tag{6}$$

where T_p and T_d denote the distance threshold on pixels and the depth, respectively. We use different thresholds for projected centers and depth as the errors in the two spaces are different. Thus, we obtain a set of potential corresponding radar detections $\{\mathbf{b}_j^r\}$ for each $\mathbf{b}_i^c$. The complexity of matching is $\mathcal{O}\left(MN\right)$, where M and N are the number of camera and radar proposals. The score based thresholding limits the running time of the downstream matching algorithm.

Depth Weighting Network. The high-level idea of RADIANT is to update the monocular depth of the boxes with better depth from radar. Although the radar depth is generally more accurate than the camera depth, the camera depth may be better for nearby objects because of the richer semantic information. Thus, always preferring radar depth over camera is not beneficial. In other words, there should be a better *weighting* mechanism between the two depths. Hence, to better determine association and depth weights for a potential camera-radar detection pair extracted with Eqs. (4) to (6), we train another depth weighting network (DWN) to output relative confidence in radar and camera depths.

The DWN is a 4 layer multi-layer perceptron that outputs a classification score α between 0 and 1 where 1 indicates radar is more accurate and 0 if monocular is more accurate. Its input is a vector comprised of head output features, raw depths, distance, and Doppler/predicted velocity consistency. Details of the input vector are described in the supplementary material. The training labels to this network are binary. We assign the GT label for training as follows. If the GT depth z of a box is closer to the radar estimated depth, the GT label $\alpha = 1$. Otherwise it is zero. In other words,

$$\alpha = \begin{cases} 1, & |\hat{z}_j^r - z| < |\hat{z}_i^c - z| \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Fused Depth Calculation. Fusion of camera and radar depths occurs as follows. Assuming there are N radar associations $\mathbf{b}_j^r$ for $j \in set(i)$ potentially associated with a given camera detection $\mathbf{b}_i^c$. We run inference over this DWN for all these N pairs to obtain a sequence of confidence scores α_j. We then calculate the fused depth $\hat{z}_{\text{fuse}}$ from radar depths, weights α_j and the monocular depth as

$$\hat{z}_{\text{fuse}} = \begin{cases} \frac{\sum_j \alpha_j \hat{z}_j^r}{\sum_j \alpha_j}, & \text{if } \exists\ j,\ \alpha_j > T_\alpha \\ \hat{z}_i^c, & \text{if } \forall\ j,\ \alpha_j \leq T_\alpha \end{cases}, \tag{8}$$

where T_α denotes depth weighting threshold.

Method	≤ 10	$10-30$	≥ 30	All
Monocular Heads	0.563	1.442	6.042	3.415
Radar Heads	**0.413**	**0.649**	**1.017**	**0.791**
Raw Radar Depth	1.056	1.082	1.361	1.204

Table 1: Depth prediction error (in meters) on nuScenes validation subset using monocular heads and radar heads on image/radar pixels labeled as object over close, medium and long range objects.

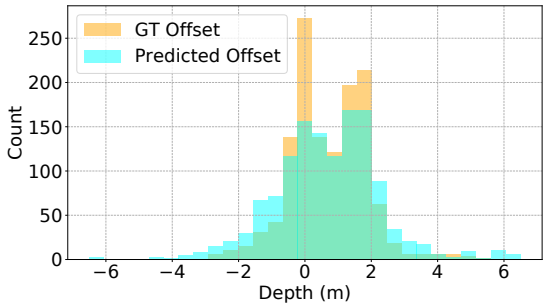


Figure 4: Histograms of predicted and GT depth offsets between radar returns and object centers. Both these distributions are in close agreement with each other.

Experiments

We apply the proposed method on the detection task of nuScenes dataset (Caesar et al. 2020), a widely used dataset with both image and radar points collected in urban driving environment. The nuScenes detection dataset consists of 28,130 training samples, 6,019 validation samples and 6,008 test samples. We experimentally show improvements in depth estimation accuracy and overall detection performance after enhancing monocular methods with the proposed strategy. The proposed method also achieves the SOTA performance in object detection via radar-camera fusion.

