RADAR+RGB FUSION FOR ROBUST OBJECT DETECTION IN AUTONOMOUS VEHICLE

Ritu Yadav, Axel Vierling, Karsten Berns

Robotics Research Lab, Technische Universität Kaiserslautern, Germany

ABSTRACT
This paper presents two variations of architecture referred to as RANet and BIRANet. The proposed architecture aims to use radar signal data along with RGB camera images to form a robust detection network that works efficiently, even in variable lighting and weather conditions such as rain, dust, fog, and others. First, radar information is fused in the feature extractor network. Second, radar points are used to generate guided anchors. Third, a method is proposed to improve region proposal network [1] targets. BIRANet yields 72.3/75.3% average AP/AR on the NuScenes [2] dataset, which is better than the performance of our base network Faster-RCNN with Feature pyramid network (FFPN) [3]. RANet gives 69/71.9% average AP/AR on the same dataset, which is reasonably acceptable performance. Also, both BIRANet and RANet are evaluated to be robust towards the noise.


1. INTRODUCTION
With increasing demand and research in autonomous vehicles, object detection has become a significant task. RGB images through cameras are the standard input of such object detectors. However, bad weather conditions or dark environment conditions distort the images or videos captured by a camera in a driving scenario. Unpredictable environmental conditions make it troublesome for single modality based object detectors to detect the object’s location, and it is category precisely. To overcome this limitation, Vehicle industries are using different sensors in autonomous vehicles such as LIDAR, radar, infrared cameras, and others. Each sensor has its own benefits and drawbacks. However, LIDAR and radar sensors have most commonly come into the picture when analyzed on the usability factor. LIDAR sensors are highly efficient in making 3d images of objects but at the cost of the reliability issue. LIDAR has more moving parts hence have room for more noise. Radar sensors, on the other hand, provide very few points hence are not suitable for 3d object image construction. However, radar sensors have a high range, due to which they can detect objects at a long distance. They are robust, reliable, and cheaper than LIDAR sensors. Long-range object detection is beneficial for autonomous vehicles, especially commercial vehicles such as trucks, where they need more time to slow down for an incoming obstacle or object. Radar sensors and RGB cameras, make a very reliable combination for object detection. Even after these significant merits of radar sensors, there is limited significant research work conducted on radar sensor and vision fusion.

In this paper, an architecture for object detection is introduced, which utilizes RGB camera images and radar sensor data. Radar points store location information of objects and their distance from the sensor with high confidence. We use this high confidence radar sensor information for better feature extraction and region proposals along with other intuitive ideas. Proposed approaches are implemented to build a robust object detection network for the autonomous driving environment. For the demonstration of the effectiveness of the proposed architecture, the NuScenes dataset is used, which has recorded data from a full autonomous sensor suite (6 RGB cameras, 1 LIDAR, 5 RADAR, GPS, IMU).

2. RELATED WORK
In autonomous vehicles, radar-based object detection is explored in work such as [4, 5, 6, 7]. All these works are based on traditional techniques such as Haar-like features [8], Extended Kalman Filter (EKF) [4], occupancy grid map [5], and others. With the introduction of Convolutional Neural Network (CNN), an informative representation of objects is learned, and it shows drastic improvement in tasks such as classification and regression. CNN based object detectors are very efficient and common these days. Two-stage object detectors generally overpower single-stage detectors in terms of accuracy. The main strength of the two-stage object detector is the region proposal network. Faster-RCNN [11] introduced region proposal network (RPN) where they made network to learn anchor proposals. RPN was a tremendous success in the field of object detection. RPN is further explored in multiple works such as GA-RPN [9], Iterative RPN [10], Cascade RPN [11], and others. RPN is much explored on RGB camera images and very less on radar data. However, there are few works such as [12], where authors presented a 3D object detection network for car detection using AVOD architecture on radar pointclouds and camera images. Authors in [13], use range-doppler spec-
trums from FMCW radar signals and camera images for 3D object detection and estimation.

Authors in [14], presented radar points-based anchors where radar points are considered towards the center, left, right, and bottom of the object. However, the authors did not provide any specific reason for not considering the radar point at the top of the object. The idea of radar point-based anchor generation is accommodated in the presented work with some intuitive changes. In [15] and [16] use feature level fusion for the detection task.

