2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON) | 978-1-6654-8643-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SECON55815.2022.9918553

Pedestrian Liveness Detection Based on mmWave Radar and Camera Fusion

Hao Li[1], Ruofeng Liu[2], Shuai Wang[1,*] Wenchao Jiang[3], and Chris Xiaoxuan Lu[4]
[1]Southeast University, Nanjing, China [2]University of Minnesota, Minnesota, United States
[3]Singapore University of Technology and Design, Singapore [4]The University of Edinburgh, Edinburgh, United Kingdom
220194384@seu.edu.cn, liux4189@umn.edu, shuaiwang@seu.edu.cn, wenchao_jiang@sutd.edu.sg, xiaoxuan.lu@ed.ac.uk

***Abstract*—Autonomous driving requires vehicles to achieve fine detection of objects in the surrounding environment, especially living pedestrians. Nevertheless, in real world road environments there are living pedestrians and roadside portrait billboards. Existing vision-based object detection technologies fail to accurately distinguish living pedestrians from human figures. As an important sensor of autonomous driving system, mmWave radar has extra help to detect living pedestrians. In this paper, we extract the radar cross section (RCS) of the object from the low-cost mmWave radar signal as a distinguishing feature between living pedestrian and portrait billboard. Based on this observation, we propose a feature fusion network of mmWave radar and computer vision based on attention mechanism, and detect living pedestrians from fusion features. We implement the design with commodity mmWave radar IWR6843ISK-ODS and RGB camera Logitech Pro C920. The evaluation results show that our method effectively detects living pedestrians with an mAP of 97.7% and outperforms existing studies.**

I. Introduction

Achieving full autonomous driving requires vehicles to have fine-grained perception capabilities while being robust in the complex environment. Pedestrian detection is especially important among perception tasks. Failing to identify pedestrians could lead to severe traffic accidents (e.g., a pedestrian being hit and killed in Uber's autonomous driving test [1]). On the other hand, misidentifying non-pedestrian objects as a pedestrian could also cause non-trivial issues. Specifically, in addition to living pedestrian targets, roadside advertisements and car body advertisements with portraits widely exist in the road traffic environment. These false pedestrian targets will interfere with the autonomous driving system, and thus misguide vehicles to adopt wrong control strategies such as unnecessary emergency braking, threatening road traffic safety and reducing the traffic efficiency (e.g., create traffic jam). Tesla, for example, was reported to misidentify human figures in car body advertisements and give misleading warning to the driver [2]. Therefore, it is of great significance for the sensors of a vehicle to reject these false positives and accurately distinguish real living pedestrians from these interference.

In commercial vehicles, RGB camera and mmWave radar are most widely equipped sensors for object detection [3, 22]. Camera provides rich visual information, allowing accurate localization of the objects in the image. However, using visual features to perform classification or segmentation are prone to errors when subjects are visually similar. As a result, it is challenging for a camera to distinguish living pedestrians from portrait billboards or portraits printed on car body, especially when the subject is in distance and shows up with only a few pixels in the image. In contrast, a radar sensor uses mmWave RF signal for detection and thus the result is not affected by visual interference (e.g., it works in harsh light condition, fog and rain). In addition, it has a superior capability of detecting subjects in a long distance. Yet, radar point clouds are sparse, noisy and have a significant lower angular resolution than camera images. As a result, it cannot capture the appearance of subjects as the clue for classification.

Motivated by the need for a robust pedestrian detection and complementary natures of camera and radar, this work proposes the first pedestrian liveness detection design through the fusion of mmWave radar and camera sensor. The key difference of our proposal from the previous mmWave and camera fusion designs [3, 14] is that we exploit Radar Cross Section (RCS), i.e., the object's ability to reflect signal as the key feature for classification rather than the absolute position or signal strength utilized in previous designs. Through a measurement study on the commodity radar device, we observe that the RCS value of a real living pedestrian is dramatically different from these interfering objects (e.g., portrait billboards) due to the distinctions in shape, material and reflection characteristics of the skin. In addition, compared to absolute signal strength which suffers from ambiguity, the RCS of a living pedestrian is consistent over a long range and across various angles, making it superior for classifying objects in distance.

Based on the above observation, we further propose a novel multi-modal neural network design that fuses the radar RCS with visual clues from camera for pedestrian liveness detection. More specifically, we extend YOLOv3, the state-of-the-art image-based object detector to support multi-modal inputs and feature extraction, and dynamically integrate the RCS features with visual features through attention mechanism.

To summarize, our work makes the following contributions:

- We present the first pedestrian liveness detection design using mmWave radar and camera, enhancing the robustness of autonomous driving perception algorithm against visual interference.
- We propose a novel design that exploits the mmWave radar RCS feature of targets and a new neural network

*Shuai Wang is the corresponding author.

for radar and image feature fusion.

- We implement our design with commodity mmWave radar (e.g. IWR6843ISK-ODS) and RGB camera (e.g. Logitech Pro C920). We evaluate the system in four different scenarios and the results demonstrate that our design is highly accurate in pedestrian liveness detection, achieving mean Average Precision (mAP) of 97.7%.

II. MOTIVATION

A. The Need for Pedestrian Liveness Detection

To make the right control strategy and ensure safety, autonomous driving system needs to accurately detect pedestrians in the surrounding environment. In the real world road environment, there are not only living pedestrians, but also many interference targets, such as portraits in roadside billboards and car body advertisements. These interference may cause autonomous driving system to execute incorrect commands and cause serious accidents. For example, a normally moving autonomous driving car brakes abruptly after detecting portraits in a roadside billboard as living pedestrians and may be rear-ended by a car behind it. On the other hand, failing to detect living pedestrian results in serious issues. For instance, it is reported that in 2018 Uber's autonomous driving test car causes collision after detecting pedestrian crossing the road as non-pedestrian target [1].