Depth Errors for Camera and Radar Heads

To show the advantage of the proposed radar head over monocular head in object depth estimation, We quantitatively compare the depth estimation accuracy of the radar head with the monocular head FCOS3D (Wang et al. 2021b) on random 900 images from the nuScenes validation set in Tab. 1.

We compute the mean absolute error for camera estimated depth, radar estimated depth and raw radar depth for monocular/radar pixels where GT depths are known. For a fair comparison, the depths are compared for objects having both positive camera and radar pixels as labels and the error for each object is averaged over all pixels associated with it and final error over all objects. In addition, we also show the

R	C	Method	mATE(↓)	AP(↑)										
				Mean	Car	Truck	Bus	Trailer	CV	Ped.	Motor.	Bicycle	TC	Barrier
	✓	MonoDIS-M (2020)	0.738	0.304	0.478	0.220	0.188	0.176	0.074	0.370	0.290	0.245	0.487	0.511
	✓	CenterNet (2019)	0.658	0.338	0.536	0.270	0.248	0.251	0.086	0.375	0.291	0.207	0.583	0.533
	✓	FCOS3D (2021b)	0.690	0.358	0.524	0.270	0.277	0.255	0.117	0.397	0.345	0.298	0.557	0.538
	✓	PGD (2021c)	0.646	0.360	0.547	0.268	0.253	0.243	0.087	0.422	0.379	0.300	0.584	0.525
✓	✓	CenterFusion (2021)	0.631	0.326	0.509	0.258	0.234	0.235	0.077	0.370	0.314	0.201	0.575	0.484
✓	✓	FCOS3D + RADIANT	0.622	0.374	0.582	0.301	0.257	**0.248**	**0.145**	0.439	0.386	0.302	0.579	0.500
✓	✓	PGD + RADIANT	**0.609**	**0.380**	**0.602**	**0.302**	**0.267**	0.242	0.107	**0.444**	**0.416**	**0.312**	**0.604**	**0.503**

Table 2: Performance comparison on nuScenes test set. R, C, CV, TC, Ped. and Motor. stand for radar, camera, construction vehicle, traffic cone, pedestrian and motorcycle, respectively.

R	C	Method	mATE(↓)	AP(↑)										
				Mean	Car	Truck	Bus	Trailer	CV	Ped.	Motor.	Bicycle	TC	Barrier
	✓	FCOS3D (2021b)	0.739	0.326	0.494	0.236	0.316	0.115	0.057	0.416	0.306	0.303	0.549	0.465
	✓	PGD (2021c)	0.658	0.368	0.546	0.290	0.378	0.148	0.063	0.441	0.374	0.343	0.595	0.504
✓	✓	CenterFusion (2021)	0.649	0.332	0.524	0.265	0.362	**0.154**	0.055	0.389	0.305	0.229	0.563	0.470
✓	✓	FCOS3D + RADIANT	0.653	0.363	0.587	0.291	0.371	0.120	**0.073**	0.447	0.364	0.333	0.581	0.467
✓	✓	PGD + RADIANT	**0.617**	**0.384**	**0.616**	**0.310**	**0.382**	0.141	0.068	**0.462**	**0.395**	**0.374**	**0.604**	**0.487**

Table 3: Performance comparison on nuScenes validation set.

error if we directly use radar depth as object depth without compensating with estimated residual depth. Tab. 1 shows that the radar heads achieve better depth estimation compared with camera heads, especially for the far objects.

We also plot the distribution of estimated and GT offset depths in Fig. 4. The estimated residual depth follows the GT distribution. It can be seen that, typically, the object center is a little farther than the measured depth of radar points. It demonstrates the usefulness of offset depth to compensate the error from direct radar measurement.

nuScenes Quantitative Results

The nuScenes (Caesar et al. 2020) leaderboard evaluates detection with metrics including mAP, mean average translation error (mATE), mean average size error, mean average orientation error and mean average velocity error. As this paper focuses on using radar to improve the monocular depth estimation of objects, mAP and mATE are the most relevant metrics and are reported. Other metrics are not reported since we did not update them with radar.

