End-to-End Target Liveness Detection via mmWave Radar and Vision Fusion for Autonomous Vehicles

SHUAI WANG, Southeast University, China
LUOYU MEI, Southeast University, China and City University of Hong Kong, China
ZHIMENG YIN, City University of Hong Kong, Hong Kong and City University of Hong Kong Shenzhen Research Institute, China
HAO LI, Southeast University, China
RUOFENG LIU, University of Minnesota Twin Cities, United States
WENCHAO JIANG, Singapore University of Technology and Design, Singapore
CHRIS XIAOXUAN LU, The University of Edinburgh, United Kingdom

The successful operation of autonomous vehicles hinges on their ability to accurately identify objects in their vicinity, particularly living targets such as bikers and pedestrians. However, visual interference inherent in real-world environments, such as omnipresent billboards, poses substantial challenges to extant vision-based detection technologies. These visual interference exhibit similar visual attributes to living targets, leading to erroneous identification. We address this problem by harnessing the capabilities of mmWave radar, a vital sensor in autonomous vehicles, in combination with vision technology, thereby contributing a unique solution for liveness target detection. We propose a methodology that extracts features from the mmWave radar signal to achieve end-to-end liveness target detection by integrating the mmWave radar and vision technology. This proposed methodology is implemented and evaluated on the commodity mmWave radar IWR6843ISK-ODS and vision sensor Logitech camera. Our extensive evaluation reveals that the proposed method accomplishes liveness target detection with a mean average precision of 98.1%, surpassing the performance of existing studies.

CCS Concepts: • **Hardware → Sensor devices and platforms**; **Sensor applications and deployments;**

Additional Key Words and Phrases: Target liveness detection, mmWave radar

S. Wang and L. Mei contributed equally to this research and should be regarded as co-first authors.

This work was supported in part by Science and Technology Innovation 2030 - Major Project 2021ZD0114202, National Natural Science Foundation of China under Grant No. 62272098, NSF China 62102332, ECS CityU 21216822, City University of Hong Kong 9610491, CityU 11206023 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 and Development Programme.

Authors' addresses: S. Wang (Corresponding author) and H. Li, Southeast University, Nanjing, China; e-mail: {shuaiwang, 220194384}@seu.edu.cn; L. Mei, Southeast University, China and City University of Hong Kong, China; e-mail: lymei-@seu.edu.cn; Z. Yin, City University of Hong Kong, Hong Kong and City University of Hong Kong Shenzhen Research Institute, China; e-mail: zhimeyin@cityu.edu.hk; R. Liu (Corresponding author), University of Minnesota Twin Cities, Minneapolis, United States; e-mail: liux4189@umn.edu; W. Jiang, Singapore University of Technology and Design, Singapore; e-mail: wenchao_jiang@sutd.edu.sg; C. X. Lu, The University of Edinburgh, Edinburgh, United Kingdom; e-mail: xiaoxuan.lu@ed.ac.uk.

1550-4859/2024/05-ART93 $15.00
https://doi.org/10.1145/3628453

ACM Reference format:
Shuai Wang, Luoyu Mei, Zhimeng Yin, Hao Li, Ruofeng Liu, Wenchao Jiang, and Chris Xiaoxuan Lu. 2024. End-to-End Target Liveness Detection via mmWave Radar and Vision Fusion for Autonomous Vehicles. *ACM Trans. Sensor Netw.* 20, 4, Article 93 (May 2024), 26 pages.
https://doi.org/10.1145/3628453

1 INTRODUCTION

Autonomous vehicles must have accurate and robust sensing abilities in complex real-world road environments. Liveness target detection is one critical sensing ability. On the one hand, misidentifying the living targets might lead to serious car accidents (e.g., a Tesla autopilot caused a fatal motorcycle accident in 2022 [6]). On the other hand, a false alarm also causes trouble. For example, roadside billboards with human portraits will bring serious visual interference to the perception and decision of autonomous vehicles, which could lead to erroneous control strategies such as sudden braking, unexpected changes in direction, and sudden speeding up [3]. In future full autonomous driving (e.g., L5), it is very significant for vehicles to prevent these mistakes and distinguish the living targets from visual interference accurately.

In the real-world environments of autonomous driving, a diverse range of visual interferences, such as LED billboards on vehicles, often exist, leading to the inaccuracy of methods solely based on visual perception. Despite this, the most common sensors in autonomous driving vehicles are mmWave radar and camera [9, 18]. Cameras offer abundant visual data for identifying road objects but struggle to differentiate between visually alike targets and often fail to discern the liveness of dynamic visual interference (e.g., a video of a biker) displayed on LED billboard advertisements. Conversely, the mmWave radar utilizes the reflection of millimeter wave signals for sensing unaffected by visual interference and robust under harsh sensing environments, e.g., dark, foggy, and rainy conditions. Nonetheless, mmWave radar possesses its own constraints. The point clouds produced by mmWave radar tend to be sparse and noisy and have a significantly lower angular resolution compared to camera imagery [35], which is insufficient for accurate target liveness detection. Moreover, the performance of radar liveness target detection depends on the target's orientation, which is difficult to obtain from the point cloud due to its sparsity.

Motivated by the necessity for liveness target detection and the potential of multi-modal fusion, this work presents the first end-to-end methodology for liveness target detection harnessing the strengths of mmWave radar and vision technologies. The key idea underpinning our design is to leverage the inherent complementarity between radar and camera technologies. In particular, radar provides liveness-related information that is impervious to visual interference, albeit with some sparsity and noise [37]. In contrast, vision data, while susceptible to visual interference, furnishes detailed visual features that enable precise target segmentation.

In light of the key insight, we propose a multi-modal data fusion machine learning scheme that collaborates the mmWave radar with vision sensors for target liveness detection. We exploit the unique features in the mmWave radar and the vision data. In the mmWave radar data, the **Radar Cross Section (RCS)** is extracted as the distinguishable characteristic for liveness detection. The RCS, influenced by numerous target features such as shape, size, and material, serves as a comprehensive metric. Utilizing the off-the-shelf mmWave radar device, we measure the RCS of the living target and visual interference. The result shows that the RCS of the living target significantly differs from the visual interference. Moreover, the RCS values are stable under the detection ranges of mmWave radar, which means as long as the targets are within the field of vision, the results are not affected by the distance from the object to the radar.

Meanwhile, we obtain detailed visual features from the visual data to estimate the target's posture, offering essential angle-of-view insights to counteract the radar signal's sensitivity to the target's posture. Finally, mmWave radar and visual features are collaboratively utilized to identify the true living targets and prevent visual interference. To achieve this, we extend MediaPipe (a cross-platform, customizable machine-learning methodology for target detection) to support mmWave radar and vision multi-modal data fusion. This fusion of technologies leads to a refined liveness target detection process, effectively handling scenarios where a human target coexists with metal, such as bikers. Consequently, our contribution transcends individual advancements in RCS and vision technologies, promoting a novel collaboration for superior liveness target detection. The approach we have established forms a solid foundation for future exploration of additional features that could further enhance detection robustness and accuracy.

This article has the following contributions:

- This article presents the first end-to-end liveness target detection methodology utilizing the mmWave radar and vision fusion, which enhances the robustness of autonomous vehicles.
- This article presents a novel design that extracts the RCS feature from mmWave radar signal and utilizes a multi-modal data fusion machine learning method for mmWave radar and vision data fusion. The proposed methodology utilizes the complementary information in the mmWave radar and vision data for robust liveness target detection, thus avoiding visual interference.
- We implement and evaluate the performance of the proposed design on the off-the-shelf mmWave radar board IWR6843ISK-ODS and vision sensor Logitech camera by experimenting with the system in different scenarios. The results demonstrate that the proposed method achieves the liveness target detection with a **mean average precision (mAP)** of 98.1%.

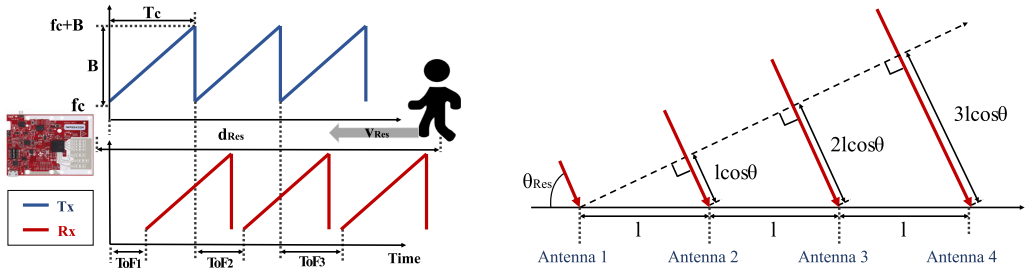
2 MOTIVATION

2.1 The Necessity of Liveness Target Detection in Autonomous Vehicles

Accurately distinguishing living targets from visual interference is important for autonomous vehicles to make correct and safe control. In the real-world road environment, the living target looks similar to the visual interference. For example, autonomous vehicles still do not have the capacity to fully distinguish a real biker from the bikers in advertisements on billboards along the road utilizing visual information only. These mistakes make autonomous vehicles execute incorrect actions and lead to serious accidents. However, being unable to distinguish living targets from visual interference also results in serious issues. For instance, Tesla's autopilot was involved in a third fatal motorcycle crash in 2022, raising questions about the driver-assist system's ability to operate safely [6]. A robust liveness target detection is needed to improve the safety of autonomous vehicles. Specifically, when the autonomous vehicle faces living targets and visual interference, our design aims to detect living and exclude visual interference (e.g., billboard).

2.2 Limitation of Existing Solutions

In recent years, research progress has been made in the field of target detection based on mmWave radar and vision data. The traditional target detection approaches, e.g., CRUW [39], CARRADA [24], nuScenes [8], LiveNet [27], and RadarScenes [31], are based on the position or intensity of the body to achieve object detection and recognition. In general, there are two kinds of commonly utilized methodologies for achieving liveness target detection. One is based on the contour feature of the face to realize liveness target detection [27, 39], but these methods are only utilized in the case of close faces, so they are not suitable for automatic driving scenes. The other method is based on the contour features of the human body to realize the detection of living



(a) Radar Multiple Chirps (b) Antennas Array

Fig. 1. (1) Compute the range information from one ToF; (2) compute the velocity information from over-chirps ToF differences; (3) compute the angle information from over-antenna ToF differences.

targets [26, 28], and the implementation effect of this method in the autonomous driving scene is relatively much better.

