IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 7, 2021, accepted May 18, 2021, date of publication May 25, 2021, date of current version June 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083503

A Novel Multi-Sensor Fusion Based Object Detection and Recognition Algorithm for Intelligent Assisted Driving

TIANBI LIU[1,5], SHANSHAN DU[2], CHENCHEN LIANG[3], BO ZHANG[4], AND RUI FENG[1,5]

[1]School of Computer Science, Fudan University, Shanghai 200433, China
[2]School of Information Science and Technology, Fudan University, Shanghai 200433, China
[3]School of Electronics and Information Engineering, Tongji University, Shanghai 200082, China
[4]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China
[5]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China

Corresponding author: Rui Feng (fengrui@fudan.edu.cn)

This work was supported in part by the Science and Technology Commission of Shanghai Municipality, Research on Key Technologies of Interpretable Intergeneration of Large-scale Intermodal Sequence Data under Grant 20511100800, and in part by the Science and Technology Commission of Shanghai Municipality, Research on Intelligent Driving Algorithm Governance in Parks and Expressways under Grant 20511101704.

ABSTRACT The object detection and recognition algorithm based on the fusion of millimeter-wave radar and high-definition video data can improve the safety of intelligent-driving vehicles effectively. However, due to the different data modalities of millimeter-wave radar and video, how to fuse the two effectively is the key point. The difficulty lies in the data fusion methods such as insufficient adaptability of image distortion in data alignment and coordinate transformation and also the mismatching of information levels of the data to be fused. To solve the problem of data fusion of millimeter wave radar and video, this paper proposes a decision-level fusion method of millimeter-wave radar and high-definition video data based on angular alignment. Specifically, through the joint calibration and approximate interpolation, projected to polar coordinate system, the radar and the camera are angularly aligned in the horizontal direction. Then objects are detected by a deep neural network model from video data, and combined with those detected by radar to make the joint decision. Finally, object detection and recognition task based on the fusion of the two kinds of data is completed. Theoretical analysis and experimental results indicate that the accuracy of the algorithm based on the two data fusion is superior to that of the single detection and recognition algorithm on the basis of millimeter-wave radar or video data.

INDEX TERMS Intelligent assisted driving, object detection and recognition, multi-sensor fusion, millimeter-wave radar, high-definition video.

I. INTRODUCTION

Automotive driving assistance can significantly facilitate the safety of driving and avoid traffic accidents. Up to now, intelligent vehicles are equipped with sensors such as MMW Radar (millimeter-wave radar), LiDAR and high-definition cameras. With real-time analysis of these sensor data, detecting and recognizing pedestrians, bicycles, motorcycles, cars and other objects outside the car, achieves the purpose of real-time perception of the external environment of the vehicle. However, there are some disadvantages in perception based on a single sensor, for example, due to the large positioning error of the object space position based on visual perception, the position of the object in the real world cannot be accurately estimated. The analysis based on MMW Radar data has insufficient object classification and recognition capabilities, etc. The method based on multi-sensor data fusion can often obtain a more comprehensive object state estimation than a single sensor data, and improve the credibility of the analysis data through the information complementarity of the different sensor data.

Most of the current research on multi-sensor fusion for intelligent driving assistance focuses on the fusion of radar and camera data. The main reason is that the perspective of radar point cloud data in the horizontal and vertical directions is easier to align with the image coordinates. There are shortcomings in radar such as high cost, susceptibility to weather and air impurities, and inability to detect object speed. Data

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Jiang.

fusion based on MMW Radar and camera has high reliability and more extensive application scenarios. The data modalities of MMW Radar and camera images are quite different, which makes it difficult to achieve effective data fusion. Most of the existing methods achieve data alignment between MMW Radar and camera images by projecting to real world coordinates. However, image distortion has a great impact and the mismatch of feature-level data fusion at the information level also limits the advantages of the two sensors.

To solve the problem that the heterogeneous data of MMW Radar and camera is difficult to fuse, this paper presents a fusion algorithm based on MMW Radar and high-definition video for object detection of intelligent driving assistance. First, the MMW Radar and camera are jointly calibrated in the polar coordinate system, and the two types of data are aligned in the horizontal direction through approximate interpolation, which effectively overcomes the influence of image distortion on coordinate transformation. The deep network model is used to implement object detection and recognition based on image data. With the objects detected by radar, the decision-level data fusion task of MMW Radar and image is fulfilled.

This paper is distinguished by the following main contributions:

- Through multi-angle joint calibration, the spatial sparse alignment of the heterogeneous data of MMW Radar and the camera in the common dimension is realized with image distortion ignored.
- A neighboring approximate interpolation method is proposed to achieve the spatial alignment of the heterogeneous data of MMW Radar and camera in the common dimension.
- While multi-camera fusion improving the recall rate of object detection task, the proposed method of decision-level fusion of MMW Radar and camera data removes false positives in the object set and improve the accuracy.

II. RELATED WORK

The data fusion of radar and video images can be divided into pixel-level, feature-level and decision-level fusion according to the degree of abstraction of the information [1]. The position detection results of space objects can be obtained directly by MMW Radar, which are distributed in the horizontal dimension, which is unlikely to achieve pixel-level fusion with image data. Feature-level fusion processes sensor data by extracting features (such as edges, shapes, regions, distances, etc.) and then fusion processing. MMW Radar data, which can directly provide the detection result of the object, cannot achieve feature-level fusion with image data. The only feasible method to implement the fusion of radar and video image data is to analyze the image features combined with radar detection result, and implement the decision-level fusion of the optimal decision based on certain criteria and decision credibility.

Object detection is one of the most fundamental and challenges in computer vision [2]. Traditional digital image processing is a typical method before 2012, such as V-J [3], [4], HOG [5], DPM [6], etc. After 2012, Deep Learning (DL) represented by Convolutional Neural Networks (CNN) has gradually become the mainstream method. Performing the task of object detection by deep learning can be divided into two technical routes: two-stage networks and single-stage ones. The two-stage networks are represented by RCNN [7]–[9], and the single-stage ones are represented by YOLO [10]–[13] and SSD [14], RetinaNet [15], etc. Considering that intelligent driving assistance has extremely high-speed requirements for object detection algorithms, most of strategies use single-stage algorithms with efficiency advantages. Especially the newly appeared YOLO V5 algorithm performs best in terms of efficiency and performance balance, which is more suitable for object detection in intelligent driving assistance.