To show the effectiveness of the proposed camera-radar fusion on both test set (Tab. 2) and validation (Tab. 3), we compare the performance, *i.e.*, mATE and mean/classwise average precision (AP), of monocular methods, *i.e.*, FCOS3D (Wang et al. 2021b) and PGD (Wang et al. 2021c), before and after being fused with the proposed radar heads outputs and it shows significant improvements over mAP and mATE after combined with the proposed radar heads. Note for fairness, we compare monocular and corresponding RADIANT with the same monocular weights because the performance of RADIANT is partly determined by the performance of underlying monocular detection.

Next, we compare with CenterFusion (Nabati and Qi 2021), the best published radar-camera fusion method on nuScenes test set. Our method outperforms CenterFusion in both mAP and mATE, which indicates that our method acquires more correct detections and smaller localization error for those positive detections. For classwise results, RADIANT shows a gain of over 18% and 20% in AP over CenterFusion (Nabati and Qi 2021) on Cars and Pedestrians, respectively, in Tab. 2. This improvement is significant as cars and pedestrians are common participants in traffic, with Cars accounting for about 50% of total objects in nuScenes detection dataset.

nuScenes Qualitative Results

Fig. 5 shows the detection of the monocular detector FCOS3D (Wang et al. 2021b) (in magenta), detections with fusion from our proposed RADIANT (in cyan) and GT bounding boxes (in dashed orange) on image and BEV, respectively. In addition, we plot estimated position offsets for radar pixels with scores larger than 0.3. It is clear that the estimated residual depths are able to compensate the depth gap between radar measurements and actual object positions. As a result, the proposed RADIANT corrects the localization error of the FCOS3D (Wang et al. 2021b) detections and achieves accurate 3D position estimates leading to better 3D detection performance. The detections from monocular and fusion have the same orientations because only depths are updated during fusion. Specific examples show the effectiveness of depth correction in both near (Fig. 5 (b)) and long (Fig. 5 (h)) ranges.

Ablation Studies

Our proposed RADIANT model uses DWN to predict the confidence of radar depths for depth fusion and therefore carries out the *intelligent merging* of the depths. We therefore carry the ablation of this component on the nuScenes validation set in Tab. 4 on both the monocular methods FCOS3D (Wang et al. 2021b) and PGD (Wang et al. 2021a). In addition, we also consider an alternative strategy of averaging out the depth of camera box proposals and neighboring radar proposals which we call it as *Average Fusion* in the table. Tab. 4

