Poster: mmBox: mmWave Bounding Box for Vehicle and Pedestrian Detection under Outdoor Environment

Zhuangzhuang Gu; Hem Regmi; Sanjib Sur
Computer Science and Engineering, University of South Carolina, Columbia, SC, USA
zg5@email.sc.edu; hregmi@email.sc.edu; sur@cse.sc.edu

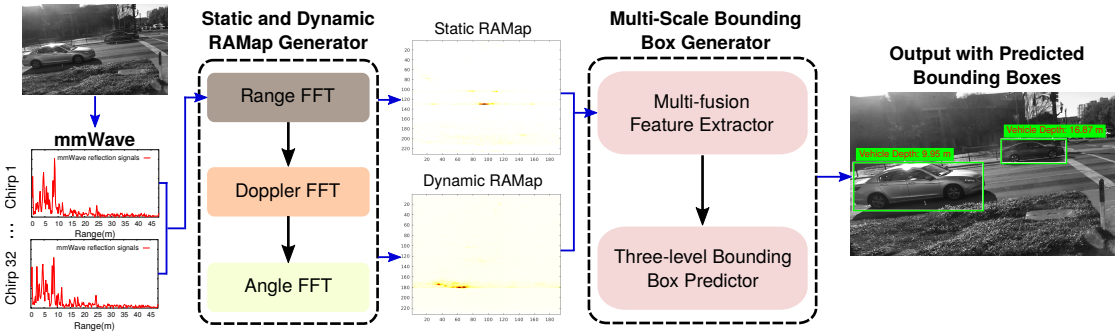

Fig. 1: *mmBox* predicts the bounding boxes from car and pedestrians from millimeter-wave reflected signals.

***Abstract*—Millimeter-wave technology's unique advantages, including low-light functionality, cost-effectiveness, and penetration of small objects, make it perfect for outdoor object detection. But traditional methods like likelihood clustering have faced challenges in determining target objects' extent and distance. This work presents *mmBox*, a two-stage system tailored for precise bounding boxes of vehicles and pedestrians outdoors. We assess *mmBox*'s effectiveness through extensive testing in outdoor street scenes using multiple metrics.**

***Index Terms*—Millimeter-Wave; Object Detection.**

I. Introduction

In recent years, the use of drones and autonomous driving has expanded across various sectors [1], [2]. As the demand for efficient navigation of drones and vehicles continues to rise, researchers have directed their attention toward the generation of bounding boxes for object detection. These boxes are essential for tasks like object tracking, instance segmentation, *etc.* [3]. [1] uses drones to detect cars from aerial images, while another introduces the MV3D deep learning model for 3D object detection in driving [2]. These accurate bounding boxes can help in data collection, traffic monitoring, and accident prevention. Although past studies have used cameras and LiDAR for this purpose [1], they falter in low-light conditions. LiDAR, being expensive and affected by outdoor interferences like fog, is not always the optimal choice.

Fortunately, millimeter-wave (mmWave) signals in 5G-and-beyond devices can penetrate minor obstacles, making them effective in challenging outdoor environments and diverse lighting conditions. Their high frequency and wide bandwidth allow for compact, high-resolution devices. The prevalence of mmWave in 5G-and-beyond devices also makes it a cost-effective solution. While some studies like [4] utilize Point Cloud Data (PCD) from high-resolution radar for 2D bounding box estimation, their scope is limited, testing only on one car. Both RODNet [5] and Radatron [6] employ deep learning models using mmWave signal heatmaps for outdoor detection. However, RODNet only predicts a likelihood cluster on the heatmap, and Radatron lacks distance and height details of objects.

We introduce *mmBox*, a two-tiered system that creates a Range-Azimuth Heatmap (RAMap) using reflected mmWave signals and subsequently predicts 2D bounding boxes for pedestrians and vehicles, with depth details (see Figure 1). Initially, the *Static and Dynamic RAMap Generator* captures both stationary and moving objects. The subsequent fusion model, the *Multi-Scale Bounding Box Generator*, extracts features from both RAMaps to predict bounding boxes and depths for various entities. By categorizing objects into dynamic and static RAMaps, *mmBox* assists the deep learning model in distinguishing overlapping entities, especially with increasing target objects. Additionally, *mmBox* offers three-tiered predictions using unique anchors at each tier to align with actual bounding boxes, effectively discerning features based on object scale and device depth. For validation, we utilized a standard mmWave device in outdoor settings like traffic intersections. Our preliminary results demonstrate the efficacy of *mmBox* in generating accurate bounding boxes with mmWave signals.

II. System Design

A. Static and Dynamic RAMap Generator

mmBox introduces a module to produce static and dynamic heatmaps from raw mmWave signals, effectively distinguishing stationary and moving objects. These heatmaps offer improved features compared to the sparse PCD method [4], presenting more cohesive clusters. Processing the raw reflections in *mmBox* involves 3 steps. *First*, *Range FFT* is applied to convert time domain signals to the frequency domain, capturing distance details. *Second*, *Doppler FFT* is applied on varying chirps in a frame to differentiate between stationary and moving entities. *Finally*, *Angle FFT* is applied on signals from non-overlapping virtual antennas to derive the azimuth angle from Range-Doppler data. From this, the static RAMap arises from stationary object reflections, and the dynamic RAMap from moving ones.