Since point cloud data is similar in distribution to the observation angle of video images, most researches on object detection based on LiDAR data mostly instead of radar. Many scholars have successively proposed similar image object detection algorithms [16]–[21], as well as VoxelNet [22], BirdNet [23], PointNet [24], [25], StarNet [26] and other object detection models for point cloud data. However, LiDAR has a large amount of output data, a high price, but poor adaptability in weather. Applied in assisted driving, this paper focuses on the data fusion of camera and MMW Radar instead of LiDAR considering the cost issue and the limited resources of edge computing. Unlike LiDAR, MMW Radar generally does not provide point cloud data, but directly offers the relative position and relative speed of the object. With poor accuracy of object recognition, most of MMW Radar cannot achieve object classification.

In terms of MMW Radar and image data fusion, whether the two can be aligned in the dimensions of time and space is the key to data fusion. The data output rate of mainstream MMW Radar is 20fps, and the image data of the camera is 25∼30fps. The time difference between the two is less than 40ms, which meets the requirements of time alignment. Spatial alignment can be achieved through calibration, and data correspondence conversion can be realized through coordinate transformation.

Zhai *et al.* [27] projected the positions of the objects detected by radar to the image data through joint calibration, and proposed a method to generate the area of interest of the objects in radar data, which completed obstacle detection based on the fusion of MMW Radar and camera images data. However, the algorithm has visualized radar information only by aligning with the image position, and the image information is not fully extracted and utilized without object detection work. Similarly, Zhiqiang *et al.* [28] first selected the objects using radar data, then established the area of interest based on the image processing method and judged whether it was a vehicle obstacle, and implemented obstacle detection through information fusion based on joint Kalman

filtering. Bi *et al.* [29] used coordinate transformation to map radar object information to image data, and then used HOG and SVM classifiers to detect the sliding window based on the radar data and the dynamic area generated by the sliding window, and implemented object recognition through matching. Zhaowei *et al.* [30] established a multi-sensor-based coordinate transformation model, mapped the depth information given by the radar to the image data, extracted the gradient histogram of the region of interest, and used SVM to achieve pedestrian detection. Lisheng *et al.* [31] also used coordinate transformation to project radar data on the image data to form a region of interest, and used image processing methods to reduce interference points. However, this method only extracts the characteristics of the taillights of the vehicle without object recognition and can only be applied at night. The above fusion strategies all use the spatial information in the radar data to reduce a lot of interference for image feature analysis, but the image data analysis is generally not deep enough, and the object recognition performance is not satisfactory.

In the case of object detection and recognition based on video image data, it can complete the decision-level data fusion with radar. Yuan *et al.* [32] implemented vehicle objects detection by searching for vehicle shadows, and then converted the radar data to the image coordinate system to verify the matching relationship. With insufficient accuracy in searching for vehicle objects, radar data is weak in object recognition, and it is difficult to integrate the advantages of the two types of data to improve the overall recognition performance. Aziz *et al.* [33] proposed a method of using 3D-CNN+LSTM to do MIMO radar data analysis, using YOLO algorithm to implement image object detection, and then using projection transformation to achieve result fusion. Because MIMO radar provides two-dimensional spatial data, it can implement object detection through convolutional neural networks. However, MIMO radar is difficult to implement joint calibration of image data, and fusion is prone to deviation in the decision-making stage, and the fusion calculation is large.

Nobis *et al.* [34] presented a fusion architecture of radar and video data based on deep learning object detection. The radar data is processed into independent channels that match the image data, and fused with the image RGB channel data, and the object detection and recognition is achieved through the deep learning model. But, it adds noise to the image data through using multi-period radar data, and this fusion strategy destroys the time alignment of the two sensor data.

Mureșan *et al.* [37] proposed a multimodality fusion framework which can flexibly accomplish both object detection and end-to-end driving policy for prediction of steering angle and speed. It had some enlightenment for the heterogeneous data fusion of MMW Radar and camera, but had no further discussion. Nie *et al.* [38] proposed a fusion architecture by applying a combination of two types of sensor data fusion methods [37] (a model-based approach using the Unscented Kalman Filter and a data-driven approach using a single-layer perceptron) to combine three-focus camera, LiDAR, and millimeter wave radar for data fusion. For the data fusion of millimeter-wave radar, it initially considered the mapping relationship of the object position in the polar coordinate system, but lacked the explanation of the method of spatial alignment, and did not pay attention to the image distortion and the noise of MMW Radar data.

In summary, although lots of scholars have done a lot of work and made some progress in the fusion of radar and image data, there are still many problems to be solved. The core issue is how to implement effective joint calibration and how to overcome the impact of image distortion.

III. MULTI-SENSOR FUSION-BASED OBJECT DETECTION AND RECOGNITION

In this paper, we propose an object detection and recognition method that fuses the information of MMW Radar and video data. Specifically, the camera and MMW Radar are jointly calibrated with multiple key points to achieve spatial alignment, and such two types of data are transformed into polar coordinates. We combine the visual perception algorithm and the object spatial position detection information of MMW Radar to achieve multi-sensor data fusion at strategy-level, that is also a model-based fusion [37], thereby improving the accuracy of object detection and recognition. The fusion algorithm framework is shown in Figure 1.

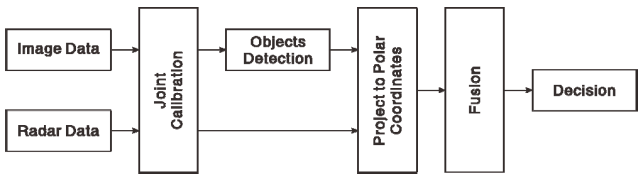


FIGURE 1. The framework of millimeter wave radar and image data fusion algorithm. The data of millimeter wave radar is an object set which is similar to image's detection result.

The joint calibration of image and MMW Radar data requires the spatial alignment between them, i.e., the image pixel data and radar data in the coordinate system are one-to-one correspondence. First, the radar can only provide the object detection results in the horizontal direction while the image pixels are distributed in the horizontal (X-axis) and vertical (Y-axis) dimensions. Hence, the radar data and the image can only be aligned in the horizontal direction. Secondly, the image data lacks depth information, and there is no polar diameter value in the polar coordinate system, so only polar angles are aligned.

In the common viewing range of the camera and MMW Radar, uniformly spaced angles are used for joint calibration. Leveraging the data form of polar angle and polar diameter from the MMW Radar, the horizontal pixel coordinates of the image are acquired through the look-up table to obtain the angle range, and the polar angle is calculated through interpolation. Compared with coordinate transformation, the comprehensive look-up table and interpolation algorithm is more efficient.

After the spatial alignment between the two types of data is accomplished, the object detection and recognition results based on the image and radar data are matched and fused in space, and the final results are comprehensively analyzed to improve the performance of object detection and recognition.