3. PROPOSED OBJECT DETECTION NETWORK

The idea of sensor fusion is inspired by humans. We humans use multiple sensors to sense different attributes or specifications, which are refined together to come to a final observation. When it comes to artificial sensor fusion, the same approach is followed. In the proposed architecture, input from RGB cameras and radar sensors are fused, refined, and processed for object detection. Radar and RGB modalities are fused at feature and data level. Two variations of the fused networks are proposed in this work, namely RANet and BIRANet. Both follow the same architecture with two differences; First, RANet works only on radar point-based anchors. However, BIRANet also utilizes anchors generated similar to that in FPN. Second, BIRANet has a different RPN target generation method, "Best of Two," which is explained in subsection 3.4. Below is a detailed description of the proposed architecture.

3.1. Radar + RGB Feature Level Fusion

Radar sensors provide points, distance, velocity, azimuth, and others. Radar points are in the 3D coordinate system (x,y,z) where x and y coordinate of radar points provide the 2D location of reflective objects in the view, and the third dimension z corresponds to the distance of the object from the radar sensor. Before feeding the radar data into the network, radar data points are first converted from vehicle coordinates to camera-view coordinates, which will align radar points with the RGB image. Second, radar points are resized and then converted into a feature map of a size equivalent to that of the input image. At each radar point location x, y of the feature map, the z distance value is embedded and assign 0 to all other locations where there is no radar point present. The resulting radar feature map is the input of radar feature extractor branch, shown in in Fig 1 and has one channel.

![Fig 1: Radar Feature extractor branch.](image)

The radar and RGB feature extractor branch are fused in the feature extractor backbone. The feature extractor used for RGB images is ResNet [17]. While we experimented with radar and RGB data fusion at different stages of the feature extractor, the proposed configuration shown in Fig 2 proved to be the best. For fusion, element-wise addition operation is performed on the output feature map of the radar feature extractor branch and second stage output feature map of the RGB feature extractor branch.

3.2. Attentive FPN

Extracted feature maps, P2 to P5 are the features on which the performance of the detection network is primarily based. Hence these features have to be as good as possible. It is experimented and found that in the radar+RGB fused network, attention help in creating focus at radar point locations and boost the features in those areas. With this motivation, an Attentive FPN is proposed, which uses Concurrent Spatial and Channel ‘Squeeze & Excitation’(scSE) blocks [18]. The scSE block acts as attention and highlights important spatial features as well as significant channels. In the attentive FPN, scSE blocks are applied to output feature maps from feature extractor as shown in Fig 2. Feature maps with boosted good are then upsampled and added to a lower layer feature map to get output feature maps C2, C3, C4, C5.

3.3. Radar Point-Based Anchors and RoIAlign Layer

Radar signals are reliable in providing an object’s location even if they are far away hence have a long-range. When radar points are superimposed on images, although radar signals do not point to all object’s in the image but incorporate most of the objects and also those in long range. Therefore, it is reasonable to use radar points for generating region proposals. Another observation is that radar points are not only present towards the center of the surface of the object but also towards the top, bottom, left, and right edges of objects. Thus,
in a good proposal scheme, one cannot just consider the radar point at the center of the bounding box but should also consider the radar point on all four edges of the bounding box.

Anchor boxes are generated with 3 ratios (0.5, 1.0, 2.0) at each radar point, considering that each radar point is at the top, down, left, right edge, and at the center (TDLRC) of an anchor box, as shown in Figure 3. In RPN, Radar points are mapped to each output feature map of FPN. Anchors for each radar point are extracted with different scales based on the level of the feature map. In RANet, only radar point-based anchors are used. In BIRANet, radar point-based anchors along with all default anchors leading to data level fusion.

In the region proposal network of FFPN, RoIPooling [3] is used to obtain fixed-size (e.g., 7x7) feature maps from each region of interest. RoIPool cause harsh quantization over features which creates alignment issue between the input image and extracted features. For a good point based proposals, it is essential to have an accurate mapping of radar points in each feature map. For feature alignment reason, the RoIPool layer of FFPN is replaced with the RoIAlign [19] layer.