This scenario motivates us to research on a reliable pedestrians liveness detection method. Specifically, when there is living pedestrians and interference from portraits in front of the autonomous driving system, our design aims at accurately detecting living pedestrians and excluding visual interference (e.g., billboard portraits).

B. Limitation of Existing Solutions

There has been a lot of impressive research in the field of object detection using mmWave radar and camera, and a number of mmWave radar datasets have been published, such as CRUW [4], CARRADA [5], nuScenes [6] and RadarScenes [7]. These related studies mainly focus on general object detection and commonly use the location or intensity of the object as the feature. Although effective, these methods are not sufficient to distinguish between living pedestrians and visual interference (e.g., portrait billboards). Specifically, the position of a object doesn't indicate its type. In addition, pedestrians and visual interference might have the similar absolute intensity. In the work, we propose new design to make the pedestrian detector robust to visual interference.

III. BACKGROUND

This section introduces the primer of mmWave radar and attention mechanism needed in this work.

A. Principles of mmWave Radar

The single-chip mmWave radar is based on the principles of frequency modulated continuous wave (FMCW) and has the ability to measure the range, relative radial speed and angle of the target. Specifically, the FMCW radar repeatedly transmits continuous chirp signals for a short period time which frequency increases linearly with time. When receiving the signal reflected by an object, the radar sensor produces Intermediate Frequency (IF) signal, which is analyzed to obtain three-dimensional position of the object.

Range Measurement. Based on the IF signal, the distance d between the object and the radar can be calculated as:

$$d = \frac{f_{IF}\, c T_c}{2\, B} \tag{1}$$

Here c is the speed of light, f_{IF} is the frequency of IF signal, B is the bandwidth swept by chirp, and T_c is the duration of chirp. To measure the range of multiple objects at different ranges, a fast Fourier transform (FFT) [8] is performed on the IF signal (i.e., range-FFT). The result of range-FFT represents the frequency response at different ranges.

Angle of Arrival Estimation. To depict the exact positions of objects in a spatial Cartesian coordinate system, the angle estimation is indispensable. The mmWave radar uses a linear antenna array to estimates the object angle. After emitting chirps with the same initial phase, RF Front-end simultaneously samples from multiple receiver antennas. Because the phases of the received signals are different between receiver antennas, the angle of the reflected signal can be estimated. Formally, the AoA can be calculated as:

$$\theta = \arcsin \frac{\lambda\, \omega}{2\, \pi\, l} \tag{2}$$

Here ω denotes the phase difference, l represents the distance between consecutive antennas and λ is the wavelength. Once obtaining the range and AoA (θ) of the targets, we get the exact positions of objects in a spatial Cartesian coordinate system.

B. Review of Object Detection

Since deep convolutional networks learn robust high-dimensional feature representations in images, deep learning strategies are widely used in object detection. Object detection methods based on deep learning are divided into two categories: two-stage object detectors such as Faster-RCNN [9] and single-stage object detectors such as YOLOv3 [10]. Two-stage object detectors have an independent module to generate region proposals. The working process of such detectors is divided into two stages. In the first stage, these models find a certain number of object proposals in the images, and then classify and locate them in the second stage.

Faster-RCNN takes a Fully Convolutional Network (FCN) as a region proposal network (RPN), which accepts any input image and outputs a set of candidate windows. Each window has a score, which determines the possibility of an object. Different from two-stage object detectors, single-stage object detectors combine extraction and detection and directly obtain the results of object detection. Compared with two-stage object detectors, they have simpler design and better real-time performance. YOLOv3 reconstructs the object detection from the classification problem in two-stage object detection to a regression problem, directly taking image pixels as objects and their boundary box attributes to predict. Multi-scale training

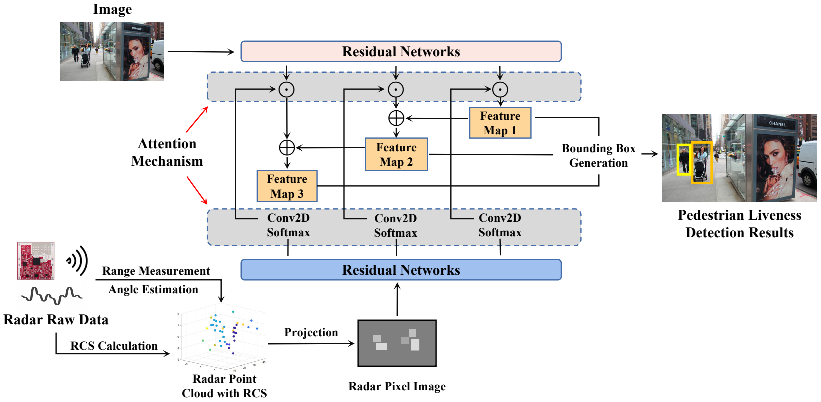


Fig. 1. System architecture.

makes YOLOv3 have good detection capability for small size objects. Although one-stage object detectors are basically used for object detection of single modal, their ideas of feature extraction, prediction result generation and multi-scale training can borrowed to our multi-modal object detection.

IV. System Overview

In this paper, we propose a mmWave and vision fusion pedestrian liveness detection system. Given the radar raw data and visual images of the road environment, the system detects living pedestrians in the road environment with visual interference using radar RCS features. Fig. 1 shows the overall architecture of our design. According to principle of mmWave radar (Section III) and proposed radar RCS calculation method (Section V), the radar raw data are processed to obtain radar point cloud with RCS. Then, radar point clouds are projected into radar pixel images and multi-scale features are extracted from radar pixel images and visual images respectively. We fuse multi-scale features of radar and vision based on attention mechanism and predict bounding box results from fusion features (Section VI).

V. Feasibility of Pedestrian Liveness Detection

This section discusses radar RCS, the key feature for our pedestrain liveness detection. We first explain the concept and how to obtain it on commodity radar. Then, a comprehensive measurement study is conducted.