The proposed method avoids the complicated calculation of transforming from the image coordinate system to the world coordinate system. Through joint calibration, the image coordinates are directly mapped to the coordinate system matching the radar data. The proposed method can adapt to different types of cameras and can effectively overcome the differences in camera optical properties and images. In addition, in order to take both the view range and image resolution into account, the intelligent auxiliary drive system usually deploys multiple cameras at the same time, such as normal, wide-angle, telephoto, etc. The proposed algorithm in this paper is compatible with multiple cameras, which realizes effective data fusion of multi-channel video and MMW Radar.

A. SPATIAL ALIGNMENT OF MULTI-SENSOR DATA

Considering the linear propagation properties of electromagnetic waves and light through the air, the MMW Radar and the camera have a consistent view of the real world, despite the fact that they work on different principles, and that external information enters the sensor by converging on a single point.

The working coverage angle of the millimeter wave radar and the camera is shown in Figure 2. Both sensors have a similar signal input field of view. The MMW Radar uses the horizontal plane to estimate the reflected signal angle, known as the (AoA). The coverage of the camera lens is called Field of View (FOV). Objects at the same angle will be obscured by objects with a small polar diameter against objects with a large polar diameter. The resolution of the raw data received by the sensor actually reflects the ability to resolve real-world angles.

The image pixel positions can be mapped back to angular positions, and the MMW Radar data contains angular information about the objects. Thus, the polar coordinates can be used as a reference to achieve spatial alignment of the two kinds of data.

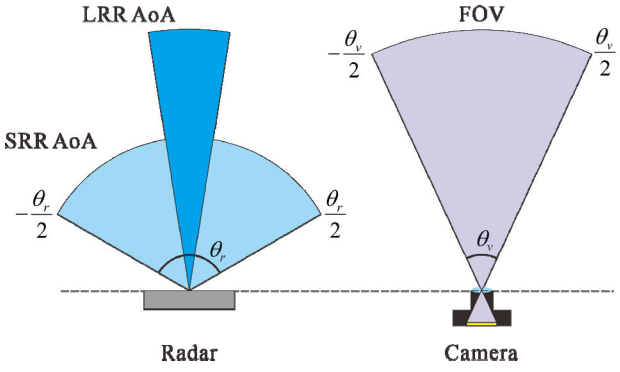


FIGURE 2. Schematic diagram of the working coverage angle of the MMW Radar and camera. The angle of arrival (AoA) of short-range radar(SRR) is θ_r which we concern about.

B. JOINT CALIBRATION ALGORITHM OF IMAGE AND RADAR DATA

Image data generally suffers from mirror image distortion and tangential distortion [34]. The real-world industry often corrects for distortion through algorithms to improve the imaging of the camera. Considering that most intelligent assisted driving includes a variety of camera types such as telephoto and wide-angle, it is difficult for aberrated and corrected images to be accurately mapped back to the objective world's angular coordinates through theoretical calculations. However, for both distorted and corrected images, the angular coordinates of objects in the camera's field of view that are located at the same polar angle are always the same. Therefore, the image pixel coordinates can be matched with the objective angle coordinates through calibration.

The camera lens and the MMW Radar receiver are placed in the same reference vertical line to ensure that the field of view origins of the multiple sensors coincide. N positions are calibrated in even angular intervals in front of the sensor via a protractor, and the camera and radar are calibrated at the same time through the radar corner reflector data.

The camera is fixed to the radar and the camera's vertical view needs to cover the radar view in the light of the larger vertical view for the camera than radar. A schematic overhead view of the joint calibration is shown in Figure 3. The marked position is used to place the radar corner reflector to realize the joint calibration of multiple sensors.

The joint calibration starts with aligning the 0° positions of all sensors, as shown in Figure 4. The visualization of the data from the radar and the image enables a fine adjustment of the camera and radar positions. It ensures that the central pixel of the image coincides with the radar corner reflector support bar and that the 0° coordinate line of the radar coincides with the signal position of the corner reflector.

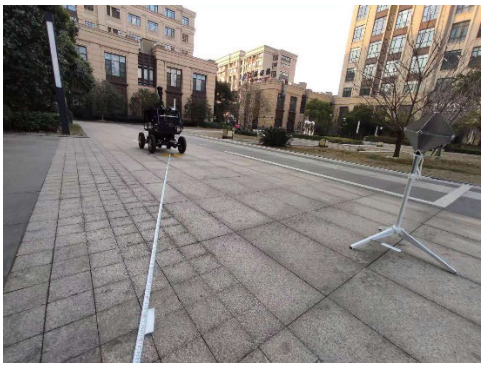
Considering the different FOV ranges of different types of cameras, generally the horizontal view is smaller than the MMW Radar, and the calibration range needs to be selected as the overlap of the two, i.e., the horizontal view range of the camera. If the horizontal view of the camera is larger than the MMW Radar, such as fisheye cameras, the radar view range can be selected for calibration.

Without loss of generality, an even number (N) of equal angle regions is selected in the horizontal direction, giving a total of $N + 1$ calibration positions for the horizontal view of the image. By placing the radar angle reflector at the calibration position, the set of radar calibration angles can be defined as:

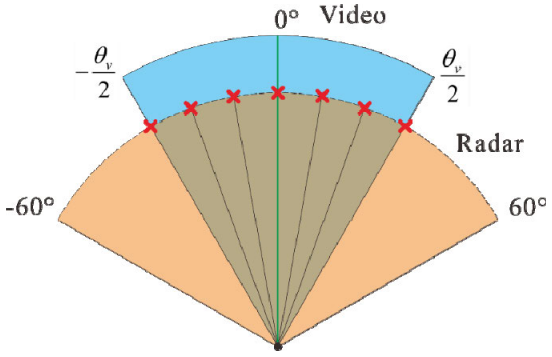
$$\Omega = \{\alpha_1, \alpha_2, \ldots, \alpha_N, \alpha_{N+1}\} \tag{1}$$