3.4. Best of Two - RPN Target Generation Method

In RPN target generation, anchors for each input image are compared to the ground truth bounding boxes, and the IOU score is calculated. Anchors, whose IOU score is greater than 0.7 are assigned with the positive label, and those with IOU score less than 0.3 are assigned with the negative label. Out of all anchor boxes, 128 positive and 128 negative labeled anchor boxes are selected for further process.

In cases where few ground truth bounding boxes of an image do not have any corresponding positive anchor box (i.e., Anchors with IOU > 0.7), the IOU scores of all corresponding anchor boxes are compared, and the one with maximum IOU score is assigned with a positive label. In such scenarios, radar point-based anchoring offers anchors with higher IOU scores provided there is a radar point corresponding to those ground truth objects. However, it is known that radar sensors do not provide reflected points for all objects in the scene; hence there are no anchors for them.

The Proposed “Best of two” method as outlined in Algorithm 1 help in overcoming this limitation where all anchors as per FFPN along with radar point-based anchors are utilized in RPN target generation. The benefits of this strategy are; first, objects which are not detected by using radar points can now be detected by using RGB anchors similar to FFPN. Second, this method increases the probability of having better anchor candidates for the RPN. The effect of this method can be seen in the performance comparison present in Table 1 and 2. We also want to highlight that the stated method causes a negligible overhead at training time and no increment in inference time.

Algorithm 1: Best of Two RPN target generation method.

1. Input: Ground-truth bounding box(gt_bbox), RGB anchors(i_anch), radar based anchors(r_anch).
2. Output: 128 positive target anchors.
3. Assign positive/negative labels to RGB anchors and radar based anchors separately.
4. for positive label RGB anchors(P_i_anch) do
5. for positive label radar anchors(P_r_anch) do
6. if IOU(P_r_anch, gt_bbox) > IOU(P_i_anch, gt_bbox) then
7. P_j_anch = P_r_anch.
8. Select random 128 positive anchors.

4. DATASET

For training and evaluation, we used the NuScenes dataset which is inspired by the KITTI dataset and covers a large variety of outdoor conditions. To our knowledge, it is the only publicly available dataset where radar data and camera images are synchronized. For object detection, NuScenes provide 3D bounding boxes with 23 classes.

Before using the NuScenes dataset, 3D bounding boxes are converted into 2D and merged relevant classes to obtain 6 classes, which are Car, Truck, Person, Motorcycle, Bicycle, and Bus. In the dataset, there are ground truth bounding box annotations that are not visible in the image due to full occlusion. Such cases are discarded by setting the visibility level parameter from NuScenes as two. Also, the dataset is split into 80:20% ratio for training and testing, respectively.

To investigate the robustness of networks, extra noise is added to evaluation images. Noise added are average blur noise with kernel size three and additive Gaussian noise with 0.05 standard deviation.

5. IMPLEMENTATION AND TRAINING

Conceptual base for the presented work is Faster R-CNN with feature pyramid network. Experiments are conducted on images of size 1024x1024 and 512x512. Scale used for anchor generation are (16, 32, 64, 128, 256) and (8, 16, 32, 64, 128) respectively. Maximum number of objects per image is set to 100. Regression and classification losses are combined with
Fig. 4: Detection sample images. Interesting part is focused on in the upper right. From left to right: Ground-Truth bounding box with Radar point superimposed on the Image; FFPN Prediction on RGB; BIRANet Prediction on RGB+Radar.

hyper parameters which are fine-tuned in the training process. Training is conducted on one TITAN Pascal GPU with batch size of two and the learning rate used is 0.001. Also, step-wise training is used for training both RANet, BIRANet.