However, these methods fail to distinguish living targets from visual images, e.g., recognizing a billboard with a human portrait as a real person. This is because living targets and visual interference may have similar visual characteristics, so visual information alone is not enough for liveness target detection. In this work, we propose a new design that takes advantage of mmWave characteristics to make the liveness target detectors robust to the disturbance of visual images.

2.3 Opportunities of mmWave Radar and Vision Fusion for Liveness Target Detection

Most autonomous vehicles have vision sensors, i.e., cameras and mmWave radars that are fused for liveness target detection besides visual and wireless sensing. The camera provides the capacities of target segmentation and posture recognition from visual data. In contrast, the mmWave radar brings the opportunity of liveness target detection to distinguish living targets from visual interference. Moreover, the RCS feature extracted from the mmWave radar signal is stable with visual interference, such as billboards, which solves the limitations of visual sensors. In conclusion, the mmWave radar and vision sensors have complementary features for robust liveness target detection.

3 BACKGROUND

This section introduces the primer knowledge of the mmWave radar and machine learning methodologies utilized in this work.

3.1 mmWave Radar Principles

The fundamental principle of mmWave radars is straightforward. A mmWave signal is generated and transmitted from the radar toward a target that scatters the transmitted mmWave signal. Some scattered waves are redirected back toward and intercepted by the receiving antenna. The received wave contains information from different dimensions, e.g., time, space, and frequency. This information is extracted and supports the mmWave radars' fundamental range, velocity, and angle measurement functions. This subsection introduces the principle for those functions in an easy-to-understand way. The key idea behind frequency-modulated continuous-wave radar is to measure the **Time-of-Flight (ToF)** of signal from antennas by comparing it with its delayed signal reflected by the target. As shown in Figure 1, the range measurement is done by multiplying half of the ToF with the signal speed. The velocity measurement is done by detecting

the cross-chirps ToF differences, as shown in Figure 1(a). Moreover, the angle measurement is done by estimating the cross-antenna ToF differences, as illustrated in Figure 1(b).

Range Measurement. Range measurement is done by measuring the propagation time of the reflected wave and multiplying it by the wave propagation speed. Specifically, the mmWave radars transmit frequency-modulated continuous wave radio to measure the range. After the antenna sends the signal, it will be blocked by objects in the transmission path and reflected. The receiving antenna of mmWave radar will capture the transmitting linear frequency modulation pulse. A mixer inside the radar combines the sent signal with the captured reflected signal to produce the intermediate frequency signal. Then the fast Fourier transform operation is performed on the intermediate frequency signal to separate different frequency components and thus get the distance between each object and the radar denoted as $d = \frac{cfT_c}{2B}$. Based on the Fourier transform theory, the range measurement resolution is limited as $d_{Res} = \frac{c}{2B}$, which depends on the wave frequency bandwidth. For example, a bandwidth of 4 GHz gives a range resolution of 3.75 cm.

Velocity Measurement. The movement of the target results in a frequency shift of the received wave determined as the Doppler shift. The Doppler shift value is utilized for velocity measurement. The principle of velocity measurement is finding the distance change amount in a short period, i.e., T_c. The FMCW technology continuously emits a set of N equally spaced linear frequency modulation pulses, called a frame, to obtain multiple phases after the distance fast Fourier transform and then perform the Doppler fast Fourier transform called slow time fast Fourier transform, on multiple phases. The phase value w of different objects are obtained and thus obtain the amount of change in distance, i.e., $\Delta d = \frac{\lambda \Delta \phi}{4\pi d}$, Finally, the velocity $v = \frac{\lambda w}{4\pi T_c}$ is obtained. Also, according to Fourier transforms theory, the velocity resolution is denoted as $v_{Res} = \frac{\lambda}{2T_f}$, where $T_f = NT_c$, which depends on the length of the observation window. Larger window size leads to higher velocity measurement resolution.

Angle Measurement. The **Angle-of-Arrival (AoA)** measurement represents the receiving wave direction of the antenna array. The principle of AoA measurement is that the same object is at different distances to different receiving antennas, and the distance difference has a trigonometric relationship with the AoA. So as long as this distance difference is calculated, the AoA is calculated based on the geometric relationship between it and the distance l between antennas. Similarly to the Doppler fast Fourier transform, it is necessary to perform a fast angular Fourier transform to distinguish multiple objects at the same distance with the same velocity to separate the components of different objects. After obtaining $\Delta\phi$, the angle θ is calculated by $\theta = sin^{-1}(\frac{\lambda \Delta \phi}{2\pi l})$. The angle measurement resolution is calculated as $\theta_{Res} = \frac{\lambda}{Ndcos(\theta)}$. Angle measurement resolution is related to the number of the antenna array. The more receiving antennas, the better the angular resolution.

3.2 Visual-based Machine Learning Principles

Visual methods based on machine learning have been widely utilized in three-dimensional (3D) target detection and liveness recognition. However, due to the need for more data and the diversity of the appearance of the recognized objects, it is a challenging research topic to detect 3D objects. Therefore, we choose the state-of-the-art 3D target recognition framework MediaPipe [21] and related algorithms to implement the 3D vision-based target recognition part of our model.

Target Segmentation. Under the development framework of MediaPipe, the model of a single-stage pipeline is utilized to achieve target segmentation. The backbone of the model is an encoder–decoder framework based on the MobileNetv2 model, and a multi-task learning method is utilized to realize object detection and object shape frame prediction. The 2D projections of eight boundary points of the object shape frame are estimated by regression. Finally, a mature pose estimation

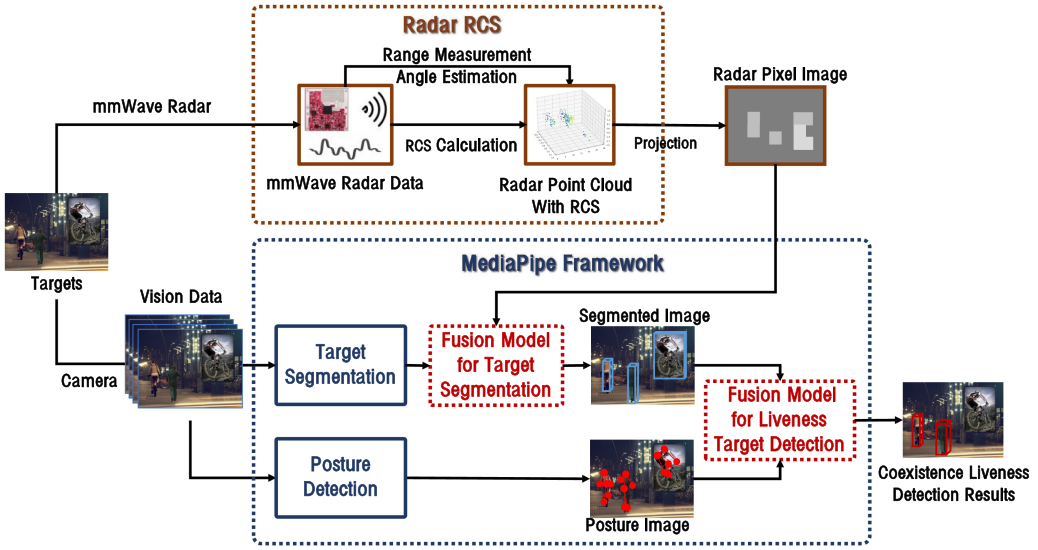


Fig. 2. System architecture.

algorithm EPnP is utilized to recover the 3D bounding box of the object to realize the detection of the object and the estimation of the shape and pose.

Posture Recognition. The MediaPipe framework also implements human pose detection based on a two-stage detector-tracker ML pipeline. The tracker first utilizes a detector to locate the person/pose **Region of Interest (ROI)** within the frame. Frames cropped to the ROI are then utilized as input to predict pose landmarks and segmentation masks within the ROI. Finally, the bounding boxes of different parts of the human body are displayed and rendered in the image, and the human pose is judged through these results.

4 SYSTEM OVERVIEW

This article introduces an end-to-end liveness target detection system with mmWave radar and vision fusion. The input of this system is signal data from the mmWave radar and vision data from the camera. Figure 2 presents the overall system architecture. The mmWave signal data from the radar are processed by the RCS calculation methodology (Section 5). The output of the RCS calculation process is the point cloud of the mmWave radar with RCS information. Then, the point clouds are projected into radar pixel images with RCS information. Meanwhile, the vision data from the camera are sent into MediaPipe [21] in the stream. During that, the vision data are divided into frames and processed for target segmentation and posture recognition. The target segmentation process employs advanced image processing techniques to identify potential targets within an image. It segments the image into different regions, each potentially containing a target object. The posture recognition process leverages machine learning techniques to recognize the posture or orientation of the detected objects. It classifies the objects based on their features, providing an additional layer of information that significantly enhances the overall object detection process. The output target segmentation results and the output target posture results are then sent together with the mmWave radar pixel image into two designed fusion models for liveness target detection. The fusion models combine mmWave radar and vision sequence data in a time-series format, marking a significant contribution to our work. They exploit the unique features (e.g., movements) within radar and image sequence data over time, which help to distinguish living targets from

dynamic visual interferences such as LED billboards, a situation where conventional state-of-the-art techniques often fall short. Two fusion models are detailed in Section 6. At last, the system outputs bounding box results for the living target to distinguish them from the visual interference.

5 FEASIBILITY OF LIVENESS TARGET DETECTION

This section introduces the RCS, the crucial feature for mmWave radar liveness target detection. This work first introduces the principle of RCS. Then we illustrate the methodology of RCS calculation in commodity mmWave radars. After that, we provide a detailed study of living targets' RCS measurements.

5.1 Radar Cross Section

The RCS measures the object's reflection area to radar. A larger RCS indicates that the object is more obvious and easier to be detected by radar [29]. The object's RCS is related to its size, material, and orientation. The RCS features were initially utilized for aircraft and missile detection and distinguishment in the military field. Moreover, we find that RCS features have some unique characteristics that bring opportunities for mmWave and vision fusion liveness target detection. (i) First, the RCS of the living target has a huge difference with its visual interference. (ii) Second, the RCS value is consistent in various sensor-to-target distances. (iii) Third, the RCS of the living target is related to its posture (the angle of view toward radar). Therefore, utilizing the vision data for target posture detection and fusion with mmWave radar data for liveness target detection improves the robustness of the autonomous vehicles sensing ability.