Determining the image pixel positions of the radar corner reflector support rod, the set of horizontal positions of the camera calibration pixels can be defined as:

$$X = \{x_1, x_2, \ldots, x_N, x_{N+1}\} \tag{2}$$

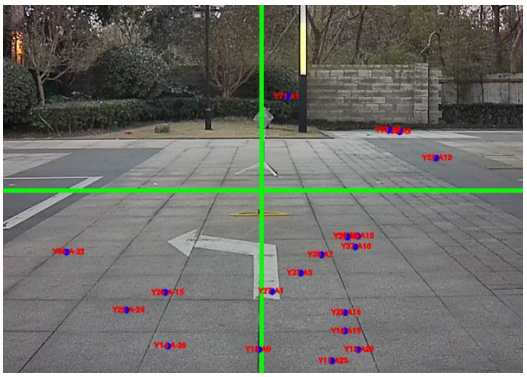


(a) Calibration using a radar angle reflector

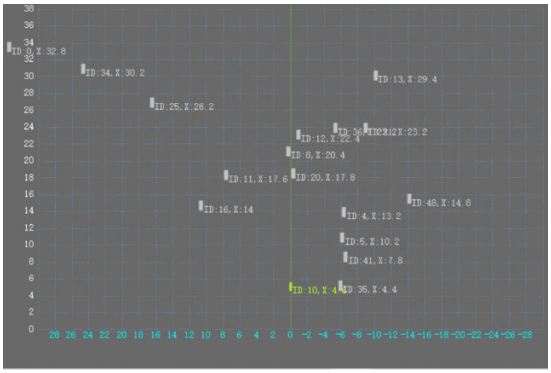


(b) Illustration of the key position of calibration

FIGURE 3. Schematic diagram of joint calibration. We set several key angles to calibrate camera and radar simultaneously.



(a) 0° position calibration of the image



(b) 0° position calibration of the radar

FIGURE 4. Joint calibration of 0° position. The green vertical line in (a) means the middle of the image which corresponds to 0°. The line must be aligned with the angle reflector. In (b) the green point lying on 0° is the signal reflected from the angle reflector.

where the X axis coordinate origin is defined at the center of the image, i.e., $x_{\frac{N}{2}+1} = 0$. The X coordinate on the left is negative and the right is positive.

Define the horizontal view of the camera as θ_v and the image resolution as $H \times W$, where H is the number of vertical pixels and W is the number of horizontal pixels, then the following is obtained.

$$\begin{cases} \theta_v = \alpha_{N+1} - \alpha_1 \\ W = x_{N+1} - x_1 \end{cases} \tag{3}$$

Since the elements in X and Ω correspond one-to-one, the X coordinate of any pixel can be determined by the look-up table method to determine its angular range, and the X coordinate transformation of the pixel uses correlation to perform approximate linear interpolation.

For aberrated images, where the degree of distortion varies continuously, the angle at which a single pixel is located can be defined as the pixel angular density ρ, and then the angular density of a pixel at any location is near to its adjacent location. The polar angle corresponding to a pixel can be set as:

$$\alpha_x = \int_0^x \rho(x)dx \tag{4}$$

Given the unknown nature of image distortion and aberration correction, $\rho(x)$ can be calculated by interpolating the pixels in this region using the angular density of the neighboring region pixels. When $x_{n-1} < x < x_n$, it meets that

$$\alpha_x = \alpha_{n-1} + \int_{x_{n-1}}^{x_n} \rho_n(x)dx \tag{5}$$

Define the angular interval between the calibration points is $\Delta\alpha$. Then,

$$\overline{\rho_n} = \frac{\Delta\alpha}{x_{n+1} - x_n} \tag{6}$$

Assuming that the angular density of pixels within $(x_{n-1}, x_n]$ varies uniformly, then

$$\rho_n(x) = \overline{\rho_{n-1}} + \frac{\overline{\rho_{n+1}} - \overline{\rho_{n-1}}}{x_{n+1} - x_n}(x - x_n) \tag{7}$$

In particular, the left and right edges of the image meet that

$$\rho_1(x) = \overline{\rho_1} + \frac{\overline{\rho_2} - \overline{\rho_1}}{x_2 - x_1}(x - x_1) \tag{8}$$

$$\rho_N(x) = \overline{\rho_{N-1}} + \frac{\overline{\rho_N} - \overline{\rho_{N-1}}}{x_{n+1} - x_n}(x - x_n) \tag{9}$$

Substituting equation (7) into equation (5) enables the transformation from X-coordinate to angular coordinate of any pixel.

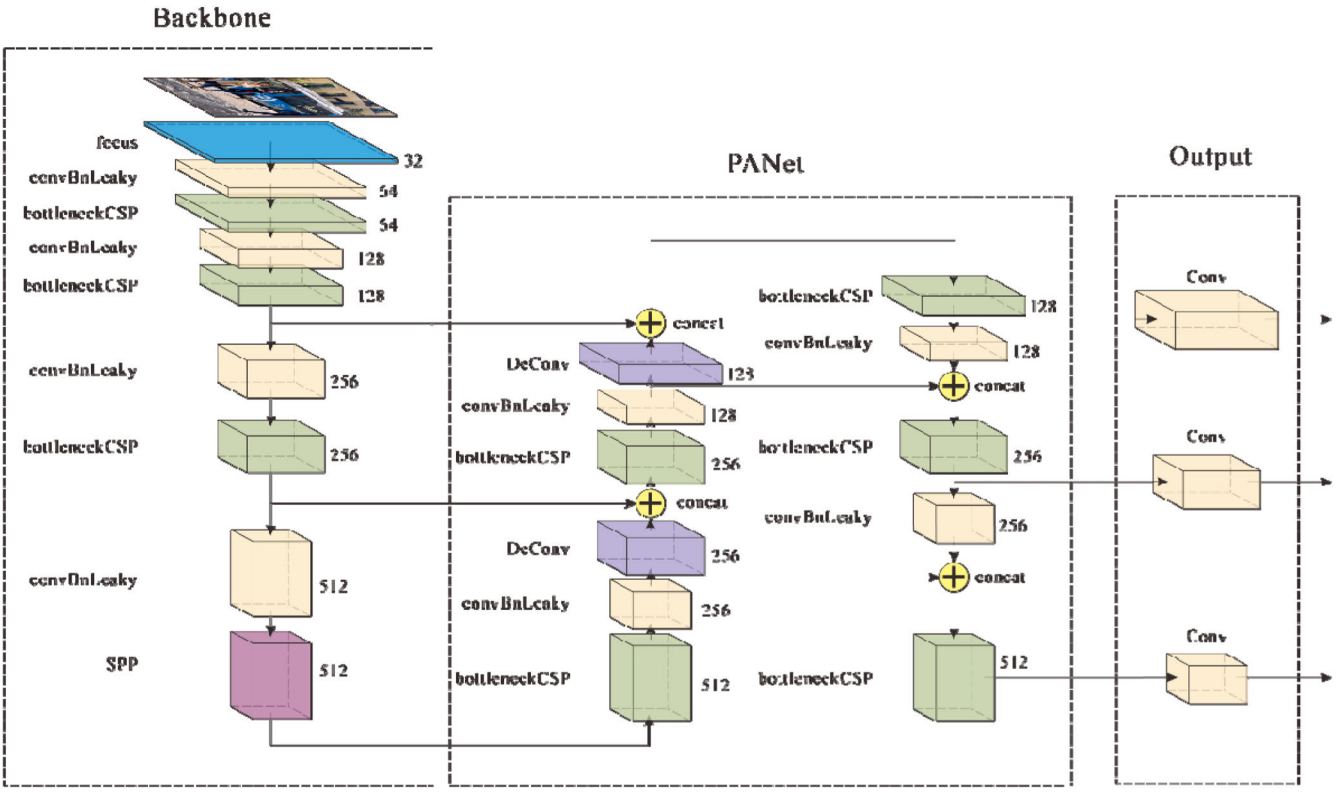


FIGURE 5. The structure of YOLOv5: the overview of YOLOv5 according to the open source code and description.