A. Radar Cross Section

Radar cross section (RCS) is a measure of a target's ability to reflect radar signals in the direction of radar reception [11], which are determined by the size, shape, material of the target, incident/reflection angle of signal, etc. It is originally used in military radar technology to classify aircrafts and missiles. We notice that RCS has two unique characteristics that make it ideal for our pedestrian detection task. (i) First, the RCS of an objective in the far field of the radar is theoretically a constant value. This is distinct from the absolute intensity (or RSSI), which depends on the power of the transmitter, the gain of the receiver, the position of an object, etc. (ii) Second, human body is dramatically distinct from typical visual interference in the road environments (e.g., billboard and vehicle) in size, material and shape, which leads to highly diffrent radar reflectivity. Therefore, RCS can be a very discriminative feature to differentiate pedestrian from other interfering objects.

B. RCS Acquisition on Commodity Radar

Standard outputs from commodity radar (e.g., TI IWR series) does not provide the RCS value. Thus, we need to design a method to obtain it. Theoretically, RCS σ can be derived using the calculation formula:

$$\sigma = \frac{(4\pi)^3 d^4 kTFSNR}{P_t G_{TX} G_{RX} \lambda^2 T_{meas}} \tag{3}$$

where, k (Boltzmann constant), T (the antenna temperature), F (the noise coefficient of RX), and λ (mmWave wavelength) are constants and P_t (output power of radar), G_{TX} (TX Antenna Gain), G_{RX} (RX Antenna Gain), T_{meas} (the measurement time), d (the distance between target and mmWave radar) can also be calculated from the radar outputs and metadata. The key challenge is to obtain SNR, i.e., the ratio of the RX average signal intensity to the average noise intensity. Specifically, the noise is mainly background noise of radar circuit. Although the average noise intensity can be considered unchanged, it is difficult directly measure from the integrated radar device.

Our idea is that RCS should be proportional to RX signal intensity when other parameters are determined. Therefore, we use leverage a corner reflector [12], a square trihedral metal device with known RCS as a reference (RCS of a corner reflector be determined by its side length L as $\frac{12\pi L^4}{\lambda^2}$). Specifically, we collect RX signal intensity of the corner reflector across various distance d and build a benchmark database $\mathcal{B}(d)$. With $\mathcal{B}(d)$, the RCS σ_p of an object in spatial Cartesian coordinate (x, y, z) and with RX signal intensity P_{r_t} can be obtained as:

$$\sigma_p = \frac{P_{r_t}}{\mathcal{B}(\sqrt{x^2+y^2+z^2})}\sigma_r \tag{4}$$

Note that we only need the corner reflection in calibration stage. The system will use collected $\mathcal{B}(d)$ during operation.

C. RCS Value: Living Pedestrian vs. Visual Interference

We conduct a empirical measurement with an IWR6843 mmWave radar to verify the feasibility of using RCS for object classification (i.e., consistency in various distances and discrepency across different types of objects). As shown in Fig. 2, we set up two groups of target scenarios. The first scenario consists of a portrait billboard and two vehicles. The billoard is 0.8m × 0.7m and built with aluminum alloy which are the most commonly used material. mmWave radar collect raw data respectively and calculate the RCS of the targets based on the calibration of $\mathcal{B}(d)$.

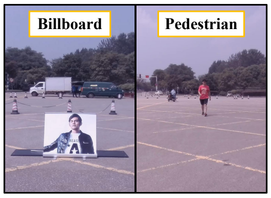


Fig. 2. Experimental scenarios of RCS value differences between living pedestrian and billboard.

The results shown in Table I demonstrates that RCS of living pedestrian target is dramatically different from billboard and vehicle. Average RCS of pedestrian is $3.5m^2$, whereas the value of billboard and vehicle are $230m^2$ and $500m^2$. Note that the unit of RCS is m^2 because RCS is formally defined as cross-sectional area of a perfectly reflecting sphere that can produce reflection of the the same strength.

TABLE I
EXPERIMENTAL SCENARIOS AND RCS VALUES OF LIVING PEDESTRIAN AND INTERFERENCE

Target Group	Target Object	RCS
1	Portrait billboard	$230m^2$
2	Vehicle	$500m^2$
3	Living pedestrian	$3.5m^2$

In Fig. 2 we further demonstrate the RCS of pedestrian and billboard when they are moving from 20 meters to 40 meters. Fig. 3 shows that the RCS of living pedestrian is very stable in the range of $2\sim5m^2$ and the RCS of portrait billboard is basically stable in the range of $210\sim250m^2$. These results confirm that the RCS obtained from commodity radar at different distances is generally consistent, which confirms the the theory that the RCS of targets in the far field of mmWave radar is their respective constant value.

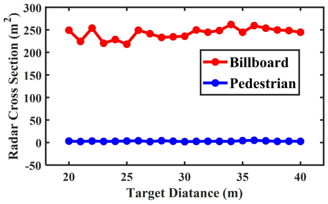


Fig. 3. Experiment result of RCS differences: living pedestrian vs. billboard.

D. Stability of RCS on Different Angles

The pedestrians are not always located directly in front of the radar, so the movements of them change the angle of arrival of the radar reflected signal. Therefore, we design experiment to verify the influence of the change of angle of arrival to the RCS value. As depicted in Fig. 4, we experiment with two pedestrian targets in the scenes. The first pedestrian walks away from the radar along the normal direction of the radar plane, while the second pedestrian walks along the trajectory that is 5 meters to the left of the horizontal direction of the mmWave radar.

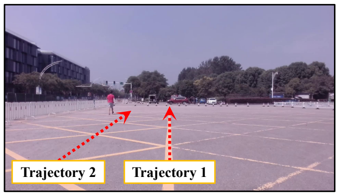


Fig. 4. Experimental scenarios of stability of RCS at different angles.