5.2 RCS Acquisition on Commodity Radar

RCS is a physical quantity that measures the intensity of the echo generated by the target illuminated by the radar wave. We obtain the RCS value of the object reflection to judge the shape characteristics of the target. But off-the-shelf radars (e.g., the TI IWR series) do not provide RCS results directly. Therefore, we need to obtain the other parameters (including the radar parameters) before calculating the RCS value.

In calculating RCS, the most critical and difficult problem is obtaining the SNR, which is the ratio of RX average signal strength and average noise strength. The noise mainly depends on the background noise of the radar circuit. Although the intensity of this noise is relatively stable, it is difficult to measure the noise value on the integrated radar equipment directly. Therefore we do not directly obtain the noise value but consider that the RCS should always be proportional to the RX signal strength when the other parameters are determined. Therefore, we decided to use the corner reflector [19] in the calibration phase to obtain the RX signal strength at different distances d and establish the benchmark database $\mathcal{B}(d)$. From $\mathcal{B}(d)$, we obtain the space Cartesian coordinates (x, y, z); if the distance $L = \sqrt{x^2 + y^2 + z^2}$, then we calculate RX signal intensity P_{r_t} based on the following formula:

$$\sigma_p = \frac{P_{r_t}}{\mathcal{B}L}\sigma_r. \tag{1}$$

We obtain the benchmark database $\mathcal{B}(d)$ during the calibration phase. From $\mathcal{B}(d)$, we obtain the signal-to-noise ratio, and then we obtain the RCS based on the following formula:

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

Here k is the Boltzmann constant, T is the antenna temperature, F is the noise coefficient of RX, λ is the millimeter-wave wavelength (constant), P_t is the radar output power, G_{TX} is the TX Antenna Gain, G_{RX} is the RX Antenna Gain, T_{meas} is the measurement time, and d is the distance

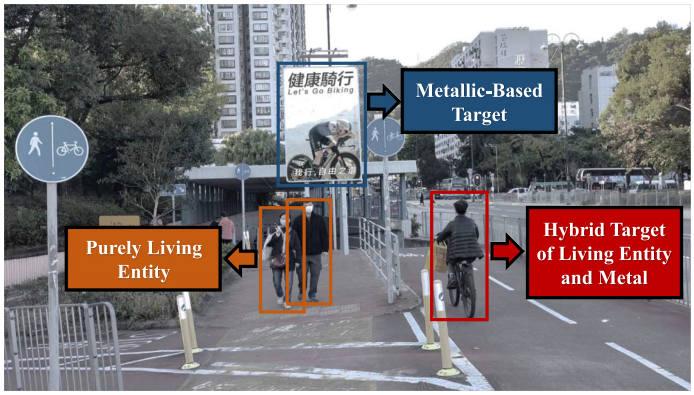


Fig. 3. Experimental scenarios of RCS value differences between living target and metal.

Table 1. Experimental Scenarios and RCS Values of Living Target and Visual Interference

Type	Scenario	Target Object	RCS
Visual Interference	**Metallic-Based Target**		
	1	vehicle	500 m^2
	1	billboard	230 m^2
Living Targets	**Hybrid Target of Living Entity and Metal**		
	2	Biker side view	50 m^2
	2	Biker front view	20 m^2
	Purely Living Target		
	3	Pedestrian	3.5 m^2

between the target and the millimeter-wave radar. d is computed from the radar output and the metadata. We obtain the RCS value of the target through the above method.

5.3 RCS Value of Living Targets and Visual Interference

We conduct experiments to measure the RCS value of the living targets and the visual interference utilizing the commercial mmWave radar IWR6843-ODS to verify the feasibility of RCS for liveness target detection. As shown in Figure 3, there are two groups of target objects scenarios. The first scenario contains the metal and living target. The second scenario contains the metal with the hybrid target of living entity and metal advertisement in the same place as the first scenario. The metal is 0.8m × 0.7m in size and has aluminum alloy in the material. We utilize mmWave radar to measure the RCS value of the living targets and the visual interference around the different ranges and different angles of view.

Table 1 illustrates the significant distinction present in the RCS between living targets and visual interferences. Three broad categories of objects are prevalent on streets: metallic-based targets (e.g., vehicles and billboards), hybrid targets of Living entity and Metal (e.g., bikers), and purely living targets (e.g., pedestrians). RCS is defined as the effective area that intercepts the transmitted radar power and then scatters that power isotropically back to the radar receiver; therefore, the unit of RCS is m^2. Quantitatively, the average RCS for a pedestrian is 3.5 m^2. The RCS value of a Biker is 50 m^2 when observed from the side and 20 m^2 when observed from the front, indicating the influence of metal in the hybrid category. However, metallic objects significantly inflate the

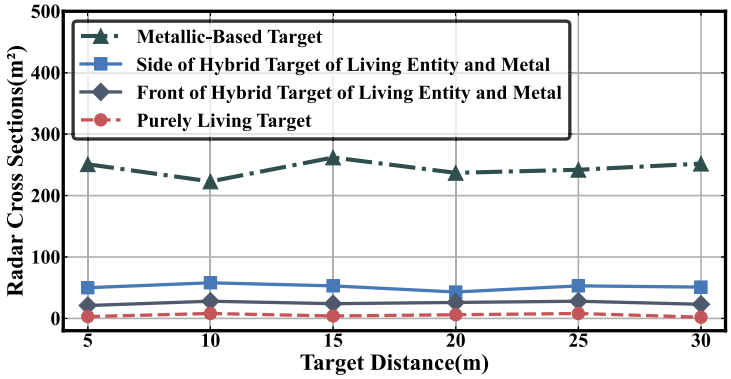


Fig. 4. Experiment result of RCS from different distances.

RCS values, with the billboard and vehicle documenting figures of 230 m^2 and 500 m^2, respectively. The divergence in these figures emphasizes the potential of liveness target detection within the landscapes of autonomous driving.

5.3.1 RCS from Different Distance. We further demonstrate the RCS of the living target and the visual interference when the sensors are at different distances, moving from 5 m to 30 m, from the object. As shown in Figure 4, the RCS of the hybrid target of living entity and metal from the side, the front view, and the living target is very stable in the range of 47–51 m^2, 17–22 m^2, and 2–5 m^2, respectively. However, the RCS of the metal is in the range of 200–300 m^2. Therefore, the RCS value of the living target differs from the visual interference. Moreover, the RCS value of the same target at different distances is generally consistent, which supports the theory that the target's RCS, in the sensing coverage of mmWave radar, is not influenced by the target distance.

5.3.2 RCS from Different Angles of View. The signal from the mmWave radar is highly sensitive to an object's posture, indicating a strong correlation between the object's viewing angle and its RCS value. To quantify this effect, we undertake a series of experiments to assess viewing angle effects on RCS. As illustrated in Figure 5, we collect radar RCS data for the living target and the visual interference across a range of viewing angles. Our findings highlight the significant variation in the RCS values of both the living target and the visual interference. Moreover, it is notable that despite the RCS value for both the hybrid target of living entity and metal and the metallic-based target varying significantly in alignment with the viewing angle, the hybrid target of living entity and metal distinctly alters based on the target posture. This insight motivates us to further integrate the mmWave radar and vision fusion cues, such as target posture, to fortify the robustness of liveness target detection.

Figure 6 shows the RCS results from different angles of view. Although RCS is changing related to the angle of view, the change range of the living target and the visual interference is different. The RCS value of the metal rises from 50 to 300 m^2 when the angle of view changes from 0° to 90°, and then the RCS value decrease to 54 m^2 when the angle of view changes from 90° to 180°. Meanwhile, the RCS value of the hybrid target of living entity and metal rises from 17 to 51 m^2 when the angle of view changes from 0° to 90°, and then the RCS value decreases to 20 m^2 when the angle of view changes from 90° to 180°.

5.4 Summary

The experiment results prove that the RCS values of the living target and the visual interference differ significantly. Moreover, the RCS value is related to the angles of view, and vision data

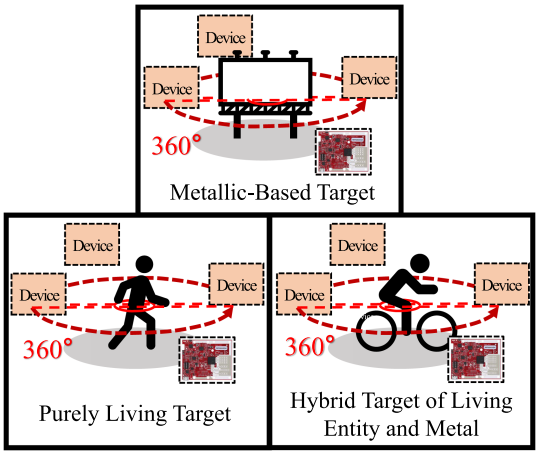


Fig. 5. Experimental scenarios of RCS from different angles of view.

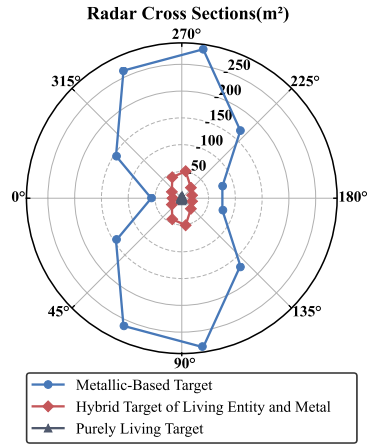


Fig. 6. RCS from different angles of view.

contains the target's angle of view information. These bring opportunities for utilizing mmWave radar and vision fusion for liveness target detection.

6 DESIGN

6.1 Design Overview

This section proposes a multi-modal data fusion liveness target detection methodology, which fuses the mmWave radar and vision data to distinguish the living target from visual interference. The common method of utilizing mmWave radar and vision data together is first to process target segmentation by the two-stage image-based detectors (e.g., Faster-RCNN), utilizing the vision data. Then map the mmWave radar data into the corresponding target to distinguish whether it is the living target or the visual interference. However, our experiments demonstrate that the two-stage image-based detectors come into a short while detecting the target at different angles of view and only achieve mAP of 57.5% (detailed results are shown in Section 8).