The approximate coordinate mapping algorithm, based on table look-up and interpolation, is well suited to a wide range of image distortions and distortion-corrected images. To improve the accuracy of the coordinate transformation, it is sufficient to reduce the calibration position interval and set more calibration points.

C. IMAGE DATA-BASED OBJECT DETECTION AND RECOGNITION

The object detection and recognition algorithm in this paper uses the single-stage YOLO V5 [9]–[12], whose network structure framework is shown in Figure 5.

The YOLOV5 network includes 4 main substructures, namely convBnLeaky, bottleneck, bottleneckCSP, and SPP, which can be used as a predefined substructure and used in the entire network. The network substructure is shown in Figure 6.

The input of the network model is a single frame of video data, and the detection and recognition results of the objects in the screen can be obtained by performing an inference operation. After the model is trained, it can realize the dynamic real-time detection and recognition of pedestrians, motor vehicles, non-motor vehicles and other objects. The algorithm supports parallel object detection and recognition of multiple video data. The video frame data of multiple cameras are combined into a batch data input model. The object detection results of all images can be obtained by performing one inference. Simultaneous analysis of multiple channels of video data will consume more video memory, but will not affect the inference speed.

D. JOINT CALIBRATION ALGORITHM OF IMAGE AND RADAR DATA

Object detection can be achieved based on both video and radar data. Between them, the object perception ability based on video data is stronger. Especially in the case of multiple cameras' cooperative sensing, the detection and recognition performance has reached a high level. Due to the lack of spatial location information, the result of video perception is difficult to support the judgment of whether the object is valid. Based on this, the decision will affect the vehicle. Therefore, it is necessary to integrate radar detection results for decision fusion. The fusion strategy framework is shown in Figure 7.

If multiple camera models are used, the integration between video sensing data should be prioritized. Generally multi-vision cameras have achieved view center alignment between cameras and different camera data can be mapped to the field of view by image alignment algorithms, which are not described in detail due to space constraints.

Object detection and recognition results based on multiple image data can be mapped to the same image coordinate system. Duplicate objects are filtered out by non-maximum suppression, and the final results is the union of all detection

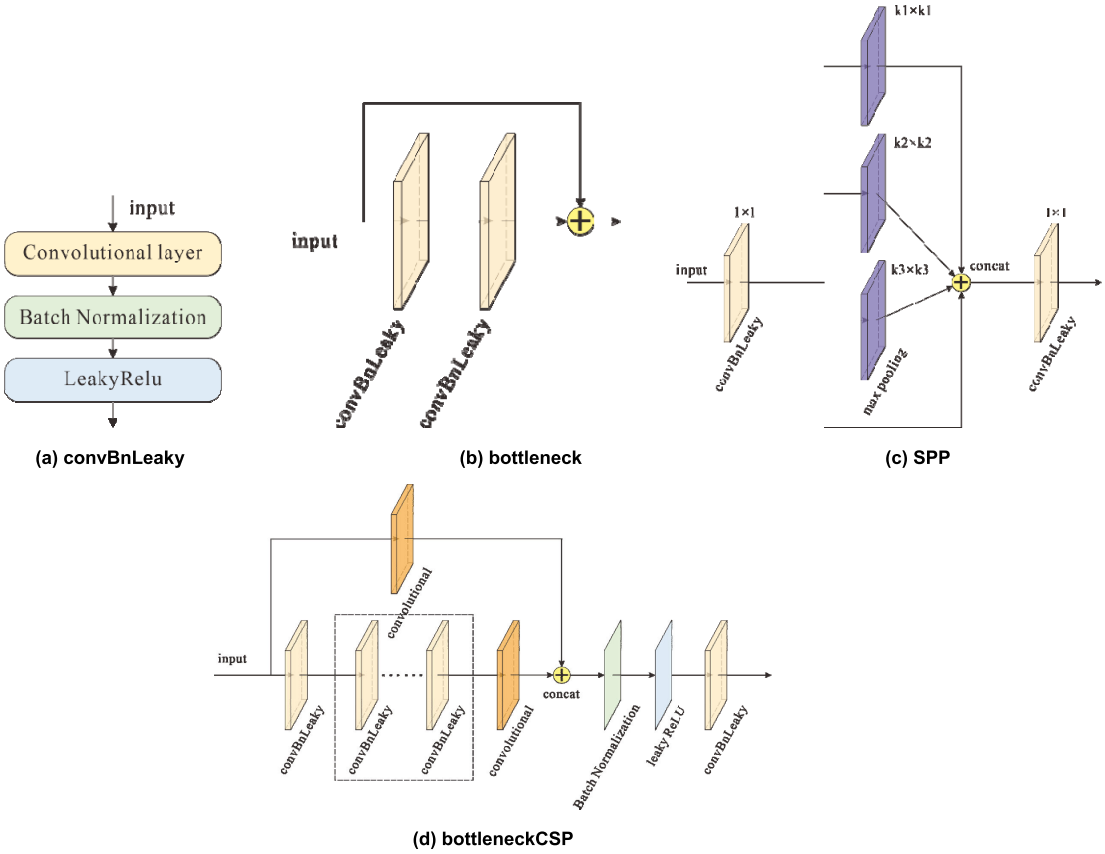


FIGURE 6. The structures of YOLOv5 substructures.