Fig. 5 shows the results of RCS of living pedestrian targets under two moving paths. Although varying angles of arrival of radar signal change the areas irradiated by mmWave radar, results of the two groups of RCS remain in the same fluctuation range, indicating that RCS of living pedestrian targets with different body angles have strong stability in the road environment.

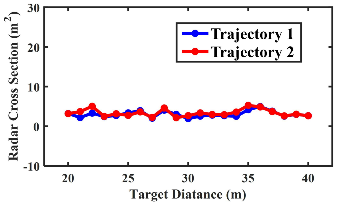


Fig. 5. Experiment result of stability of RCS at different angles.

E. Summary

The empirical measurement results prove that RCS is consistent over varying distances and angles while being distinguishable among pedestrians and other interfering objects in road environment. These observations motivate us to design a framework that incorporates radar RCS with image-based pedestrian detector for a robust detection.

VI. DESIGN

A. Design Overview

In this section, we develop a multi-modal pedestrian liveness detector that fuses radar RCS and images features to make decisions. A naive method to fuse images with radar RCS is to first use a two-stage image-based detector (e.g., Faster-RCNN) to obtain all the local living pedestrian candidates in the image, and then find the radar point clouds corresponding to the living pedestrian candidates and judge whether the candidate is living pedestrian according to the RCS value of the point clouds. However, our empirical experiment shows that the two-stage approach is not good at detecting distant pedestrians and only achieves mAP of 54.5% (more detailed in Section VIII). Thus, it does not satisfy the requirements of autonomous driving where pedestrians commonly appear in distance.

To reliably detect pedestrian across various distances, we got inspiration from recent multi-scale feature extraction design in the single-stage image-based object detector (e.g., YOLOv3). Specifically, features are extracted from different scales within an images and thus it can detect an object regardless of its size in the image. For example, our experiment shows that YOLOv3 has better detection accuracy (mAP of 82%) than Faster-RCNN (mAP of 51.4%) for small scale objects with a long distance. Therefore, in order to make our algorithm robust to visual interference while also being accurate in various distances, we design a multi-scale detection network with multi-modal feature fusion.

B. mmWave Radar Pixel Image

Our RCS calculation algorithm provides one RCS value for each voxel in 3D space. Thus, all RCS value form is a 3D heatmap. However, due to sparsity of radar point cloud, there are a lot of voxels that don't contain useful information for pedestrian detection. In order to extract multi-scale features from radar RCS more efficiently, we project the RCS from the Cartesian coordinate system of radar to the pixel coordinate system identical to the RGB camera by squeezing it on the depth direction (y-axis) as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{y} M_1 M_2 \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{5}$$

where M_1 is the 3×3 internal parameter matrix of RGB camera, M_2 is the 3×4 external parameter matrix of RGB camera, (u, v) is the pixel location in the RGB camera pixel coordinate system. Since the mmWave radar and RGB camera are fixed on autonomous vehicle, their relative positions remain static during the movement. Therefore, M_1 and M_2 can be calculated in advance.

By doing this, we construct a 3-channel radar pixel image with the same height and width as a RGB image, with the initial value of the pixel to be 0. The value of pixel is set according to the RCS value of the radar point that this radar pixel is projected from. Fig. 6 shows a RGB image and radar pixel image collected at the same time. Since the RCS value of the living pedestrian is lower than that of the billboard, the pixel blocks representing the living pedestrian are less bright than the pixel blocks representing the billboard.

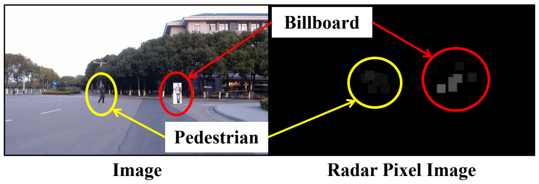


Fig. 6. Image and radar pixel image.

C. Multi-modal Feature Extraction

As we discuss in Section VI-A in autonomous driving scenarios, objects might be located at a long range of distances from the sensor. The scale of the object in the image and the pixel block distribution area in the radar pixel image will decrease with the increase of the distance. Moreover, due to limited resolution of radar, when there are multiple close objects, the boundary of their pixel block region may be unclear. Therefore, we develop multi-scale feature extraction for both RGB image and RCS pixel image that can obtain the high-dimensional features of objects in different scales. This is especially useful for maintaining accurate position prediction of small scale objects.

As the blue and pink block in Fig.1 depicts, we use two 5-layer residual networks to extract features from radar pixel image and visual image. Each layer of residual network is composed of multiple residual units. Each residual unit contains a 1×1 two-dimensional convolution and a 3×3 two-dimensional convolution. As the data complexity of radar pixel image is lower than that of visual image, the number of residual units in the residual network used to extract features of radar pixel image is less than that of visual image, which are $(1, 2, 4, 4, 4)$ and $(1, 2, 8, 8, 4)$ respectively. The features extracted from the last three layers, from shallow to deep, represent the feature of three different scales with sizes being 52×52, 26×26 and 13×13. In other word, each pixel in the deeper layer captures the features in a larger scale. These three features are used for the multi-modal feature fusion discussed in the next section.

D. Multi-modal Feature Fusion

Feature extraction module yields multi-scale feature maps for both the RGB image and radar pixel image. We then fuse the multi-modal features to detect living pedestrians. A classic operation of feature fusion is tensor concatenation. After concatenation, the multi-dimensional linear relationship between features is further extracted using multi-layer convolution operation. However, we find that since the features of radar pixel image and visual image come from different modalities, it is challenging for the convolution network to figure out their linear relationship. Our preliminary experiment also proves that the performance of tensor concatenation method has no obvious advantage compared with single-modal detector (e.g., YOLOv3) as detailed in Section VIII.