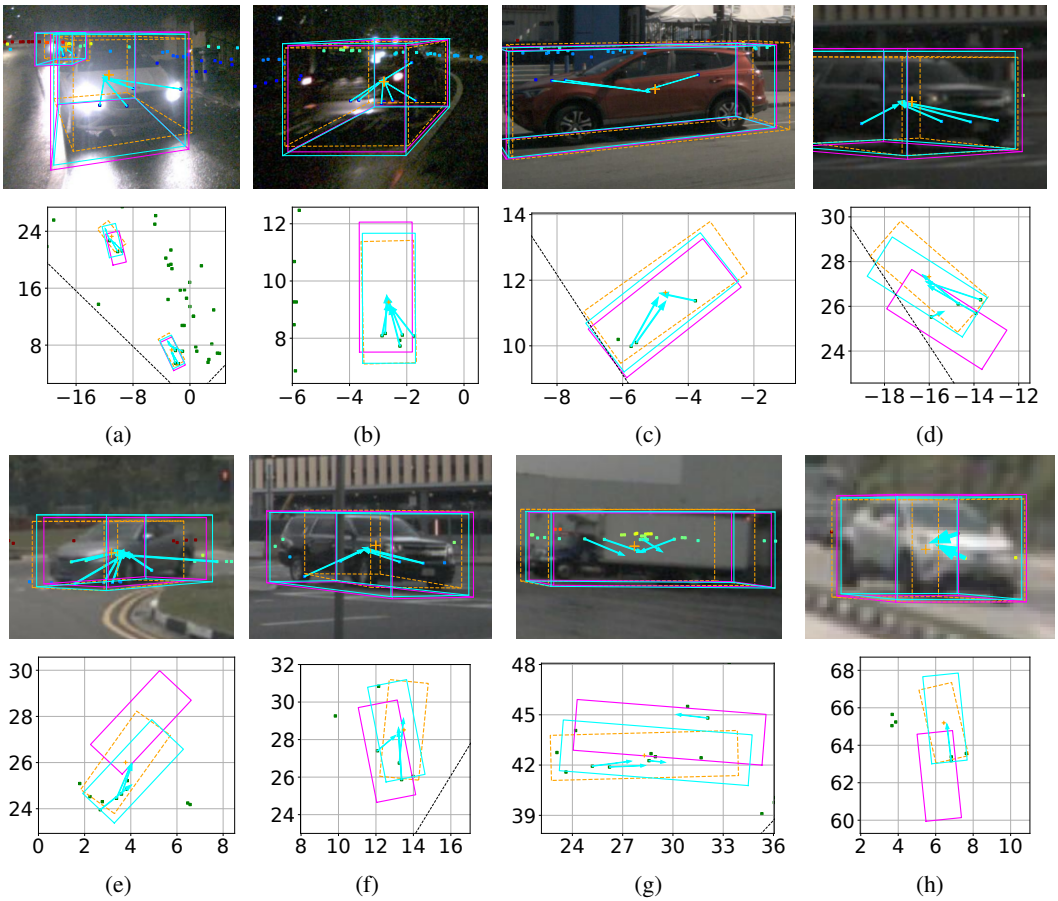


Figure 5: Qualitative Results. Visualization of predicted radar offsets (cyan arrows) to object centers and detections on image and BEV. The radar association of the RADIANT corrects the localization error of the FCOS3D (Wang et al. 2021b), improving detection performance. Orange, magenta and cyan boxes are GT bounding boxes, monocular detections, and RADIANT detections, respectively. The vertical axis in BEV indicates object distance in meters.

Monocular Method	Fusion	mATE (↓)	mAP (↑)
FCOS3D (2021b)	None	0.739	0.326
	Average	0.711	0.342
	DWN	**0.653**	**0.363**
PGD (2021c)	None	0.658	0.368
	Average	0.647	0.371
	DWN	**0.617**	**0.384**

Table 4: Ablation on fusion strategies, *i.e.*, average fusion and depth weighting network (DWN). The best results under the same monocular component are in bold.

results show that fusion methods outperform the monocular counterparts (non-fusion) methods. This is expected because the depth remains the hardest parameter to estimate for the monocular methods (Ma et al. 2021). More importantly, the fusion with DWN outperforms the Average fusion by a significant amount on both the metrics mATE and mAP on both the monocular methods. This suggests that DWN carries out the intelligent fusion of radar and camera depths instead of blindly averaging them, proving the effectiveness of DWN.

Conclusions

Fusing radar with camera-based detectors has proven challenging due in part to its low spatial resolution. Our network, RADIANT, provides a new way to fuse radar with 3D monocular image detectors. Using a radar branch in parallel to an image branch, RADIANT fuses both mid-level features and final detections. The mid-level features provide context to radar returns predicting the object centers offsets. RADIANT uses these offsets for association with the image detector, and to obtain accurate depth estimates for object detections, which are fused with the camera detectors. We show that the parallel branch fusion approach in RADIANT works with two monocular detectors FCOS3D and PGD. Finally, RADIANT achieves SOTA fused radar-camera detection on the nuScenes dataset.