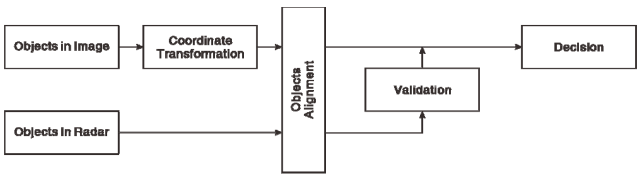


FIGURE 7. Decision fusion strategy framework based on image and MMW Radar data. Since the data of Radar is the result of the detected objects, it is directly aligned with the objects in the image. Based on the objects set in image, the Radar data is used for Validation.

and recognition results, that is, the image object set in Figure 7. Define the set of object detection and recognition based on image data as O_v. If M cameras are used to achieve object detection and recognition, then it can obtain that

$$O_v = O_{v_1} \cup O_{v_2} \ldots \cup O_{v_M} \tag{10}$$

The position and size of the object detection and recognition are mapped to the polar coordinate system through coordinate transformation, and the viewing angle range of each object can be obtained. In the polar coordinate system, the image detection object set O_v and the radar detection object set O_r are aligned and fused, and the decision steps to obtain the detection result are as follows:

Step 1: Obtain the actual location information of a single object

For the object o_{vi} detected based on image data, the viewing angle range is $[\theta_1, \theta_2]$. Given that $o_r \in O_r$, φ is the polar angle of o_r. It can be obtained the radar detection object set O_r' corresponding to o_{vi}.

$$O_r' = \{O_r | \varphi \in [\theta_1, \theta_2]\} \tag{11}$$

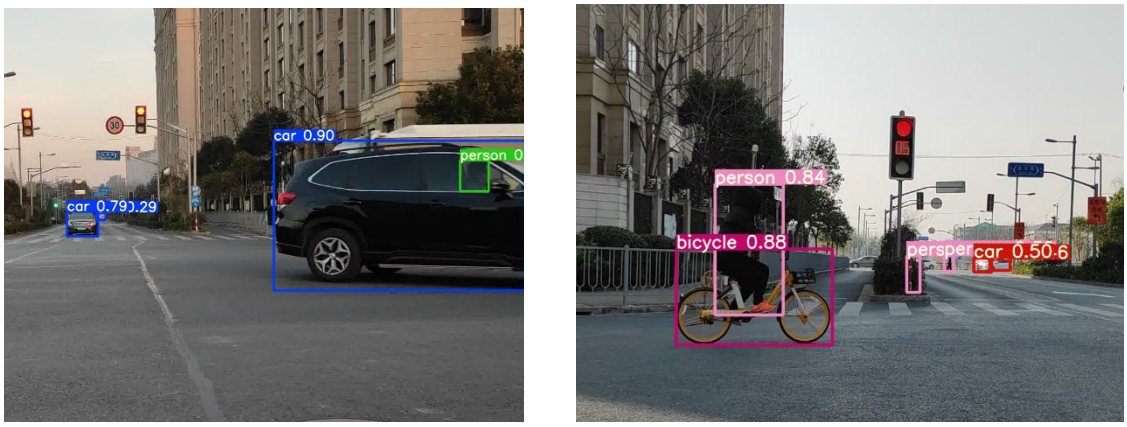
Considering the diffraction characteristics of the radar signal, the object within the viewing angle range may be hollow or small in size, so there may be multiple objects with different polar diameters in the O_r' set. In addition, the radar may misidentify a single object that is close to multiple objects. Therefore, it is necessary to cluster O_r' according to the possible size of the object, and select the object closest to the vehicle to match with o_{vi}. The threshold d_0 can be set according to the actual size of the object in the real world. We select the object $o_{r_{nearest}}$ with the smallest extreme diameter value in $O_r{}'$, and filter the distant objects. Given that (x_r, y_r) is the actual radar coordinates of o_r, then the distance between o_r and $o_{r_{Nearest}}$ can be defined as:

$$d = \sqrt{(x_r - x_{r_{nearest}})^2 + (y_r - y_{r_{nearest}})^2} \tag{12}$$

The object set that matches o_{vi} is:

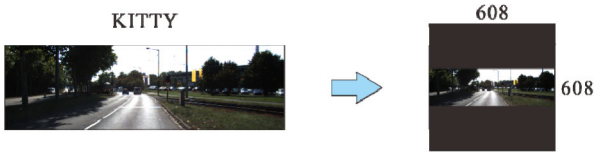
$$O_r{}'' = \{O_r{}' | d < d_0\} \tag{13}$$

Considering that $O_r{}'' = \{o_1{}'', o_2{}'', \ldots, o_m{}''\}$, ρ'' is the polar diameter of o'', and φ'' is the polar angle of o''.



(a) vehicle and person can be detected simultaneously (b) bicycle and person can be detected respectively

FIGURE 8. Examples of object analysis based on data fusion. In (a) the person sitting in the car was detected, but it is unnecessary because the car object is more important. In (b) the person and the bicycle were all detected, but actually they should be regarded as one object.



(a) Image preprocessing on KITTY dataset

(b) Normal 16:9 image preprocessing

FIGURE 9. Input image preprocessing in YOLOv5 training procedure. As shown in (a), after resized with the same aspect ratio, the images in KITTY contribute fewer pixel to YOLOv5 model in contrast to normal 16:9 images shown as in (b).

If m > 0, the position data of o_{vi} is

$$\begin{cases} \rho_{vi} = \dfrac{1}{m}\sum_{j=1}^{m} \rho_j'' \\ \varphi_{vi} = \dfrac{1}{m}\sum_{j=1}^{m} \varphi_j'' \end{cases} \tag{14}$$

The relative position of o_{vi} and the vehicle on the ground satisfies that

$$\begin{cases} x_{vi} = \rho_{vi}\cos\varphi_{vi} \\ y_{vi} = \rho_{vi}\sin\varphi_{vi} \end{cases} \tag{15}$$

step2: validation of the object

For an object o_{vi} detected on the basis of image data, which corresponds to $O_r'' = \varnothing$, i.e. when $m = 0$. It means that the radar does not detect an object in the o_{vi} view range, o_{vi} should be a false positive object.

According to the radar detection data, rules can be set for the detection area, and the positioning and speed information of the object in the radar data can be used to define business rules for intelligent assisted driving. Objects outside the rules can be regarded as invalid objects. For a single object o_r detected by radar, if it does not appear in the O_v coverage angle of view, the object does not belong to the range of object types detected and recognized and can be considered invalid.

IV. EXPERIMENTS AND ANALYSIS

The MMW Radar and camera used in our experiments are set up on the vehicle and face straight ahead. We adopt the camera with a FOV of 60° and the MMW Radar of German Mainland ARS 408-21.

In the experiment, the urban highway is set as the scenario source of test data, and 455 scenes of different time and place are extracted from the real video to evaluate the performance of object detection and recognition based on multi-sensor fusion method. We adopt YOLOv5 s and YOLOv5 l algorithms for object detection and recognition on video image data, which are denoted as 5s and 5l, respectively. The pre-trained model is based on coco [35] dataset, including six types of objects, i.e., pedestrian, bicycle, motorcycle, car, bus and truck.