Meanwhile, we find that the radar pixel image and the visual image can be regarded as the observation of the same object at the same position and perspective but with different modalities (mmWave radar and visual modal respectively). Therefore, we need guide the network associate features of two modalities using their spatial relationship. Specifically, we use the features extracted radar pixel images to help neural networks to identify the area of interest for living pedestrians in the images. This process is similar to the mechanism of human brain ignoring irrelevant information and focusing on key information when processing information overload and the features of radar pixel images can be regarded as the weight matrix when observing images. Technically, it is achieved by fusing radar pixel image features and visual image features using the attention mechanism [13]. The feature fusion formula of attention mechanism is as follows:

$$\mathcal{F}_{i,j} = \varphi(\mathcal{V}_{i,j}) \odot \psi(\mathcal{R}_{i,j}) \tag{6}$$

where $\mathcal{F}_{i,j}$ is the value of the fused feature at the index (i, j), $\mathcal{V}$ is the visual image feature, $\mathcal{R}$ is the radar pixel image feature, φ stands for linear mapping operation, ψ stands for two-dimensional convolution and Softmax operation and $\odot$ stands for Hadamada product operation. φ and ψ are learnable parameters that obtained from training data.

The attention operations are operated on features of different scales. Furthermore, we borrow the idea of feature pyramid network and feed the features captured in the global scale to back to the ones of local scale to provide extra context. More specifically, the fusion feature of the layer 5 is up-sampled and then concatenated with the fusion feature of the layer 4 to generate the final fusion features of layer 4. The new fusion feature of the layer 4 is up-sampled and concatenated with the fusion feature of layer 3 to generate the final fusion feature of layer 3. Finally we generate pedestrian liveness prediction results from the fusion features of these three layers.

VII. Implementation and Data Collection

A. Data Collection

a) Data Collection Platform: For the radar and camera data collection, we design a mobile data collection platform with a commercial and off-the-shelf mmWave radar IWR6843 and a commercial high-definition RGB camera Logitech Pro C920 (as shown in Fig. 7) to simulate an on-board system. The radar operates in a frequency band from 60GHz to 64GHz whose wavelength is $\sim 4mm$. It has three transmitting antennas and four receiving antennas that form a 60 degree azimuth FoV and 60 degree elevation FoV whose angle resolution is $\sim 15°$. For reproduction, the detailed configuration parameters of the device are provided as follows: the device is set to transmit 64 chirps per frame. The start frequency of the chirp is set to 60GHz. The frequency bandwidth is set to 1009.82MHz. The Frequency slope is set to be 21.038MHz/us. The RGB camera records video at a resolution of 1920×1080 at 30fps.

b) Experiment Site: For experiments, we simulate the vehicle driving scenario and collect data from the vehicle lane on campus. Portrait billboard is placed at one end of the motorway, on the left or right side and living pedestrians walk near the portrait billboard. The data collection platform moves towards the billboard and pedestrians from the other end of the motorway, about 50 meters away. In the process of data collection platform movement, mmWave radar and camera record data simultaneously. We collect 13800 frames image data and 2760 frames radar data on four different motorways. The proportion of the dataset containing single, two, and three living pedestrian scenarios is 20%, 70% and 10% respectively and all the scenarios have a portrait billboard as interference. The ratio of training and validation data to testing data in our dataset is 9:1. Since billboard portraits are the same in all scenarios, to avoid over-fitting of the neural network during training, billboard portraits in the validation data and testing data are replaced with other portraits not present in the dataset, respectively, to achieve data augmentation. It took about 60 days to collect the data.

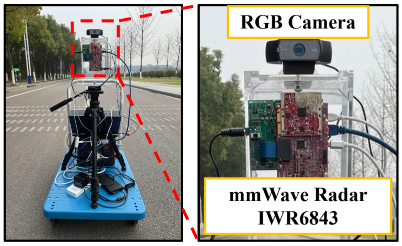


Fig. 7. Data collection platform.

B. Object Segmentation for Radar Data

The radar data was collected in outdoor environments that include static obstructions such as trees and ground, together with the portrait billboard and living pedestrians. In such an environment, however, multipath noise is non-negligible, which is a common issue for almost all RF technologies. Due to the reflection of ambient objects and beam spreading, the propagation of mmWave signals between objects and transceivers tends to travel through multiple paths. Consequently, unwanted points often appear in the radar point cloud which are widely known as the ghost points. In order to mitigate the impact of these noisy points and segment portrait billboard and living pedestrians out, we implement clustering based point segmentation.

We apply the DBScan algorithm to acquire the cluster of points of billboard and pedestrian such that the noise can be suppressed. DBScan is a density-aware clustering algorithm that can divide a point cloud based on the distance and the density described based on a set of neighborhoods in the 3D space. As it does not require the number of clusters to be specified a priori and can automatically mark outliers that are noise, DBScan has been used to separate individual objects from mmWave radar point clouds. Our implementation separates the radar points into different clusters and selects effective objects according to the number of points in cluster. Regarding the hyperparameters settings of DBScan, we empirically set the maximum distance (radius) between two points falling into the same cluster to 1 and set the minimum point number in a cluster to 3.

VIII. Evaluation

A. Evaluation Methodology

a) Evaluation Metrics: Precision, Recall, F1 score and mAP are main evaluation metrics for object detection tasks.