To achieve robustness in distinguishing the living target from visual interference at different angles of view. This article proposes an end-to-end liveness target detection system with mmWave radar and vision data. The system utilizes mmWave signal data for liveness target detection and vision data for target segmentation and posture detection. During the target segmentation process, this article utilizes machine learning target segmentation methodology for live and streaming media (e.g., MediaPipe [21]). Specifically, visual data streams from the camera are input into the single-stage pipeline in MediaPipe [4]. The pipeline estimates the real-time 3D object poses through the machine learning model. Then the radar pixel images from mmWave radar data are fused into the output of the single-stage pipeline for getting the target segmentation results with the RCS features. Meanwhile, during the posture detection process, the vision data are sent into the two-step detector-tracker pipeline in MediaPipe [5]. The pipeline first locates the target's region of interest in the vision data, and then the tracker subsequently predicts the target's posture within the ROI utilizing the ROI-cropped frame. The output in the target segmentation process and the posture detection process are then sent together into the designed fusion model for getting liveness target detection results. The system outputs bounding boxes around the living targets to distinguish them from visual interference.

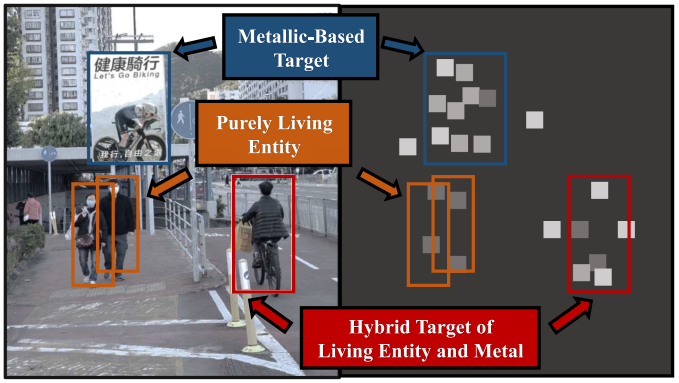


Fig. 7. Image and radar pixel image.

6.2 mmWave Radar Data Procession

The RCS values are extracted from the mmWave radar data. This subsection provides the RCS calculation algorithm for obtaining the RCS value of the target object in the 3D space. Because of the sparsity of mmWave radar point cloud, there are noise voxels without useful information for liveness target detection. To solve this problem and extract multi-scale features from mmWave radar data more efficiently, this article projects the mmWave radar data from the Cartesian coordinate system to the pixel coordinate system of the vision data by squeezing it on the depth direction (y-axis) ss shown in the following equation:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{y} M_1 M_2 \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{3}$$

Here M_1 is the 3×3 internal parameter matrix of the vision data, M_2 is the 3×4 external parameter matrix of the vision data, and (u, v) is the pixel location in the vision data's pixel coordinate system. Because the mmWave radar and vision sensors are fixed at an unchangeable place on the autonomous vehicle, the relative positions between mmWave radar and vision data remain static during the measurement stages. Therefore, M_1 and M_2 are parameters that are known ahead.

After that, the mmWave radar data are processed into a three-channel radar pixel image the same size as the vision data. Then, the target's RCS value is utilized to fill each pixel of the radar pixel image corresponding to the target location on the vision data. As shown in Figure 7, the left figure is the vision data collected by the vision sensor, and the right is the mmWave radar pixel data calculated from the mmWave radar data. The vision and mmWave radar pixel data are collected and calculated simultaneously.

6.3 Multi-modal Feature Extraction

As discussed in Section 6.1, the target objects are located at various angles of view and distances from the autonomous vehicle. The target objects' size in both the vision data and the mmWave radar pixel image decrease when the distance from the target to the sensors increases. To accurately detect the targets and obtain features from sparse mmWave radar data, this article provides the multi-scale feature extraction methodology for obtaining high-dimensional features from mmWave radar and vision data in different scales. This feature extraction methodology is utilized for the accurate detection of small-scale objects.

As shown in the blue block in Figure 2, this article utilizes the single-stage pipeline in MediaPipe for target segmentation and posture recognition. The signal-stage pipeline contains the encoder and decoder architecture. The multi-task learning approach is employed for jointly processing target segmentation and posture recognition. This article utilizes the annotated bounding boxes for the target segmentation task, and the box size is standardized to be the same as the deviation proportion. To obtain the bounding box's final 3D coordinates, we leverage the pose estimation algorithm [1] for point correspondence, which draws the 3D bounding box of the objects without prior knowledge of the object's shape. Then, this article further utilizes the multi-modal feature fusion methodology for the liveness target distinguishment.

6.4 Multi-modal Feature Fusion

The mentioned feature extraction module yields the features from the mmWave radar and the vision data. Then, the multi-modal features are utilized for liveness target detection. The common methodology for feature fusion is tensor concatenation. The multi-dimensional linear features' relationships are further extracted utilizing the multi-layer convolution operation. However, since the mmWave radar and the vision data are from different modalities, the convolution network comes short when learning the linear relationship between those two kinds of data. Our preliminary experiments, as shown in Section 8, provide evidence that the single-stage MediaPipe pipeline is better.

Because the mmWave radar and vision sensors are both on the autonomous vehicle and toward the same direction, the mmWave radar and vision data are regarded as different modalities of data of the same target from the same distance and angles of view. Therefore, this article proposes the multi-modal features fusion methodology for associating those data in two different modalities, utilizing their relationship. Specifically, this article utilizes the features extracted from the mmWave radar pixel data to detect the locations of the living target in the vision data. The segmentation and the detection process are similar to the process in the human brain while ignoring useless information and focusing on important information. The mmWave radar pixel data is regarded as the weight matrix for the segmentation and detection process. Technically, the segmentation and detection process is done by fusing the vision and the mmWave pixel data utilizing the attention mechanism [36]. The attention mechanism's multi-modal data fusion model is shown as follows:

$$\mathcal{F}_{i,j} = \varphi(\mathcal{V}_{i,j}) \odot \psi(\mathcal{R}_{i,j}). \tag{4}$$

Here the $\mathcal{F}_{i,j}$ is the fused data feature at the index (i, j), $\mathcal{R}$ is the feature from mmWave radar pixel data, $\mathcal{V}$ is the feature from vision data, the linear mapping operation is represented by φ, the two-dimensional convolution together with Softmax operation is represented by ψ, and $\odot$ represents for the Hadamada product operation. φ and ψ are the known parameters calculated from the sensors deployment setting.

The features in the vision and the mmWave radar pixel data are operated by the above-mentioned attention operation. Moreover, this article takes advantage of the feature pyramid network to provide extra context by feeding the features captured on the global scale to the local scale. Specifically, as shown in Figure 8, the fusion data features in layer five are up-sampled and then concatenated with the fusion data features in layer four. Layer four contains a new up-sampling and concatenation algorithm for generating the data features for layer three. After that, the result of liveness target detection and distinguishment is generated in these three layers.

In this article, two fusion models are presented utilizing the idea of the feature pyramid network. As shown in Figure 8, the first fusion model is designed based on radar pixel images and visual images for target segmentation in multi-target scenes. The fusion model puts the two image data into the residual network respectively and sets up an attention mechanism to improve the fusion

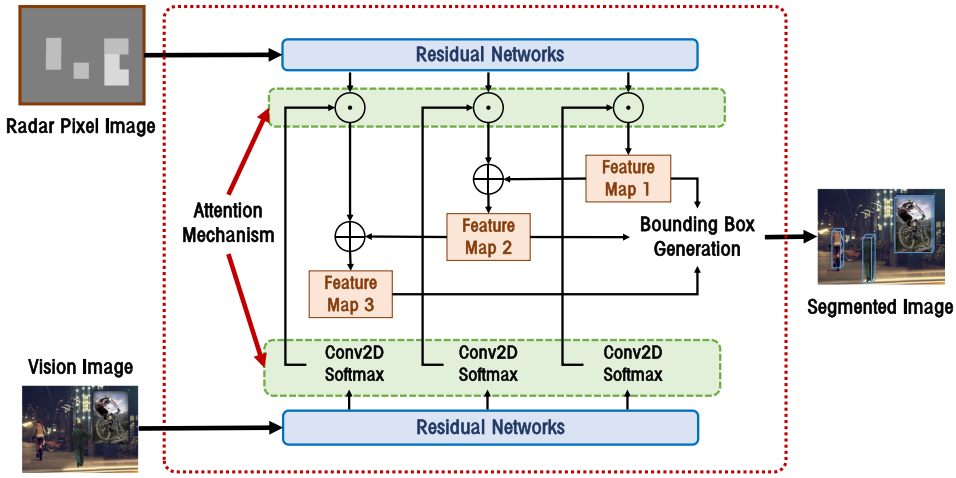


Fig. 8. Fusion model for target segmentation.

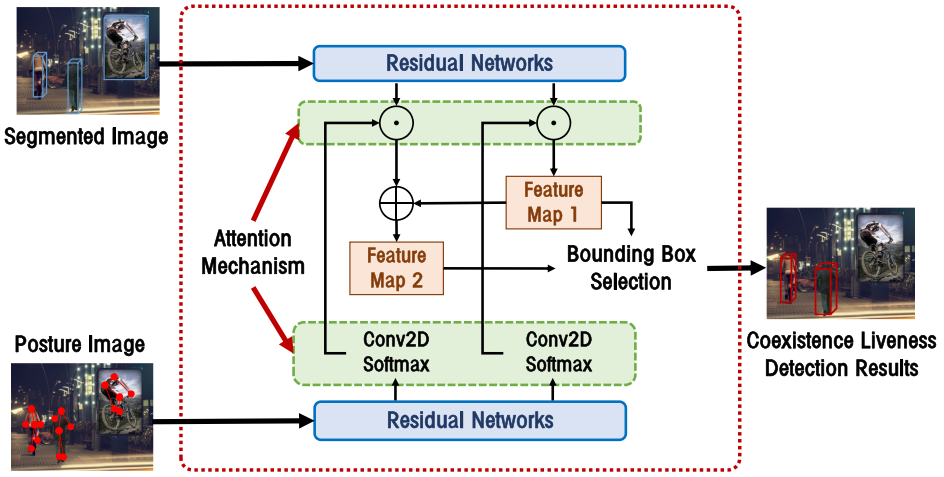


Fig. 9. Fusion model for liveness target detection.

performance. For the processing of visual images, this article utilizes the Conv2D Softmax model to extract image details. Then, the information obtained from both sides is merged and discriminated at the pixel level, and the fusion results of different residual levels are utilized to generate features. Finally, the object's bounding box is generated based on these different levels of features.