979-8-3503-0322-3/23/$31.00 ©2023 IEEE

2023 IEEE 31st International Conference on Network Protocols (ICNP) | 979-8-3503-0322-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICNP59255.2023.10355609

B. Multi-Scale Bounding Box Generator

Feature Extractor: *mmBox* predicts bounding boxes by fusing features from both static and dynamic RAMaps across multiple scales. Initial heatmaps are sliced into four lower-resolution images, maintaining and fusing distinctive features from both inputs. Features from both RAMaps are fused, passed through a Multi Layer Perception (MLP) to generate multi-level feature maps, and further refined using Dark Block and Spatial Pyramid Pooling (SPP) [7] to capture both global and local object details.

Three-level Bounding Box Predictor: Since the bounding box size of objects can be significantly varied with categories and distance, the small objects in feature map might be ignored with reducing size while large objects pretend to be easier for extracting global characteristics by convolution filter in reduced feature map. Therefore, we design this module to predict different scale objects. The predictor takes use of 3 feature maps from *Feature Extractor* and finally outputs 3 scale predictions. The small size predictions mainly focus on the large bounding box, while large-scale predictions consider large bounding boxes more. In particular, this module first upsamples small features, which include global object features and concatenates with large features, which capture more detailed local information. Then, it downsamples the combination of all features and concatenates with every level feature. Finally, MLP is applied to refine the features and accurately predict bounding box with depth.

Predefined Anchors: Since the variety of bounding boxes for different objects, directly predicting exact shape of vehicle and pedestrian will cost a long time to converge. *mmBox* proposes a predefined anchor-based prediction to speed up the training process. K-means is applied to find 3x3 predefined anchors from the height and width of ground truth. These 3x3 anchors are matched with 3-level prediction in 3 different sizes. Therefore, the width and height of generated bounding box can be calculated from the predicted offsets based on the predefined anchors.

Loss Function: Directly comparing the predictor's three-scale outputs with ground truth isn't effective due to distinct error impacts at different levels. *mmBox*'s approach maps ground truth boxes to three levels like predictions. This multi-level mapping aids in evaluating the loss at each level using several components: bounding box, confidence, classification, and depth loss. The EIOU metric [8], which considers the Intersection over Union (IOU), center points distance, and difference of height and width, is applied to measure the loss of bounding box. Classification, confidence, and depth loss use Binary Cross-entropy (BCE) to quantify the difference with the following loss function.

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \tag{1}$$

where b and b^{gt} denote the center point of predicted and ground truth boxes, w^c and h^c represent the width and height of the smallest enclosing box covering the two boxes, and w, h, w^{gt}, and h^{gt} are the width and height of predicted and ground truth boxes.

III. Preliminary Results

We employed a mmWave cascade device combined with a ZED stereo camera to capture reflections, gray-scale images, and depth images in outdoor street scenes. Our dataset comprises 10,440 training and 2,280 testing samples. An example of our predicted bounding boxes on a gray-scale image is depicted in Figure 1, where objects are accurately encapsulated in green boxes with depth values indicated. To evaluate *mmBox*'s performance, we computed various metrics, as shown in Table I. Standard metrics in object detection like Average Precision (AP) and Classification Accuracy (CA) are utilized. To test the accuracy of the bounding boxes, we introduce Average Center Distance (ACD), Average Height/Width Ratio (AWR/AHR), and Average Depth Difference (ADD) for assessing center alignment, size ratio, and depth prediction, respectively. Notably, *mmBox* showcases remarkable precision in object shape, position prediction, and depth estimation, underscoring its practical application potential.

TABLE I: *Performance analysis of mmBox.*

	CA	AP_{50}	ACD	AHR	AWR	ADD (m)
Vehicle	100%	42%	20 pix.	0.998	1.009	0.80
Pedestrian	100%	24%	11 pix.	0.995	1.007	0.51

IV. Conclusion and Future works

This work introduces *mmBox*, a system that processes mmWave reflections to produce dynamic and static RAMaps. These maps are then utilized by a deep learning model to precisely delineate bounding boxes for vehicles and pedestrians. While our present version predicts bounding boxes from a camera view, the generated heatmaps are from a top view. Future enhancements will focus on creating camera-view heatmaps and expanding mmBox to accommodate multiple inputs for improved accuracy.

Acknowledgments

We sincerely thank the reviewers for their comments. This work is partially supported by the NSF under grants CAREER-2144505, CNS-1910853, and MRI-2018966.

References

[1] B. B. et al., "Car Detection using UAV: Comparison between Faster R-CNN and YOLOv3," in *IEEE UVS*, 2019.

[2] X. Chen and et al., "Multi-View 3D Object Detection Network for Autonomous Driving," in *IEEE/CVF CVPR*, 2017.

[3] Z. Zou and et al., "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE*, 2023.

[4] A. Danzer and et al., "2D Car Detection in Radar Data with PointNets," in *IEEE ITSC*, 2019.

[5] Y. Wang and et al., "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*, vol. 15, no. 4, 2021.

[6] S. Madani and et al., "Radatron: Accurate Detection Using Multi-Resolution Cascaded MIMO Radar," in *Springer ECCV*, 2022.

[7] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[8] Y.-F. Zhang and et al., "Focal and Efficient IOU Loss for Accurate Bounding Box Regression," *Elsevier Neurocomputing*, vol. 506, 2022.