Precision is the percentage of truly positive samples that are predicted to be positive. Recall refers to the percentage of positive samples that are correctly predicted. The true or false attribute of the sample bounding box is determined by its intersection-over-union (IoU) with the ground truth box, which is the ratio of intersection and union between sample bounding box and ground truth box. For instance, when the

IoU threshold is 0.3, if the IoU between sample bounding box and ground truth box is greater than 0.3, the sample is considered to be true. F1 score is a metric that measures the accuracy of binary classification model, which takes into account both the precision and recall of classification model. It is regarded as a harmonic mean of model precision and recall.

mAP is the mean of each class AP. AP is the area under Precision and Recall curves generated by confidence threshold changes, and represents the overall performance of the detection method under different confidence threshold. To calculate the AP, we first use the trained model to obtain the confidence score of all bounding boxes and rank them according to the confidence score. Then, we select the top-1 to the top-n results from the ranked boxes to calculate the precision and recall corresponding to the number of boxes respectively. All pairs of precision and recall are combined to form a P-R curve and the AP is the area under the P-R curve.

b) Competing Approaches:

- **Faster-RCNN and Radar RCS (Faster+RCS):** This is the fusion method of two-stage detector and radar point clouds with RCS information mentioned in Section VI. We apply this method to our dataset to calculate its detection performance for living pedestrians.
- **w/o attention mechanism (No-Attention):** Our proposed feature fusion method based on attention mechanism is replaced by tensor concatenation fusion. This setting aims to examines the importance of attention mechanism to the performance.
- **w/o RCS (No-RCS):** To examines the effectiveness of RCS in pedestrian liveness detection in our proposed method, RCS features in radar data is replaced with RX signal intensity, and a new radar pixel image dataset is made to calculate the pedestrian liveness detection performance of our proposed method.
- **milliEye [14]:** MilliEye is a method proposed by Xian et al. in 2021 to achieve object detection in dark light environment by using the fusion of camera and mmWave radar. MilliEye takes advantage of mmWave radar's ability to be unaffected by light and combines 3d radar point clouds with visual images to detect objects under dark light conditions. It can basically be regarded as a fusion method of two-stage detector and radar point clouds without RCS information. We apply milliEye to our dataset to calculate its detection performance for living pedestrians.

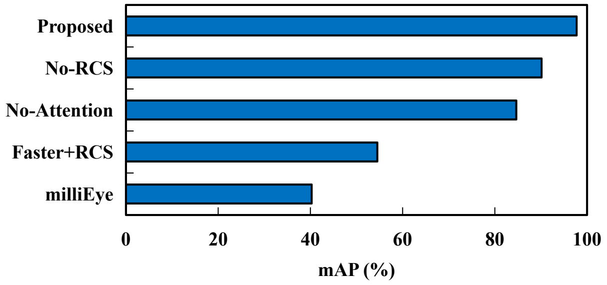


Fig. 8. Overall performance.

B. Overall Performance

As Fig. 8 depicts, our method achieves an mAP of 97.7% in the pedestrian liveness detection, the best performance of all methods. **No-RCS** achieves an mAP of 90.1% which is 7.6% lower then our proposed. This demonstrates that radar RCS is superior to absolute intensity in help distinguishing real living pedestrians from visual interference. **No-Attention** achieves an mAP of 84.6% which is 13.1% lower then our proposed. This proves that the attention mechanism can promote the efficient fusion of pixel image and image features. **Faster+RCS** achieves an mAP of 54.5% which is 43.2% lower then our proposed and only 3.1% higher than Faster-RCNN and **milliEye** achieves an mAP of 40.2% which is 57.5% lower then our proposed. These results demonstrates the the weakness of two-stage approaches for object detection in autonomous driving scenarios (e.g., inaccurate in detecting distant pedestrian with small scale in the image). Two examples of pedestrian liveness detection by our method are shown in Fig.9. Red dots represent point cloud with high RCS values of billboard and yellow dots represent point cloud with low RCS values of living pedestrians.

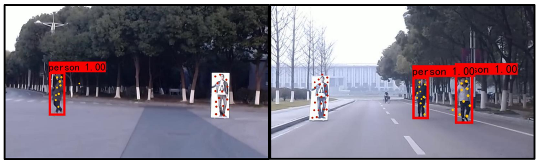


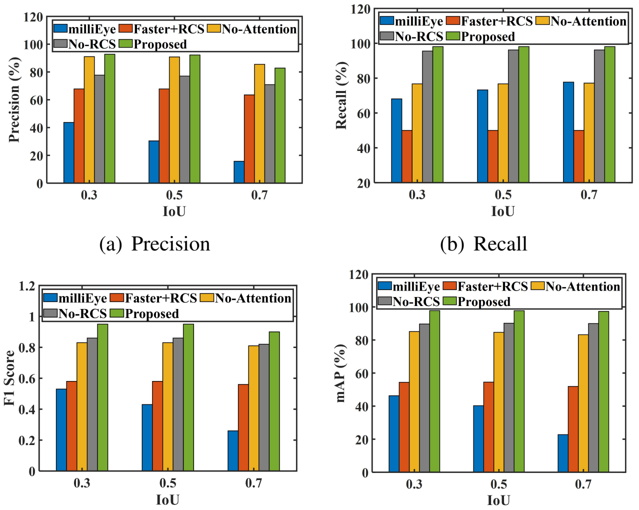
Fig. 9. Examples of pedestrian liveness detection.

C. Sensitivity Analysis

a) Precision, Recall, and F1 Score: We break down the overall accuracy and compare precision, recall, and F1 score of our method with baselines. The results in Fig. 10, our method outperforms baselines in all the metrics. As can be seen from Fig. 10(a), due to the replacement of RCS with intensity, **NO-RCS** loses the unambiguous metrics for distinguishing living pedestrians and thus is prone to false positive errors, resulting in a Precision reduction of more than 13% compared with **Proposed**. Furthermore, Fig. 10(b) shows that due to the lack of multi-scale feature extraction, **Faster+RCS** is prone to miss detection, which leads to a relatively low Recall. Finally, as shown in Fig. 10(c), **milliEye**, due to the lack of both radar RCS features and multi-scale method, results in a significant decline in comprehensive performance.

b) Impact of IoU threshold: In our evaluation, we calculate the IoU of predicted bounding boxes and ground truth boxes. The pedestrian is detected when is higher than the predefined threshold. Thus, we set IoU thresholds to various values to observe its impact on evaluation results.