In addition to realizing the object segmentation function of the image, this article also implements the posture acquisition function through the posture detection method integrated with the MeidaPipe framework [5]. However, because this method is based on conventional image visual features, it cannot distinguish the difference between the living target and the metal from the fundamental principle, which leads to the wrong judgment of the human posture information on the metal. To solve this problem, this article also proposes a fusion model based on posture information and object segmentation information obtained from the above-mentioned fusion model. As shown in Figure 9, this model adopts a similar implementation mechanism as the previous fusion model, and the input data are posture information

ALGORITHM 1: Liveness Target Detection Algorithm

Input: *Images*, Camera Vision Images, *Radar*, Radar Pixel Image
Output: *Box*, Draw bounding boxes on the image

```
predict = self.net(Image, Radar)
output = self.bbox_util.decode_box(predict)
results = self.bbox_util.non_max_suppression(torch.cat(output, 1)
if results are empty then
    return Images
end if
top_label = np.array(results[0][:, 6], dtype = 'int32')
top_conf = results[0][:, 4] * results[0][:, 5]
top_boxes = results[0][:, :4]
% Draw bounding boxes on the image with results
return Images with bounding boxes
```

and target segmentation information, which are obtained from the above-mentioned target segmentation and posture detection process separately. Finally, several obtained bounding box schemes were screened and judged to achieve the correct bounding box of the living target.

6.5 Liveness Target Detection Complexity

The pseudocode for our proposed liveness target detection algorithm is presented in Algorithm 1. The algorithm's time complexity primarily depends on three operations: prediction model execution, bounding box decoding, and Non-Max Suppression. The prediction phase has a time complexity of $O(n)$, with n being the total number of image pixels. Bounding box decoding, processed on each predicted bounding box, has a time complexity of $O(m)$, where m denotes the total number of predicted bounding boxes, i.e., the number of living targets. Following this, the Non-Max Suppression eliminates the inaccurate bounding boxes, characterized by a time complexity of $O(m^2)$. Therefore, the overall time complexity of the algorithm is $O(n + m^2)$. This balanced time complexity showcases an optimal tradeoff between detection accuracy and computational efficiency, thereby meeting the crucial requirement of real-time response in autonomous driving systems.

To evaluate the latency of the proposed methodology in a real-world scenario, we conduct target liveness detection on a laptop equipped with an NVIDIA GeForce RTX 4060 GPU and an AMC Ryzen 9 7845HX, which exhibits computational capabilities comparable to those of the hardware utilized in autonomous vehicles, such as the AMD Ryzen in the Tesla Model 3. The total system latency averages around 50 ms. The processing of the mmWave radar signal and the extraction of the RCS pixel image contribute about 15 ms to this latency. Meanwhile, the target liveness detection introduces approximately 35 ms of latency. These metrics collectively attest to the system's viability for real-world applications.

7 IMPLEMENTATION AND DATA COLLECTION

7.1 Data Collection

Data Collection Platform. Based on the existing commercial high-definition camera Logitech Pro C920 and commercial mmWave radar IWR6843, this article designs and fabricates a mobile data collection platform to simulate an on-board system as shown in Figure 10. The configuration of the radar and camera is as follows: (1) The radar used in the experiment has three transmitting

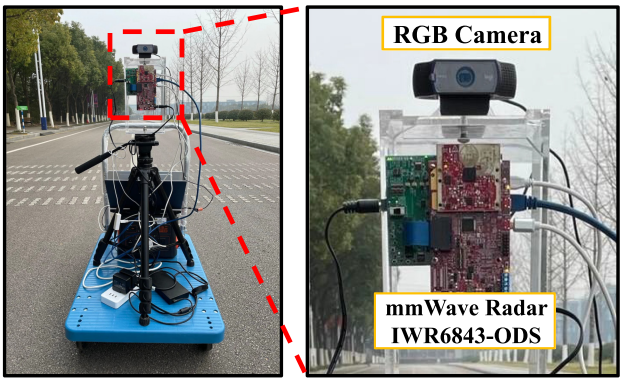


Fig. 10. Data collection platform.

antennas and four receiving antennas. The operating frequency band is from 60 to 64 GHz, the operating wavelength is ∼4 mm, the azimuth FoV is 60°, the elevation FoV is 60°, and the angular resolution is ∼15°. The radar transmits 64 chirps per frame. The chirp starting frequency is 60 GHz, the bandwidth is 1009.82 MHz, and the frequency slope is 21.038 MHz/μs. (2) The camera's resolution used in the experiment is 1920×1080, and its frame rate is 30 fps.

Experiment Site. This article conducts experiments using the data collection platform designed in the previous section. It simulates the scenario in question on an urban road in Hong Kong, where metals with images of hybrid target of living entity and metal are placed on the side of the road, and real-live hybrid target of living entity and metal are also standing in the middle of the road. The platform is slowly moved from a distance to the billboard and the hybrid target of living entity and metal, and the distance is about 30 m. While moving, two devices (a millimeter wave radar and a camera) collect data simultaneously. This article conducts eight sets of experiments, collecting a total of 26,050 frames of image data and 5,210 frames of radar data. In the process of training the model, the ratio of the validation set to the test set of the training data is set to 9:1. Finally, to avoid the situation of training overfitting caused by the image unchanged in the metal, this article replaces the metal rider pattern in the data with other rider images that do not exist in the dataset.

7.2 Object Segmentation for Radar Data

The problem of noise generated by multipath effects is difficult to solve for all RF techniques [16]. In this article, the main scene is outdoors, in which, in addition to the identified targets (portrait metal and hybrid target of living entity and metal), there are also some other static obstacles (trees and ground). The reflection and beam spread generated by these surrounding objects will lead to the multipath effect and eventually lead to a large number of interference information called ghost points in the point cloud information. This article uses the point segmentation strategy based on the DBSCAN clustering method to solve these ghost points and implement image segmentation. DBSCAN is a density-based clustering algorithm based on space distance and domain density to achieve point cloud division, so it is not sensitive to noise. Therefore, it alleviates the multipath effect of noise information on the experimental image. DBSCAN does not need to set the number of clusters in advance and automatically mark noise points and outliers, so it is often used to separate objects in processing millimeter wave radar point cloud data. Based on experience, in this article, the maximum distance (radius) between two points falling into the same cluster is set to 1, and the minimum number of points within a cluster is set to 3.

8 EVALUATION

8.1 Evaluation Methodology

Evaluation Metrics. In this article, we choose four parameters, Precision, Recall, F1 Score, and mAP, as evaluation indicators to measure the effect of the target detection task. Accuracy is equal to the number of true positive samples predicted as positive samples divided by the number of true positive samples. Recall equals the number of correctly predicted positive samples divided by the number of all positive samples. The F1 score is a comprehensive consideration of the precision and recall of the classification model. It is an indicator utilized to measure the classification accuracy of the binary classification model, which is regarded as the harmonic mean of the model's precision and recall. In this article, IOU is set as the intersection and union ratio of the sample 3D bounding box and the real-value 3D bounding box and the true and false attribute values of the sample 3D bounding box are determined based on the intersection and union ratio of the sample 3D bounding box and the real-value 3D bounding box. When the IOU of the sample 3D bounding box and the real-value 3D bounding box are greater than the set threshold, the sample is judged to be true.

Average Precision (AP) is equal to the area under the precision and recall curves determined by the confidence threshold, so it is utilized to show the overall performance of detection methods under different confidence thresholds. To obtain AP, the first step is to utilize the trained model to find a confidence score for all the bounding boxes and rank them based on that score. The second step is to select the top-1 to top-n results and compute precision and recall for the number of boxes. The last step is to plot all the precision and recall to form the P–R curve and then find the area under the P–R curve is the AP value. The mAP is the average AP of each class.

Competing Approaches.

- *Pure RCS:* This approach provides a fundamental insight into the potential of utilizing RCS alone for liveness target detection. The pure RCS approach does not take into account the multi-faceted influences of an object's orientation, shape, and viewing angle on the RCS values. Therefore, the accuracy is not good enough. However, the limitations of RCS are able to be mitigated by collaborating it with vision data. This fusion leverages the strengths of both modalities, creating a more comprehensive and reliable detection mechanism.
- *milliEye* [34]: Xian et al. proposed the milliEye in 2021. milliEye achieves target detection in low-light environments utilizing the vision and mmWave radar fusion. milliEye exploits the millimeter-wave radar's insensitivity to light and combines 3D radar point clouds with visual images for liveness target detection. It is considered a two-stage detector and mmWave radar point cloud fusion method without RCS information. In this article, milliEye is applied to the dataset to evaluate its performance.
- *Faster-RCNN and mmWave Radar RCS (Faster+RCS):* This is a two-stage detector and radar point clouds fusion method with RCS information, which is applied to the dataset in this article for calculating the liveness target detection performance.
- *w/o attention mechanism (No-Attention):* This article replaces the proposed attention mechanism-based feature fusion method with the tensor concatenation, which aims to investigate the influence of attention mechanism on performance.
- *w/o RCS (No-RCS):* This method replaces the RCS features in the mmWave radar data with the received signal intensity and makes a new radar pixel image dataset for liveness target detection. This experiment group aims to verify the effectiveness of RCS in liveness target detection.