Limitations. Although RADIANT improves depth estimation of monocular methods, it does not enhance other detection parameters such as object sizes and we leave velocity improvements to future work.

Acknowledgments

This work was supported by the Ford-MSU Alliance.

References

Barnes, D.; Gadd, M.; Murcutt, P.; Newman, P.; and Posner, I. 2020. The Oxford Radar RobotCar Dataset: A radar extension to the Oxford RobotCar Dataset. In *ICRA*.

Brazil, G.; and Liu, X. 2019a. M3D-RPN: Monocular 3D region proposal network for object detection. In *ICCV*.

Brazil, G.; and Liu, X. 2019b. Pedestrian detection with autoregressive network phases. In *CVPR*.

Brazil, G.; Pons-Moll, G.; Liu, X.; and Schiele, B. 2020. Kinematic 3D object detection in monocular video. In *ECCV*.

Caesar, H.; Bankiti, V.; Lang, A.; Vora, S.; Liong, V.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*.

Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; and Chateau, T. 2017. Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In *CVPR*.

Ding, M.; Huo, Y.; Yi, H.; Wang, Z.; Shi, J.; Lu, Z.; and Luo, P. 2020. Learning depth-guided convolutions for monocular 3D object detection. In *CVPR Workshops*.

Dong, X.; Wang, P.; Zhang, P.; and Liu, L. 2020. Probabilistic oriented object detection in automotive radar. In *CVPR Workshops*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Kumar, A.; Brazil, G.; Corona, E.; Parchami, A.; and Liu, X. 2022. DEVIANT: Depth EquiVarIAnt NeTwork for Monocular 3D Object Detection. In *ECCV*.

Kumar, A.; Brazil, G.; and Liu, X. 2021. GrooMeD-NMS: Grouped mathematically differentiable NMS for monocular 3D object detection. In *CVPR*.

Lee, W.; Jovanov, L.; and Philips, W. 2021. Semantic-guided radar-vision fusion for depth estimation and object detection. In *ICCV*.

Li, L.; and Xie, Y. 2020. A feature pyramid fusion detection algorithm based on radar and camera sensor. In *ICSP*.

Lien, J.; Gillian, N.; Karagozler, E.; Amihood, P.; Schwesig, C.; Olson, E.; Raja, H.; and Poupyrev, I. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *TOG*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*.

Liu, F.; and Liu, X. 2021. Voxel-based 3D detection and reconstruction of multiple objects from a single image. In *NeurIPS*.

Liu, L.; Lu, J.; Xu, C.; Tian, Q.; and Zhou, J. 2019. Deep fitting degree scoring network for monocular 3D object detection. In *CVPR*.

Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; and Zhang, L. 2021. AutoShape: Real-time shape-aware monocular 3D object detection. In *ICCV*.

Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; and Narayanan, P. 2021a. Full-velocity radar returns by radar-camera fusion. In *ICCV*.

Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; and Narayanan, P. 2021b. Radar-camera pixel depth association for depth completion. In *CVPR*.

Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; and Ouyang, W. 2021. Geometry uncertainty projection network for monocular 3D object detection. In *ICCV*.

Ma, X.; Liu, S.; Xia, Z.; Zhang, H.; Zeng, X.; and Ouyang, W. 2020. Rethinking Pseudo-LiDAR representation. In *ECCV*.

Ma, X.; Wang, Z.; Li, H.; Zhang, P.; Ouyang, W.; and Fan, X. 2019. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. In *ICCV*.

Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; and Ouyang, W. 2021. Delving into localization errors for monocular 3D object detection. In *CVPR*.

Major, B.; Fontijne, D.; Ansari, A.; Teja Sukhavasi, R.; Gowaikar, R.; Hamilton, M.; Lee, S.; Grzechnik, S.; and Subramanian, S. 2019. Vehicle detection with automotive radar using deep learning on range-azimuth-Doppler tensors. In *ICCV Workshops*.