<table>
<thead>
<tr>
<th>Method Used</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>AP85</th>
<th>AP(s)</th>
<th>AP(m)</th>
<th>AP(l)</th>
<th>AR</th>
<th>AR(s)</th>
<th>AR(m)</th>
<th>AR(l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFPN on RGB</td>
<td>69.7</td>
<td>88.2</td>
<td>82.0</td>
<td>60.9</td>
<td>50.3</td>
<td>68.0</td>
<td>73.1</td>
<td>73.0</td>
<td>53.2</td>
<td>71.3</td>
<td>76.4</td>
</tr>
<tr>
<td>RANet (Radar)</td>
<td>69.0</td>
<td>83.9</td>
<td>80.1</td>
<td>64.4</td>
<td>44.8</td>
<td>67.8</td>
<td>73.3</td>
<td>71.9</td>
<td>47.3</td>
<td>70.9</td>
<td>76.2</td>
</tr>
<tr>
<td>BIRANet (RGB+Radar)</td>
<td>72.3</td>
<td>88.9</td>
<td>84.3</td>
<td>65.7</td>
<td>53.5</td>
<td>70.1</td>
<td>76.9</td>
<td>75.3</td>
<td>56.2</td>
<td>73.2</td>
<td>79.8</td>
</tr>
</tbody>
</table>

Table 1: Comparison of detection results on 1024x1024 image size.

<table>
<thead>
<tr>
<th>Method Used</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>AP85</th>
<th>AP(s)</th>
<th>AP(m)</th>
<th>AP(l)</th>
<th>AR</th>
<th>AR(s)</th>
<th>AR(m)</th>
<th>AR(l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFPN on RGB</td>
<td>68.9</td>
<td>88.1</td>
<td>81.7</td>
<td>59.0</td>
<td>49.7</td>
<td>67.3</td>
<td>72.2</td>
<td>72.1</td>
<td>52.4</td>
<td>70.6</td>
<td>75.3</td>
</tr>
<tr>
<td>RANet (Radar)</td>
<td>68.3</td>
<td>83.3</td>
<td>79.1</td>
<td>62.6</td>
<td>45.0</td>
<td>66.9</td>
<td>73.0</td>
<td>71.3</td>
<td>47.1</td>
<td>70.0</td>
<td>75.9</td>
</tr>
<tr>
<td>BIRANet (RGB+Radar)</td>
<td>71.9</td>
<td>88.9</td>
<td>83.9</td>
<td>65.1</td>
<td>51.0</td>
<td>70.2</td>
<td>76.6</td>
<td>74.4</td>
<td>53.5</td>
<td>73.4</td>
<td>79.5</td>
</tr>
</tbody>
</table>

Table 2: Comparison of detection results on 512x512 image size.

For evaluation, COCO evaluation metric [20] is used with an additional AP over 0.85 IOU for a more detailed evaluation. In the presented result Tables 1, 2, better results of RANet and BIRANet under AP85 shows that the radar point based anchors help in getting better bounding boxes. The maximum lack in performance of RANet in comparison to FFPN is ∼0.7% in AP and ∼1.4% in AR, which is a reasonably good performance. Generally, we do not get radar points corresponding to each object in the scene due to the limited field of view, multiple surface reflection, and other reasons. So, it is logical to say that a network trained only on radar point-based anchor method won’t be able to detect objects for which there is no corresponding radar point. BIRANet took care of such cases and also achieved better detection in comparison to FFPN. Out of many, few such cases are shown in Fig 4, where BIRANet is able to detect more number of objects than FFPN.

On further noise addition, we see an expected decrease in performance. However, the drop in performance of BIRANet is less in comparison to that of FFPN. Both RANet and BIRANet show a minimal drop in performance, which proves their robustness. Also, in comparison to high-resolution images, detection on low-resolution images is ∼3% less, but this is along with the benefit of 55% less inference time. Hence, the proposed BIRANet is also an faster and accurate object detector for vehicles with low-resolution cameras.

Table 3: Performance comparison with existing methods.

As mentioned earlier, there are very few radar and RGB fused detection work available on NuScenes. A performance comparison with the existing work is presented in Table 3, where the results of proposed networks are compared on different image dimensions, as mentioned in their corresponding published work. From the presented comparison, it is clear that both proposed networks BIRANet and RANet perform significantly better among the mentioned methods.

Based on the stated findings, we can say that radar data help significantly in making more robust detection networks and also show performance improvement. Proposed radar and RGB camera image fusion methods are implemented on FFPN object detection network. We expect similar behavior on the state of the art object detection network.
7. REFERENCES