We set two thresholds, i.e., 50% and 75%, for the ProbExists of MMW Radar hardware parameters, which indicates the object is detected when its detection signal ProbExists exceeds the threshold.

The experimental results are shown in Table 1. Due to the data characteristics of MMW Radar, the calculation of recall rate and accuracy rate only consider the existence of object but without recognition. In addition, we set the classification accuracy as a single metric for ARS 408-21 MMW Radar.

As the MMW Radar can detect all objects within the visual angle range, the recall rate is close to 100%. However, it also introduces a large number of false positives, thus it is unable to complete the task of object detection and recognition. Through the fusion of video and radar data, the accuracy

TABLE 1. The performance of the multi-sensor fusion algorithm for object detection and recognition.

Sensor Type	Settings	Recall	Accuracy	FPR	Classification Accuracy
Camera	5s	86.50%	98.30%	1.49%	92.60%
	5l	92.80%	99.60%	0.34%	96.40%
MMW Radar	Threshold 0.5	99% (no classification)	17.7% (no classification)	82.30%	11.20%
	Threshold 0.75	97.7% (no classification)	22.7% (no classification)	77.30%	14.50%
Multi-sensor fusion	5s+threshold 0.5	86.50%	99.10%	0.76%	92.90%
	5l+threshold 0.5	92.80%	99.80%	0.18%	96.40%
	5s+threshold 0.75	86.50%	99.30%	0.61%	93.00%
	5l+threshold 0.75	92.80%	99.90%	0.12%	96.50%

TABLE 2. The performance of the multi-camera and MMW radar fusion algorithm for object detection and recognition.

Sensor Type	Settings	Recall	Accuracy	FPR	Classification Accuracy
Multi-camera fusion	5s	87.30%	98.40%	1.49%	92.80%
	5l	93.20%	99.60%	0.34%	96.50%
MMW Radar	Threshold 0.5	99% (no classification)	17.7% (no classification)	82.30%	11.20%
	Threshold 0.75	97.7% (no classification)	22.7% (no classification)	77.30%	14.50%
Multi-sensor fusion	5s+ threshold 0.5	87.30%	99.20%	0.76%	93.10%
	5l+threshold 0.5	93.20%	99.80%	0.22%	96.60%
	5s+threshold 0.75	87.30%	99.30%	0.61%	93.20%
	5l+threshold 0.75	93.20%	99.90%	0.12%	96.60%

of object detection is greatly improved, the false-positive problem of the two types of sensors is alleviated, and the accuracy of object recognition is also improved.

The object detection based on radar and image data fusion is more helpful for assistant driving. For example, in our experiments, the image detection and recognition algorithm can simultaneously detect the vehicle and the person in the vehicle, and the bicycle and rider respectively, which can effectively remedy the defect that the radar "binds" two objects into one object.

As shown in Table 1, our proposed fusion algorithm does not improve the recall rate because most undetected objects are distant and thus have small scales. As the MMW Radar is poor at recognizing objects, we cannot simply add it to the detection result set. In the experiments, we found that birds, leaves, and remaining stuff on the ground become the detected results of the radar, i.e., obstruction objects. Exploiting multi-camera can effectively boost the recall rate. By adding a telephoto camera with a FOV of 20°, we first fuse the detection results based on image data, and then make comparison with the fusion results of radar and image data. The experimental results can be found in Table 2.

As demonstrated in Table 2, the image data algorithm based on multi-camera can boost the recall rate. The multi-sensor fusion method also achieves an improved recall rate. Besides, our image data algorithm is pre-trained on coco dataset rather than the driving scenes. The performance of our approach is expected to increase when applied to real-scene datasets. For instance, we can achieve the detection result for the class 'cyclist' in KITTY dataset [36] instead of the separated detection results, i.e., 'pedestrian' and 'bicycle'.

TABLE 3. The effects of YOLO v5 trained by different dataset.

Dataset	Accuracy	Recall
COCO	98.40%	87.30%
KITTY	96.50%	74.20%

We experimentally find that YOLOv5 achieves inferior performance on KITTY dataset. The model also obtains a low recall rate of 74.2% on our collected data. After detail analysis, we find that the ratio of image width and image height is 10:3 in KITTY dataset. For YOLOv5, the input data is resized with equal ratio to the size of 608, which is 608×182 pixels in our case, leading to the lack of useful information in many tiny objects.

In the experiments, we use the collected real-scene data. The training set contains 6 categories, i.e. vehicle, bus, truck, pedestrian, sitting person, and cyclists. We re-train the YOLOv5 model and the experimental results over 455 scenes are reported in Table 4.

Table 5 shows the results after adding a telephoto camera and performing data fusion on the binocular camera and the MMV Radar.

TABLE 4. The performance of the multi-sensor fusion algorithm for object detection and recognition on our constructed dataset.

Sensor Type	Settings	Recall	Accuracy	FPR	Classification Accuracy
Camera	5s	89.30%	97.50%	3.10%	92.50%
	5l	90.40%	97.80%	2.83%	93.20%
MMW Radar	Threshold 0.5	99% (no classification)	17.7% (no classification)	82.30%	11.20%
	Threshold 0.75	97.7% (no classification)	22.7% (no classification)	77.30%	14.50%
Multi-sensor fusion	5s+ threshold 0.5	89.30%	98.70%	1.58%	93.10%
	5l+threshold 0.5	90.40%	98.60%	1.71%	93.70%
	5s+threshold 0.75	89.30%	99.00%	1.15%	93.30%
	5l+threshold 0.75	90.40%	99.10%	1.15%	93.90%

TABLE 5. The performance of the multi-camera and MMW radar fusion algorithm for object detection and recognition on our constructed dataset.

Sensor Type	Settings	Recall	Accuracy	FPR	Classification Accuracy
Multi-camera fusion	5s	94.30%	97.70%	3.26%	95.30%
	5l	94.50%	97.90%	2.83%	95.60%
MMW Radar	Threshold 0.5	99.00% (no classification)	17.70% (no classification)	82.30%	11.20%
	Threshold 0.75	97.70% (no classification)	22.70% (no classification)	77.30%	14.50%
Multi-sensor fusion	5s+ threshold 0.5	94.30%	98.30%	2.39%	95.70%
	5l+threshold 0.5	94.50%	98.70%	1.72%	96.10%
	5s+threshold 0.75	94.30%	98.70%	1.81%	95.90%
	5l+threshold 0.75	94.50%	99.10%	1.15%	96.30%

As observed from the experimental results, multi-sensor-based data fusion algorithm outperforms all single-sensor algorithms on object detection and recognition.