Nabati, R.; Harris, L.; and Qi, H. 2021. CFTrack: Center-based radar and camera fusion for 3D multi-object tracking. In *Intelligent Vehicles Symposium Workshops*.

Nabati, R.; and Qi, H. 2019. RRPN: Radar region proposal network for object detection in autonomous vehicles. In *ICIP*.

Nabati, R.; and Qi, H. 2021. CenterFusion: Center-based radar and camera fusion for 3D object detection. In *WACV*.

Ouaknine, A.; Newson, A.; Rebut, J.; Tupin, F.; and Perez, P. 2021. CARRADA dataset: Camera and automotive radar with range-angle-Doppler annotations. In *ICPR*.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is Pseudo-LiDAR needed for monocular 3D object detection? In *ICCV*.

Rematas, K.; Kemelmacher-Shlizerman, I.; Curless, B.; and Seitz, S. 2018. Soccer on your tabletop. In *CVPR*.

Saxena, A.; Driemeyer, J.; and Ng, A. 2008. Robotic grasping of novel objects using vision. *IJRR*.

Shah, M.; Huang, Z.; Laddha, A.; Langford, M.; Barber, B.; Zhang, S.; Vallespi-Gonzalez, C.; and Urtasun, R. 2020. LiRaNet: End-to-end trajectory prediction using spatio-temporal radar fusion. *CoRL*.

Shi, S.; Wang, X.; and Li, H. 2019. PointRCNN: 3D object proposal generation and detection from point cloud. In *CVPR*.

Shuai, X.; Shen, Y.; Tang, Y.; Shi, S.; Ji, L.; and Xing, G. 2021. milliEye: A lightweight mmWave radar and camera fusion system for robust object detection. In *International Conference on Internet-of-Things Design and Implementation*.

Simonelli, A.; Bulò, S.; Porzi, L.; Antequera, M.; and Kontschieder, P. 2020. Disentangling monocular 3D object detection: From single to multi-class recognition. *TPAMI*.

Simonelli, A.; Bulò, S.; Porzi, L.; Kontschieder, P.; and Ricci, E. 2021. Are we missing confidence in Pseudo-LiDAR methods for monocular 3D object detection? In *ICCV*.

Tang, Y.; Dorn, S.; and Savani, C. 2020. Center3D: Center-based monocular 3D object detection with joint depth understanding. In *GCPR*.

Wang, L.; Zhang, L.; Zhu, Y.; Zhang, Z.; He, T.; Li, M.; and Xue, X. 2021a. Progressive coordinate transforms for monocular 3D object detection. In *NeurIPS*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021b. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *ICCV Workshops*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021c. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. 2019. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *CVPR*.

Wang, Y.; Guizilini, V.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2021d. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *CoRL*.

Wang, Y.; Jiang, Z.; Li, Y.; Hwang, J.-N.; Xing, G.; and Liu, H. 2021e. RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. *IEEE Journal of Selected Topics in Signal Processing*.

Wang, Y.; Mao, Q.; Zhu, H.; Zhang, Y.; Ji, J.; and Zhang, Y. 2021f. Multi-modal 3D object detection in autonomous driving: a survey. *arXiv preprint arXiv:2106.12735*.

Xu, Q.; Zhou, Y.; Wang, W.; Qi, C.; and Anguelov, D. 2021. SPG: Unsupervised domain adaptation for 3D object detection via semantic point generation. In *ICCV*.

Yadav, R.; Vierling, A.; and Berns, K. 2020. Radar+RGB fusion for robust object detection in autonomous vehicle. In *ICIP*.

Yang, B.; Guo, R.; Liang, M.; Casas, S.; and Urtasun, R. 2020. RadarNet: Exploiting Radar for robust perception of dynamic objects. In *ECCV*.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhou, Y.; He, Y.; Zhu, H.; Wang, C.; Li, H.; and Jiang, Q. 2021. MonoEF: Extrinsic parameter free monocular 3D object detection. *TPAMI*.