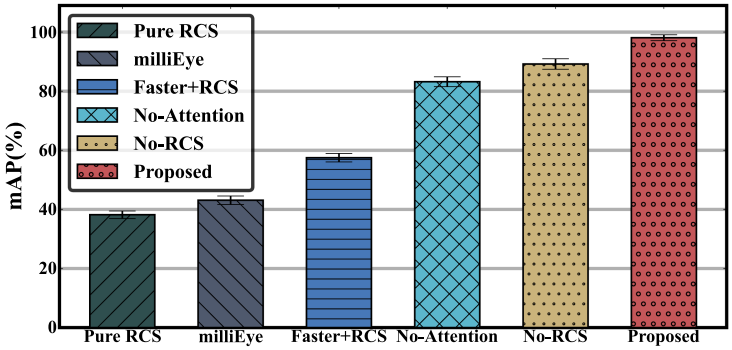
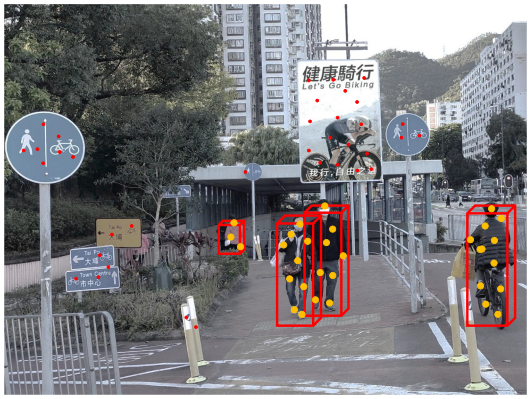


Fig. 11. Overall performance.



(a) Example on an open road.

(b) Example at a shopping plaza.

Fig. 12. Examples of liveness target detection.

8.2 Overall Performance

The proposed approach and four other baseline methods are compared, and a picture of the overall performance is plotted as shown in Figure 11. It is seen from the figure that the mAP of the proposed method reaches 98.1%, which is higher than all the other four baseline methods. The mAP of the proposed approach is 8.3% higher than that of the *No-RCS* method and is 14.9% higher than the *No-Attention* method. It shows that the radar's RCS value is better than its absolute intensity value in distinguishing between the living target and visual interference. The attention mechanism also helps to improve the fusion effect. Finally, both *Faster+RCS* and *milliEye* have low mAP, which indicates that there are obvious problems with the current two-stage method for detecting living targets in autonomous driving scenarios, such as inaccurate detection results of small-scale and long-range activity. Figure 12 shows two examples of liveness target detection in this article, where yellow points indicate the point cloud of live targets with low RCS values and red points indicate the point cloud of the billboard with high RCS values. Figure 12(a) demonstrates the liveness target detection on an open road, with bikers and pedestrians as the living targets, and a billboard with a biker figure as the interference. The result indicates that living targets are detected accurately. Furthermore, to test the maximum number of detectable targets, we evaluate our method in a crowded environment. Figure 12(b) illustrates the liveness target detection at a shopping plaza, with pedestrians as the living targets, and a metal Christmas

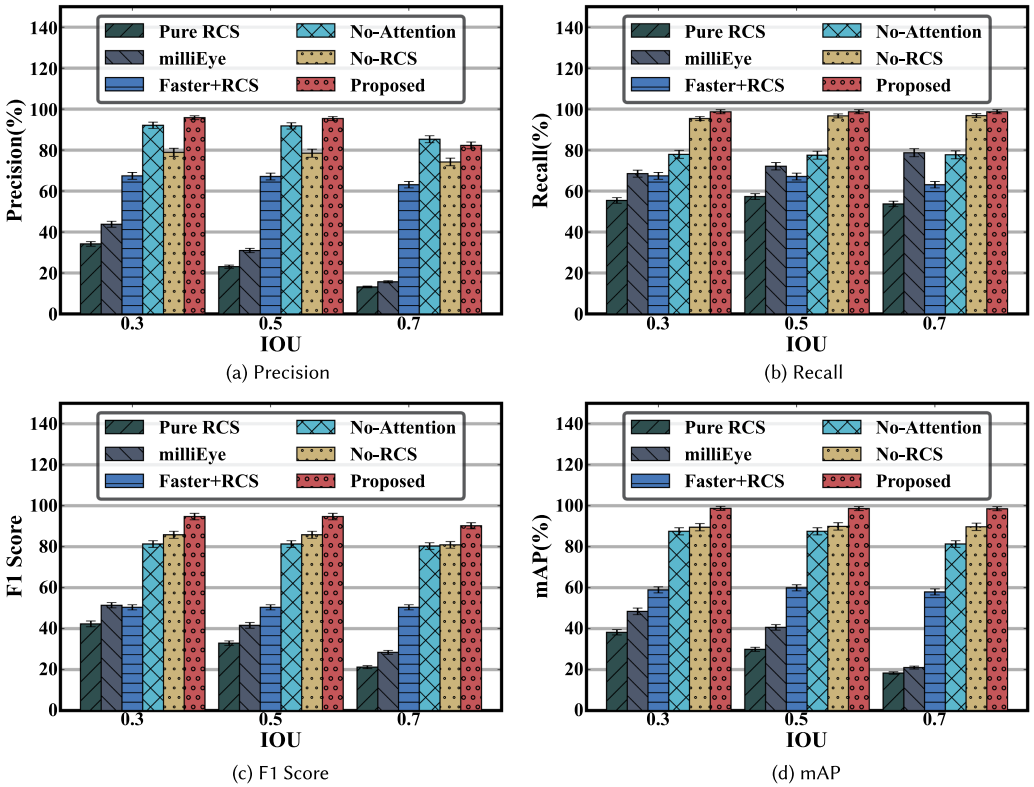


Fig. 13. Performance of proposed and baselines in different IoU thresholds.

tree as the interference. The result indicates that our method performs accurate liveness target detection for more than 10 targets simultaneously in a crowded environment.

8.3 Sensitivity Analysis

Precision, Recall, F1 Score, and mAP. Four evaluation indexes, namely precision, recall, F1 score, and mAP, are utilized to measure the performance of the proposed method. As shown in Figure 13, the proposed method is compared with five baseline methods. As a whole, the proposed method outperforms all the baseline methods in every metric. Specifically, regarding the *Pure RCS* method, it provides a fundamental perspective on utilizing RCS alone for liveness target detection. However, its performance is less satisfactory due to its failure to consider factors such as an object's orientation, shape, and viewing angle on the RCS values, which are accounted for in our proposed method. Furthermore, as illustrated in Figure 13(a), the accuracy of the proposed method, denoted as *Proposed*, exceeds that of the *NO-RCS* method. This is attributable to the replacement of radar signal strength with RCS value, which allows for a more effective distinction and measurement of living target characteristics. As shown in Figure 13(b), the recall rate of *Faster+RCS* method is relatively low, because this method cannot accurately carry out the target teaching material and segmentation. Figure 13(c) illustrates that the overall performance of *milliEye* method is significantly lower than other methods due to the lack of multi-modal fusion strategy and the RCS of mmWave radar, which also confirms that the method based on computer vision is susceptible to the interference of the external environment.

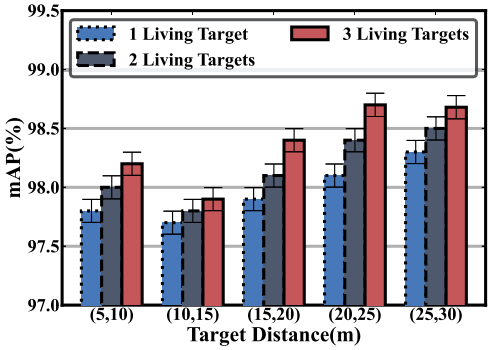


Fig. 14. Performance with target distance.

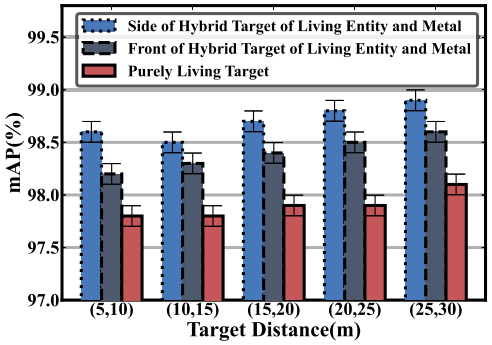


Fig. 15. Performance with target angle-of-view.

Impact of IoU threshold. The IoU threshold of the predicted active target and the real active target is set in the evaluation process, and the performance of each method in the case of different IoU values is evaluated. When the IoU threshold is larger, the bounding boxes of small-scale targets are more likely to be lost. As shown in Figure 13(d), the mAP of *Proposed* always remains above 97% and is better than the other four baseline methods, which indicates that the proposed method has high robustness. In addition, it is seen from the data in the figure that the performance of the method based on the single-stage multi-scale detector is significantly higher than that based on the two-stage detector, which proves that the multi-scale method realize the detection of small-scale targets in long-distance scenes.

Impact of target distance. In this article, the performance results for different distance cases are collected, and the amount of data to keep different distance cases is the same. As shown in Figure 14, although the distance between the target and the experimental platform will affect the proportion of the target imaging in the two kinds of sensing devices, the mAP of the proposed method is stable at 97% in the experimental scenarios with different numbers of living targets and different distances. Importantly, beyond a 30-m range, the detection performance of the commercial mmWave radar dramatically declines. This 30-m range is sufficient for an autonomous vehicle to perform essential braking or evasive maneuvers [23], setting the optimal functioning range of the system. Despite this constraint, our method demonstrates significant robustness to images captured at varying distances within the commercial mmWave radar's detection scope.

Impact of the number of living targets number. This article also collects the experimental results under different living target numbers, divides the dataset according to the different living target numbers, and evaluates the performance of the proposed method under these conditions. As shown in Figure 14, when the distance between the experimental target and the experimental platform is maintained to a certain extent, the overall mAP of the proposed method always remains above 97%, which indicates that the proposed method is robust to changes in the number of living targets. Because when the target gets closer to the mmWave radar, the display of the target on the mmWave radar is incomplete, and the performance increases when the target distance increases.

Moreover, we also evaluate the performance of liveness target detection under different target angle-of-view. To show the impact of the angle-of-view more clearly, we select the three most representative target angles-of-view as an example. As shown in Figure 15, because the side view of the hybrid target of living entity and metal has more information than the front, the proposed model has higher performance when facing the side view of the hybrid target of living entity and

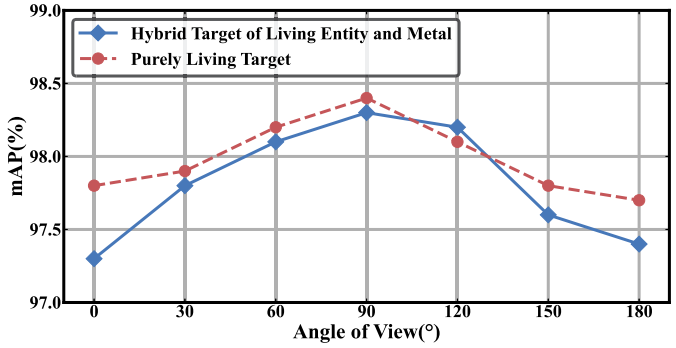


Fig. 16. Performance with different angle of view.