Excluding the decoding time of the video frames, the speed of the algorithm on the target computing platform NVIDIA Jetson TX2 based on the TensorRT framework reaches 28.17fps with batch size = 1 and 26.31bps with batch size = 2. It is enough for real-time applications on edge devices.

V. CONCLUSION

In this paper, an object detection and recognition algorithm based on multi-sensor fusion for intelligent driving assistance has been proposed. Specifically, on the basis of the latest research, we took use of MMW Radar and camera to perform coordinate transformation so as to achieve spatial alignment of the heterogeneous data of the two in the polar coordinate system by multi-angle joint calibration and neighboring approximate interpolation. Besides, the detection and recognition results of radar and image can be achieved at decision-making level. The proposed method is a strong competitor to any single sensor, because the experiments and analysis showed that this approach of multi-sensor fusion has a significant increase in the accuracy of the object detection and recognition.

It can also use more other image-based methods instead of YOLOv5, such as SSD [14], PPYOLO [39] etc., to improve the effect of object detection and recognition in intelligent assisted driving based on the data fusion algorithm in this paper.

ACKNOWLEDGMENT

(Tianbi Liu and Shanshan Du contributed equally to this work.)

REFERENCES

[1] C. Liu and N. Pobox, "Feature-level data fusion methods of using image and radar," *Comput. Digit. Eng.*, vol. 2013, no. 11, pp. 47–49, 2013.

[2] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. 1–11.

[4] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

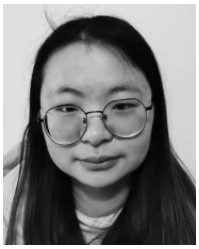
[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934
[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[16] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3D lidar using fully convolutional network," 2016, *arXiv:1608.07916*. [Online]. Available: http://arxiv.org/abs/1608.07916
[17] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.
[18] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast point R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9775–9784.
[19] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. E. Sallab, "YOLO3D: End-to-end real-time 3D oriented object bounding box detection from LiDAR point cloud," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 716–728.
[20] A. Sallab, A. A. Al Sallab, I. Sobh, M. Zidan, M. Zahran, and S. Abdelkarim, "YOLO4D: A spatio-temporal approach for real-time multi-object detection and classification from LiDAR point clouds," in *Proc. Neural Inf. Process. Syst. (NIPS), Mach. Learn. Intell. Transp. MLITS Workshop*, Dec. 2018.
[21] M. Simony, S. Milzy, K. Amendey, and H. M. Gross, "Complex-YOLO: An euler-region-proposal for real-time 3D object detection on point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 197–209.
[22] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.
[23] J. Beltran, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "BirdNet: A 3D object detection framework from LiDAR information," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3517–3523.
[24] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.
[25] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*. [Online]. Available: http://arxiv.org/abs/1706.02413
[26] J. Ngiam, B. Caine, W. Han, B. Yang, Y. Chai, P. Sun, Y. Zhou, X. Yi, O. Alsharif, P. Nguyen, Z. Chen, J. Shlens, and V. Vasudevan, "StarNet: Targeted computation for object detection in point clouds," 2019, *arXiv:1908.11069*. [Online]. Available: http://arxiv.org/abs/1908.11069
[27] G. Zhai and R. J. C. Zhang, "Tramway obstacles detection based on information fusion of MMV radar and machine vision," *Chin. J. Internet Things*, vol. 1, no. 2, pp. 76–83, 2017.
[28] L. Zhiqiang *et al.*, "The study on detection of obstacles ahead vehicle," *Machinery Des. Manuf.*, vol. 2018, no. 7, pp. 31–33 and 37, 2018.
[29] B. Xin *et al.*, "A new method of target detection based on autonomous radar and camera data fusion," in *Proc. Intell. Connected Vehicles Symp.*, 2017.
[30] Z. W. Qu *et al.*, "Pedestrian detection by radar vision data fusion," *J. Jilin Univ. (Engineer Technol. Ed.)*, vol. 43, no. 5, pp. 1230–1234, May 2013.
[31] J. Lisheng, C. Lei, and C. Bo, "Leading vehicle detection at night based on millimeter-wave radar and machine vision," *J. Automot. Saf. Energy*, vol. 7, no. 2, p. 167, 2016.
[32] Y. Yuan, L. Xiao, and Z. Jinhuan, "A vehicle recognition method using image and radar data fusion," in *Proc. 16th Automot. Saf. Technol. Conf. China Soc. Automot. Eng.* 2013, pp. 104–109.
[33] K. Aziz, E. De Greef, M. Rykunov, A. Bourdoux, and H. Sahli, "Radar-camera fusion for road target classification," in *Proc. IEEE Radar Conf. (RadarConf)*, Sep. 2020, pp. 1–6.
[34] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *Proc. Sensor Data Fusion, Trends, Solutions, Appl. (SDF)*, Oct. 2019, pp. 1–7.
[35] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*. [Online]. Available: http://arxiv.org/abs/1504.00325
[36] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[37] M. P. Muresan, I. Giosan, and S. Nedevschi, "Stabilization and validation of 3D object position using multimodal sensor fusion and semantic segmentation," *Sensors*, vol. 20, no. 4, p. 1110, Feb. 2020.
[38] J. Nie, J. Yan, H. Yin, L. Ren, and Q. Meng, "A multimodality fusion deep neural network and safety test strategy for intelligent vehicles," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 2, pp. 310–322, Jun. 2021.
[39] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding, and S. Wen, "PP-YOLO: An effective and efficient implementation of object detector," 2020, *arXiv:2007.12099*. [Online]. Available: http://arxiv.org/abs/2007.12099

TIANBI LIU received the master's degree from the School of Electronics and Information Engineering, Tongji University, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Application and Technology, Fudan University, China. His research interests include computer vision, multimedia, and artificial intelligence.

SHANSHAN DU received the master's degree from Fudan University, China, in 2015, where she is currently pursuing the Ph.D. degree majoring in computer science and technology. Her research interests include computer version and multimedia.



CHENCHEN LIANG received the bachelor's degree from the Taiyuan University of Technology, Shanxi, China, in 2019. She is currently pursuing the master's degree with Tongji University, China, majoring in electronics and communication engineering. Her research interests include broadband wireless communication and embedded systems.

BO ZHANG received the master's degree from the Communication University of China, in 2011. He is currently pursuing the Ph.D. degree with the Academy for Engineering and Technology, Fudan University. His research interests include computer vision, multimedia and pattern recognition, and digital twins.

RUI FENG received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2003. In 2003, he joined Fudan University, Shanghai, China, where he is currently a Research Professor. His research interests include computer vision, multimedia, and pattern recognition.