When the IoU threshold increases, bounding boxes of small scale objects are more likely to be lost. Fig. 10(d) shows that mAP of **Proposed** is always above 97% and remains optimal, which proves its robustness. The performance of the method based on two-stage detector is lower than that of the method based on one-stage multi-scale detector, which proves that the multi-scale method can detect small scale objects in long-distance scenes.



(a) Precision (b) Recall

(c) F1 Score (d) mAP

Fig. 10. Performance of proposed and baselines in different IoU thresholds

c) Impact of object distance: The distance between the object and the experimental platform affects the scale of the object in the image and the radar pixel image. In our dataset, the proportions of data from different distances are basically the same. Fig. 11 shows that the mAP metric of our method is basically stable at 97% on different distance data, regardless of the number of live pedestrians in the experimental scenarios, indicating the robustness of our method for pedestrian liveness detection at different scales.

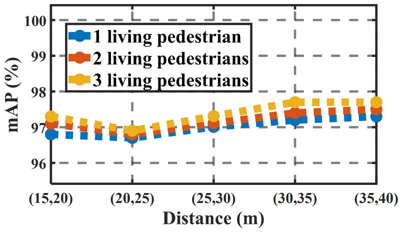


Fig. 11. Performance with different distance and number of living pedestrians.

d) Impact of the number of living pedestrians: The number of living pedestrians in the experimental scene affects the evaluation metrics of our method. Our dataset is segmented by the number of living pedestrians to separately evaluate the detection performance of our method. Fig. 11 shows that with the increase of the number of living pedestrians, the overall mAP of our method remains above 97% (actually rises slightly because multiple pedestrians in long distance are more detectable by radar) over the same object distance range, indicating that our method is robust in multi-object scenes.

e) Impact of different road scenes: In different experimental road scenes, the differences of road condition and the distribution differences of trees and road lamps on both sides of the road together constitute different radar electromagnetic environments, which affect the collection of radar raw data. The detailed road environment is shown in Table II.

We conduct experiments in four different road scenes with different numbers of living pedestrians. The results show that the change of road scenes has little effect on proposed performance. In these four scenes, the mAP metrics of proposed are 97.4%, 97.3%, 97.1%, and 97.7%, respectively. The evaluation validates that the RCS calculation method of our design is resilient to environmental heterogeneity. The reason lies in that the ambient noise is not enough to affect the radar SNR and our preprocessing algorithm can remove the static reflection points in different environments.

TABLE II
DIFFERENCES OF ROAD SCENES

Scene	Condition	Location of Trees	Lamps	Pedestrians
1	Rough	End of road	Not exist	1
2	Rough	Not exist	Not exist	1 & 2
3	Flat	Both sides of road	Exist	1 & 2
4	Flat	End of road	Exist	2 & 3

f) Impact of sunlight condition: To evaluate the generalization performance of the model, we conduct experiments in four scenes of different sunlight conditions as shown in Table III. In scene 1&2, the camera is interfered by direct or reflected sunlight from objects making it harder than normal to detect objects. In scene 4, object image blur caused by low light is also not conducive to object detection. The results show that the change of sunlight condition has little effect on proposed performance. Our method achieves mAP of over 97% in all four scenes. The verification confirms that our method makes full use of the characteristics of mmWave radar insensitive to environmental weather, which is helpful for the practical application of our method.

TABLE III
DIFFERENCES OF SUNLIGHT CONDITIONS

Scene	Experimental Platform	Objects
1	Face the sun, strong light	Back to the sun, shadow
2	Back to the sun, shadow	Face the sun, strong light
3	Cloudy, ordinary light brightness	
4	Nightfall, low light brightness	

IX. RELATED WORK

A. Single-modal Object Detection

Both visual-based and radar-based detection techniques have been separately studied in the literature. In computer vision, human contour features are usually used to detect pedestrians from RGB images [9, 10]. These methods do not consider liveness detection. Liveness detection is achieved through facial contour features in the images [15]. However, the method only when there is a close face in the image and thus doesn't work for autonomous driving scenarios.

On the other hand, mmWave radar is proposed as an environmentally insensitive detection method in autonomous driving. In [16], radar Doppler spectrum is analyzed as the motion characteristics to distinguish vehicles and pedestrians. Radar heatmap is input into neural networks in [17, 18] for the classification of different types of objects on road environment. Radar point cloud is used for road object classification and vehicle detection in [19, 20]. However, the sparse and noisy characteristics of radar limits their performance.

In contrast to these single-modal detector, our work combine image features with the radar RCS to improve the accuracy and robustness of pedestrian detection.

B. Multi-modal Object Detection

Multi-modal object detection has been studied by the autonomous driving community based on on-board sensors for years, including mmWave radar and camera fusion and LiDAR and camera fusion. A large number of studies fuse mmWave radar point cloud [21–24] or Doppler heatmap [25, 26] with visual images to achieve object detection, but as our evaluation shows, point cloud and heatmap are not sufficient to achieve pedestrian liveness detection. Although object detection based on LiDAR and visual image fusion has been studied [27–32], pedestrian living detection remains largely unexplored, with 3DCNN [33] being the only work in this field. However, 3DCNN is a technology based on LiDAR point clouds, making it difficult to distinguish real pedestrians from stereo humanoid object interference. By studying the emerging low-cost mmWave radars and RGB cameras, our proposed method utilizes the RCS characteristics of mmWave radars to achieve pedestrian liveness detection even in the presence of stereo humanoid object interference due to significant differences in size, material and shape.

X. Conclusion

This paper presents a novel system design that detects living pedestrians with data captured by mmWave radar and RGB camera. To make pedestrian liveness detection feasible, we introduced radar RCS as the key feature of our approach. In order to ensure the detection ability of long-distance small-scale objects, we construct a global multi-scale feature extraction network. Our system also propose a multi-modal feature fusion method based on attention mechanism to efficiently fuse multi-modal features. We believe our system demonstrates the potential of multi-modal pedestrian liveness detection and envision it to serve as a key solution for precise perception in the autonomous driving.