Table 2. Differences of Road Scenes

Scene	Condition	Trees Location	Lamps	Existing Living Target
1	Rough	Road End	Not exist	1
2	Rough	Not exist	Not exist	2
3	Flat	Both sides	Exist	1 & 2
4	Flat	Road End	Exist	2 & 3

metal than the front view of hybrid target of living entity and metal. The living target has less detection accuracy, because it has less RCS and information.

Impact of the Angle of View. Figure 16 illustrates the performance variation across different angles of view. Two distinct target scenarios are considered in the experiment: a hybrid target of living entity and metal and a target representing a purely living target. The observation angles, extending from 0° to 180° at 30° intervals, distinctly affect the detection performance for both scenarios. At the observation angles of 0° and 180°, there is a smaller amount of data available, leading to a minor decrease in performance. Conversely, the accuracy exhibits a proportional increase with the observation angle, culminating at a peak at 90°. This peak is attributable to the optimal target visibility at this angle, which maximizes data capture for the millimeter-wave radar and camera systems, thereby optimizing the mAP. Importantly, despite these variations, the mAP consistently remains above the 97% benchmark throughout our experiments, underlining the system's robustness and reliability across different viewing angles.

Impact of different road scenes. In the actual road scene of vehicle driving, the radar data will be significantly affected by the road conditions, the distribution of trees, and street lights on both sides of the road. As shown in Table 2, the detailed road environment information is collected in this article.

In this article, different numbers of living targets are experimentally evaluated in four road scenarios. The results show that the mAP of the proposed method is 97.4%, 97.3%, 97.1%, and 97.7% in four scenarios, respectively. The experimental results show that the proposed method based on RCS is adaptive to the change of road scene and environmental heterogeneity, because the influence of environmental noise on radar SNR is small, and the preprocessing algorithm proposed in this article removes static reflection points in different environments.

Impact of sunlight condition. Figure 17 demonstrates the efficacy of the proposed target liveness detection method functioning across a spectrum of illumination conditions. It emphasizes the mAP

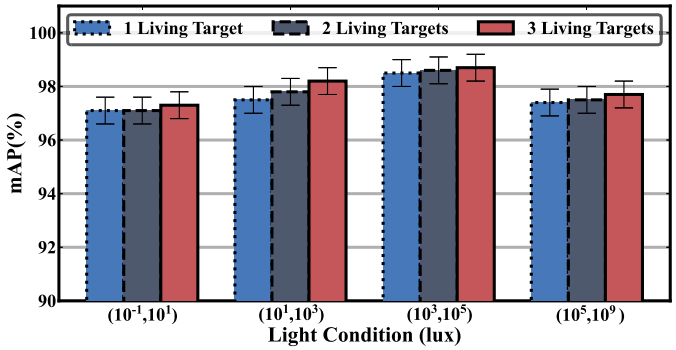


Fig. 17. Light condition.

Table 3. Differences of Sunlight Conditions

Scene	Experimental Condition	Targets' Lighting Condition
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

as a metric for identifying one to three live targets. As the ambient illumination transitions from low-light to highly lit environments, a corresponding increase in the system's precision is noted. This is exemplified when the light intensity escalates from a minimal 0.1 lux to a significantly brighter 10^5 lux, catalyzing an enhancement in the mAP from 97.1% to 98.7%. This progression is attributed to abundant visual information available under high-illumination scenarios. However, over-intense lighting induces interference, primarily due to the mirror-reflection effect. This interference becomes evident when the light intensity exceeds 10^5 lux, leading to a drop in the mAP to 97.7%. Despite these variations in illumination, the proposed algorithm exhibits remarkable robustness, consistently maintaining an accuracy level above 97%. This underlines the liveness target detection algorithm's ability to uphold high precision across a wide spectrum of lighting conditions.

In addition, to better evaluate the model's generalization ability, this article also conducts an experimental evaluation on scenes with four different lighting conditions, shown in Table 3. For example, in scenes 1 and 2, the camera will be affected by the direct or indirect reflection of sunlight from the target, and the target detection accuracy is low. In scene 4, the blur of the target image caused by weak light also affects the effect of target detection. However, the above experiments show that the mAP of the proposed method reaches more than 97% in all four scenes, which indicates that the proposed method is robust to different light conditions. The result also verifies that millimeter wave radar is not sensitive to ambient weather.

Impact of occlusion condition. The occlusion of both living targets and non-living targets influences detection accuracy. In cases of occlusion among living targets, an increase in the target count inevitably leads to overlaps within the sensing scope, consequently impacting the operational limits of the system. This scenario commonly arises when the radar's sensing range is simultaneously occupied by over 20 human targets, resulting in an occlusion rate that exceeds 60%. In terms of occlusion between living targets and other objects, mmWave radar signals demonstrate the ability to penetrate certain types of occlusions, such as wood, but are virtually incapable of penetrating

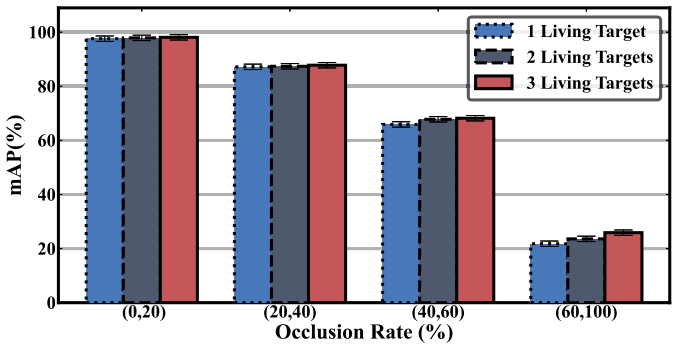


Fig. 18. Occlusion rate.

Table 4. Related Target Detection Approaches

	RCS based	No RCS based
Single-model	[11, 12, 14, 17, 33]	[7, 10, 15, 25–28, 30]
Multi-modal	This Work	[13, 20, 22, 32, 38, 40]

metallic occlusions. To evaluate the system under stringent conditions while maintaining experimental uniformity, we introduce metallic occlusions during occlusion experiments. A metal plate serves as the occluder to obscure the liveness target at varying occlusion rates. Figure 18 presents the mAP accuracy for the detection of one, two, and three living targets amidst varying levels of occlusion. An escalation in occlusion rate results in a diminished quantity of information from both mmWave radar and vision data, thereby causing a decrease in mAP. As occlusion rates range from 0% to 100%, a concomitant decline in mAP scores manifests, highlighting the adverse effect of occlusion on target detection.

9 DISCUSSION

In 2020, 30.6% of fatal crashes were caused by cars colliding with moving objects, e.g., animals, pedestrians, and bikers [2]. The end-to-end liveness target detection methodology proposed in this article is crucial for improving the robustness of autonomous vehicles to prevent accidents caused by visual interference. Instead of proposed pedestrians and bikers, animals and other kinds of liveness have similar RCS features in the mmWave radar data. Therefore, the proposed methodology is suitable for detecting different kinds of living targets and improving the safety of autonomous vehicles.

10 RELATED WORK

Table 4 provides an overview of the current state of target object detection methodologies, categorizing them into two main streams: Single-modal and Multi-modal approaches. The majority of existing techniques that utilize RCS are predominantly single-modal, focusing primarily on target detection. However, these approaches do not extend to the more intricate task of target liveness detection. In contrast, while multi-modal techniques show promise in providing a more holistic view of the scene, they typically do not incorporate RCS and also fall short in addressing target liveness detection in the proposed scenario. Significantly, our work stands out as the first to introduce a multi-modal target liveness detection method that incorporates RCS. This unique contribution bridges the gap in current methodologies, expanding the application of RCS from Single-modal to

Multi-modal detection and extending the scope of detection from mere target identification to the more advanced task of target liveness detection.

10.1 Single-modal Detection

Because the road environment is often very complex in the actual driving scene, achieving high-quality liveness target detection is crucial to the autonomous driving technology of vehicles. Therefore, there has been a lot of research and progress in this regard, and the technology of using various sensors to achieve object detection is relatively very mature. In the aspect of computer vision, there are two commonly used methods to realize liveness target detection. One is based on the contour feature of the face to realize liveness target detection [27], but this method is only used in the case of close faces, so it is not suitable for automatic driving scenes. Another type of method is based on the contour features of the human body to realize the detection of living targets [26, 28], and the implementation effect of this method in the autonomous driving scene is relatively much better. However, these methods do not consider the characteristics of living factors, so although they distinguish people from other non-human objects, they will produce wrong judgments when distinguishing between real people and dummies on metals. However, because millimeter wave radar is not sensitive to the environment, there are also many related studies to detect living targets for vehicular millimeter wave radar. For example, in reference [7, 25], the heat map information of the radar point cloud is input into the neural network to classify different types of objects on the road. In Reference [15], the radar Doppler spectrum was used as the motion feature to realize the distinction between vehicles and living targets. In literature [10, 30], radar point clouds are used for road target classification and vehicle detection.

RCS, a fundamental parameter within the domain of mmWave radar, encapsulates the inherent reflective properties of an object when exposed to radar waves, thus facilitating its application in target detection. The capacity to measure the intensity of the signal reflected by a target, along with the consistent persistence of this reflected intensity for a specific object, emphasizes its reliability. Furthermore, the reflective phenomena of RCS show considerable differences between living and non-living targets, suggesting its potential applicability in liveness target detection. Nevertheless, despite these promising attributes, to the best of our knowledge, there exists no research that leverages RCS for liveness target detection. Existing studies primarily concentrate on utilizing RCS for distinct objectives. For instance, Fu et al. [12] propose an identification and classification technique for unique drone types based on their RCS signatures. They offer a deep learning–based methodology capable of differentiating between various drone models and orientations utilizing a CNN trained on simulated RCS data. Furthermore, Shibao et al. [33] analyze the RCS patterns of diverse types of road debris and put forward a debris detection algorithm that reduces false alarms induced by clutter. However, the utilize of RCS for distinguishing living targets remains uncharted territory. This gap in the current literature highlights the novelty and significance of our research. In this article, we pioneer a novel application of RCS for liveness target detection.

10.2 Multi-modal Detection

Although the research on target detection based on single mode has made much progress, the method based on vision is easily affected by the environment, weather, and object material, and the method based on millimeter wave radar is limited by the sparse and noise characteristics of radar. Therefore, using multi-modal fusion methods to make up for the shortcomings of different sensors has become the mainstream strategy in the field of autonomous driving, including the fusion of millimeter wave radar and camera. Among them, there are two types of fusion of millimeter wave radar and visual image. One is to utilize millimeter wave radar point cloud [22, 32, 40] to fuse with the visual image for target detection, and the other is to utilize Doppler heat map of millimeter

wave radar [20, 38] to fuse with the visual image for target detection. However, these two types of methods do not consider the detection characteristics of liveness. Although they have achieved good results in object detection, they often have false recognition when distinguishing liveness and virtual images. In addition, 3DCNN [13] is the only work that addresses liveness target detection at present, but 3DCNN is a method based on lidar point cloud and visual image fusion, and there is still a problem whereby it cannot distinguish between real and fake when facing virtual human images such as metals.

To solve the problem of liveness target detection in unmanned driving scenarios, this work proposes a detection method using millimeter-wave radar and vision fusion and, for the first time, utilizes RCS as a recognizable feature for liveness target detection. Our experimental results also show that there is a vast difference between the RCS of an active target and its visual interference, such as an LED billboard, due to material, pose, and other differences in the reflection properties. In addition, the RCS values are stable in the detection range of the mmWave radar, which means that the results are not affected by the distance from the target to the radar.

11 CONCLUSION

This article presents an end-to-end liveness target detection methodology with mmWave radar and vision fusion. The proposed method improves the robustness of autonomous vehicles while facing visual interference. The RCS feature extracted from the mmWave radar data is the crucial distinguishing characteristic of the liveness target and the visual interference. We propose a multi-modal feature extraction method to extract features from the mmWave radar and the vision data. Then, we propose multi-modal feature fusion methodologies to fuse these features to distinguish living targets from visual interference. The proposed system in this article improves the robustness of autonomous vehicles by enabling the capabilities of precise liveness target detection.

REFERENCES

[1] Vincent Lepetit, Francesc Moreno-Noguer, and P. Fua. 2009. EPnP: Efficient perspective-n-point camera pose estimation. *Int. J. Comput. Vis.* 81, 2 (2009), 155–166.

[2] Highway Safety. 2022. Facts + Statistics: Highway safety. [Website]. https://www.iii.org/fact-statistic/facts-statistics-highway-safety

[3] 2022. Tesla misidentified the poster as a pedestrian. [Website]. https://xw.qq.com/cmsid/20220309V063U300

[4] 2023. MediaPipe Objectron. [Website]. https://google.github.io/mediapipe/solutions/objectron.html

[5] 2023. MediaPipe Pose. [Website]. https://google.github.io/mediapipe/solutions/pose.html

[6] 2023. Tesla Autopilot's Crash. [Website]. https://edition.cnn.com/2022/10/17/business/tesla-motorcycle-crashes-autopilot/index.html

[7] Aleksandar Angelov, Andrew Robertson, Roderick Murray-Smith, and Francesco Fioranelli. 2018. Practical classification of different moving targets using automotive radar and deep neural networks. *IET Radar, Sonar Nav.* 12, 10 (2018), 1082–1089.

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, WA, 11621–11631.

[9] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 6, 3 (2022), 1–25.

[10] Andreas Danzer, Thomas Griebel, Martin Bach, and Klaus Dietmayer. 2019. 2d car detection in radar data with pointnets. In *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC'19)*. IEEE, Auckland, 61–66.

[11] Chuanwei Ding, Hong Hong, Yu Zou, Hui Chu, Xiaohua Zhu, Francesco Fioranelli, Julien Le Kernec, and Changzhi Li. 2019. Continuous human motion recognition with a dynamic range-doppler trajectory method based on FMCW radar. *IEEE Trans. Geosci. Remote Sens.* 57, 9 (2019), 6821–6831. https://doi.org/10.1109/TGRS.2019.2908758

[12] Rui Fu, Mohammed Abdulhakim Al-Absi, Ki-Hwan Kim, Young-Sil Lee, Ahmed Abdulhakim Al-Absi, and Hoon-Jae Lee. 2021. Deep learning-based drone classification using radar cross section signatures at mmWave frequencies. *IEEE Access* 9 (2021), 161431–161444. https://doi.org/10.1109/ACCESS.2021.3115805

[13] Francisco Gomez-Donoso, Edmanuel Cruz, Miguel Cazorla, Stewart Worrall, and Eduardo Nebot. 2020. Using a 3d cnn for rejecting false positives on pedestrian detection. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'20)*. IEEE, Glasgow, 1–6.

[14] Danièle Hauser, Cédric Tourain, Laura Hermozo, D. Alraddawi, L. Aouf, B. Chapron, A. Dalphinet, L. Delaye, M. Dalila, E. Dormy, F. Gouillon, V. Gressani, A. Grouazel, G. Guitton, R. Husson, A. Mironov, A. Mouche, A. Ollivier, L. Oruba, F. Piras, R. Rodriguez Suquet, P. Schippers, C. Tison, and Ngan Tran. 2021. New observations from the SWIM radar on-board CFOSAT: Instrument validation and ocean wave measurement assessment. *IEEE Trans. Geosci. Remote Sens.* 59, 1 (2021), 5–26. https://doi.org/10.1109/TGRS.2020.2994372

[15] Steffen Heuel and Hermann Rohling. 2011. Two-stage pedestrian classification in automotive radar systems. In *Proceedings of the 12th International Radar Symposium (IRS'11)*. IEEE, Leipzig, 477–484.

[16] Wenchao Jiang, Feng Li, Luoyu Mei, Ruofeng Liu, and Shuai Wang. 2022. VisBLE: Vision-enhanced BLE device tracking. In *Proceedings of the 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON'22)*, IEEE, Glasgow, 217–225. https://doi.org/10.1109/SECON55815.2022.9918581

[17] Joel T. Johnson, Jakov V. Toporkov, Paul A. Hwang, and Jeffrey D. Ouellette. 2023. Efficient calculation of the kirchhoff integral for predicting the bistatic normalized radar cross section of ocean-like surfaces. *IEEE Trans. Geosci. Remote Sens.* 61 (2023), 1–14. https://doi.org/10.1109/TGRS.2023.3268619

[18] Daejun Kang and Dongsuk Kum. 2020. Camera and radar sensor fusion for robust vehicle localization via vehicle part localization. *IEEE Access* 8 (2020), 75223–75236.

[19] Eugene F. Knott, John F. Shaeffer, and Michael T. Tuley. 1985. Radar cross section: Its prediction measurement and reduction.

[20] Teck-Yian Lim, Spencer A. Markowitz, and Minh N. Do. 2021. Radical: A synchronized fmcw radar, depth, imu and rgb camera data dataset with low-level fmcw radar signals. *IEEE J. Select. Top. Sign. Process.* 15, 4 (2021), 941–953.

[21] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Proceeding of the 3rd Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, Vol. 2019, CVPR, Long Beach, CA, 1–4.

[22] Ramin Nabati and Hairong Qi. 2021. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1527–1536.

[23] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in autonomous driving: A survey. *IEEE Trans. Intell. Transport. Syst.* 23, 8 (2022), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865

[24] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Perez. 2021. Carrada dataset: Camera and automotive radar with range-angle-doppler annotations. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. IEEE, 5068–5075.

[25] Kanil Patel, Kilian Rambach, Tristan Visentin, Daniel Rusev, Michael Pfeiffer, and Bin Yang. 2019. Deep learning-based object classification on automotive radar spectra. In *Proceedings of the IEEE Radar Conference (RadarConf'19)*. IEEE, 1–6.

[26] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. arXiv:1804.02767. Retrieved from https://arxiv.org/abs/1804.02767

[27] Yasar Abbas Ur Rehman, Lai Man Po, and Mengyang Liu. 2018. LiveNet: Improving features generalization for face liveness detection using convolution neural networks. *Expert Syst. Appl.* 108 (2018), 159–169.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.). Vol. 28. Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[29] Mark A. Richards. 2014. *Fundamentals of Radar Signal Processing*. McGraw–Hill Education.

[30] Ole Schumann, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. 2018. Semantic segmentation on radar point clouds. In *Proceedings of the 21st International Conference on Information Fusion (FUSION'18)*. IEEE, 2179–2186.

[31] Ole Schumann, Markus Hahn, Nicolas Scheiner, Fabio Weishaupt, Julius F. Tilly, Jürgen Dickmann, and Christian Wöhler. 2021. RadarScenes: A real-world radar point cloud data set for automotive applications. In *Proceedings of the IEEE 24th International Conference on Information Fusion (FUSION'21)*. IEEE, 1–8.

[32] Arindam Sengupta, Atsushi Yoshizawa, and Siyang Cao. 2022. Automatic radar-camera dataset generation for sensor-fusion applications. *IEEE Robot. Autom. Lett.* 7, 2 (2022), 2875–2882.

[33] M. Shibao, K. Uchiyama, and A. Kajiwara. 2019. RCS characteristics of road debris at 79GHz millimeter-wave radar. In *Proceedings of the IEEE Radio and Wireless Symposium (RWS'19)*. 1–4. https://doi.org/10.1109/RWS.2019.8714559

[34] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 145–157.

[35] Pengfei Song, Luoyu Mei, and Han Cheng. 2023. Human semantic segmentation using millimeter-wave radar sparse point clouds. In *Proceedings of the 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD'23)*. 1275–1280. https://doi.org/10.1109/CSCWD57460.2023.10152726

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30, Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[37] Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. 2023. Human parsing with joint learning for dynamic mmwave radar point cloud. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 7, 1 (2023), 22 pages. https://doi.org/10.1145/3580779

[38] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. 2021. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 504–513.

[39] Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. 2021. RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. *IEEE J. Select. Top. Sign. Process.* 15, 4 (2021), 954–967.

[40] Ritu Yadav, Axel Vierling, and Karsten Berns. 2020. Radar+ rgb fusion for robust object detection in autonomous vehicle. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'20)*. IEEE, 1986–1990.

Received 7 February 2023; revised 28 July 2023; accepted 5 October 2023