Acknowledgment

This work was supported by China National Key R&D Program 2018YFB2100302, National Natural Science Foundation of China under Grant No. 61902066 and Natural Science Foundation of Jiangsu Province under Grant No. BK20190336 and the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20221-0017), National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme (FCP-SUTD-RG-2021-009).

References

[1] E. Sacks, "Self-driving Uber car involved in fatal accident in Arizona," NBC News. Available: https://www.nbcnews.com/tech/innovation/self-driving-uber-car-involved-fatal-accident-arizona-n857941. March 20, 2018.

[2] Tesla misidentified the poster as a pedestrian. [Online]. Available: https://xw.qq.com/cmsid/20220309V063U300. March 9, 2022.

[3] D. Kang and D. Kum, "Camera and Radar Sensor Fusion for Robust Vehicle Localization via Vehicle Part Localization," in *IEEE Access*, vol. 8, pp. 75223-75236, 2020.

[4] Y. Wang, Z. Jiang, Y. Li, J. -N. Hwang, G. Xing and H. Liu, "RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 954-967, June 2021.

[5] A. Ouaknine, A. Newson, J. Rebut, F. Tupin and P. Pérez, "CARRADA Dataset: Camera and Automotive Radar with Range- Angle- Doppler Annotations," *2020 25th ICPR*, 2021, pp. 5068-5075.

[6] H. Caesar et al., "nuScenes: A Multimodal Dataset for Autonomous Driving," *2020 IEEE/CVF CVPR*, 2020, pp. 11618-11628.

[7] O. Schumann et al., "RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications," *IEEE 24th International Conference on Information Fusion*, 2021, pp. 1-8.

[8] K. Sun and Z. Yin and W. Chen and S. Wang and Z. Zhang and T. He, "Partial Symbol Recovery for Interference Resilience in Low-Power Wide Area Networks," *IEEE 29th International Conference on Network Protocols*, 2021, pp. 1-11.

[9] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement." *arXiv preprint* arXiv:1804.02767, 2018.

[11] M. A. Richards, *Fundamentals of radar signal processing.* McGraw-Hill Education, 2014.

[12] E. F. Knott, J. F. Shaeffer and M. T. Tuley, "Radar cross section: Its prediction measurement and reduction." Dedham, 1985.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez and et al., "Attention is all you need", *Advances in NIPS*, pp. 6000-6010, 2017.

[14] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji and G. Xing, "MilliEye: A Lightweight mmWave Radar and Camera Fusion System for Robust Object Detection." In *Proceedings of the International Conference on IoTDI*, 2021, pp. 145–157.

[15] Y. A. U. Rehman, L. M. Po and M. Liu, "LiveNet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Systems with Applications*, 2018, 108: 159-169.

[16] S. Heuel and H. Rohling, "Two-stage pedestrian classification in automotive radar systems," *2011 12th IRS*, 2011, pp. 477-484.

[17] K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer and B. Yang, "Deep Learning-based Object Classification on Automotive Radar Spectra," *2019 IEEE RadarConf*, 2019, pp. 1-6.

[18] A. Angelov, A. Robertson, R. Murray-Smith and F. Francesco, "Practical classification of different moving targets using automotive radar and deep neural networks," *IET Radar, Sonar & Navigation*, 2018, 12(10), pp. 1082-1089.

[19] O. Schumann, M. Hahn, J. Dickmann and C. Wöhler, "Semantic Segmentation on Radar Point Clouds," *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 2179-2186.

[20] A. Danzer, T. Griebel, M. Bach and K. Dietmayer, "2D Car Detection in Radar Data with PointNets," *2019 IEEE ITSC*, 2019, pp. 61-66.

[21] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," In *Proceedings of the IEEE/CVF WACV*, 2021, pp. 1527-1536.

[22] D. Cao, R. Liu, H. Li, S. Wang, W. Jiang, and C. X. Lu, "Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars," In *Proceedings of the ACM UbiComp*, 2022.

[23] R. Yadav, A. Vierling and K. Berns, "Radar + RGB Fusion For Robust Object Detection In Autonomous Vehicle," *2020 IEEE ICIP*, 2020, pp. 1986-1990.

[24] A. Sengupta, A. Yoshizawa and S. Cao, "Automatic Radar-Camera Dataset Generation for Sensor-Fusion Applications," in *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2875-2882, April 2022.

[25] Y. Wang, Z. Jiang, X. Gao, J. N. Hwang, G. Xing and H. Liu, "Rodnet: Radar object detection using cross-modal supervision," In *Proceedings of the IEEE/CVF WACV*, 2021, pp. 504-513.

[26] T. -Y. Lim, S. A. Markowitz and M. N. Do, "RaDICaL: A Synchronized FMCW Radar, Depth, IMU and RGB Camera Data Dataset With Low-Level FMCW Radar Signals," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 941-953, June 2021.

[27] J. R. Schoenberg, A. Nathan and M. Campbell, "Segmentation of dense range information in complex urban scenes," *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 2033-2038.

[28] H. Cho, Y. -W. Seo, B. V. K. V. Kumar and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," *2014 IEEE ICRA*, 2014, pp. 1836-1843.

[29] J. Schlosser, C. K. Chow and Z. Kira, "Fusing LIDAR and images for pedestrian detection using convolutional neural networks," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2198-2205.

[30] 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*, 2017, pp. 1907-1915.

[31] J. Ku, M. Mozifian, J. Lee, A. Harakeh and S. L. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1-8.

[32] M. Liang, B. Yang, S. Wang and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641-656.

[33] F. Gomez-Donoso, E. Cruz, M. Cazorla, S. Worrall and E. Nebot, "Using a 3D CNN for Rejecting False Positives on Pedestrian Detection," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-6.