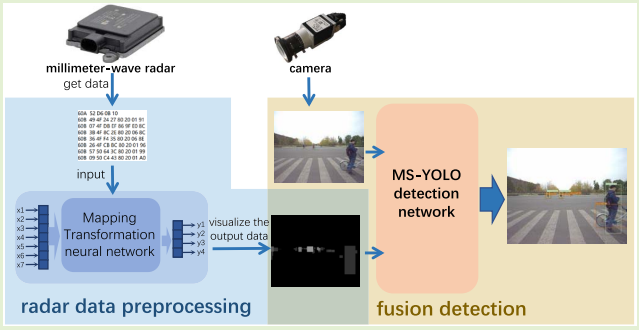
MS-YOLO: Object Detection Based on YOLOv5 Optimized Fusion Millimeter-Wave Radar and Machine Vision

Yunyun Song, Zhengyu Xie, Xinwei Wang, and Yingquan Zou

***Abstract*—Millimeter-wave radar and machine vision are both important means for intelligent vehicles to perceive the surrounding environment. Aiming at the problem of multi-sensor fusion, this paper proposes the object detection method of millimeter-wave radar and vision fusion. Radar and camera complement each other, and radar data fusion in machine vision network can effectively reduce the rate of missed detection under insufficient light conditions, and improve the accuracy of remote small object detection. The radar information is processed by mapping transformation neural network to obtain the mask map, so that radar information and visual information in the same scale. A multi-data source deep learning object detection network (MS-YOLO) based on millimeter-wave radar and vision fusion was proposed. Homemade datasets were used for training and testing. This maximized the use of sensor information and improved the detection accuracy under the premise of ensuring the detection speed. Compared with the original YOLOv5 (the fifth version of the You Only Look Once) network, the results show that the MS-YOLO network meets the accuracy requirements better. Among the models, the large model of MS-YOLO has the highest accuracy with an mAP reaching 0.888. The small model of MS-YOLO has good accuracy and speed, and the mAP reaches 0.841 while maintaining a high frame rate of 65 fps.**



***Index Terms*—MS-YOLO, object detection, multi-sensor fusion, deep learning.**

I. Introduction

With the development of autonomous vehicles, various autonomous driving technologies are gradually maturing. Among them, environmental perception, as the most critical step of autonomous driving, is the basis of decision-making and control, and many researchers are conducting further research in this area. At present, there are many sensors commonly used by researchers for intelligent vehicle environment perception, but a single sensor can only obtain partial characteristics of the surrounding environment, which makes it difficult to meet the needs of intelligent vehicle environment perception. Multi-sensor fusion can maximize the use of the information obtained by different sensors and provide more comprehensive and accurate characteristics of the surrounding environment; therefore, multi-sensor fusion has gradually become the mainstream intelligent vehicle environmental perception field.

Machine vision has become an indispensable means of object detection in intelligent driving. It can perform well in a series of different scenes. However, vision requires certain lighting conditions, and monocular devices cannot accurately obtain the position information of a object. Millimeter-wave radar, another commonly used sensor for intelligent driving, has strong complementarity with vision. It is robust to weather changes and can obtain object position information and speed information. However, millimeter-wave radar cannot classify objects, and it is difficult to determine the object size. Many studies have shown that the fusion of two sensors is more accurate than the use of vision alone in object detection.

Considering comprehensively, radar data processing and visual deep learning networks are researched in this paper to conduct object detection. On the premise of ensuring a certain detection speed, the object detection accuracy is improved to meet the environmental detection requirements of intelligent vehicles. The main work of this paper is as follows: (1) The mapping transformation neural network is constructed to process the millimeter-wave radar information. The network

Manuscript received 17 February 2022; revised 4 April 2022; accepted 4 April 2022. Date of publication 22 April 2022; date of current version 1 August 2022. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61403316. The associate editor coordinating the review of this article and approving it for publication was Dr. Brajesh Kumar Kaushik. *(Corresponding author: Yingquan Zou.)*

The authors are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China (e-mail: zouyingq@home.swjtu.edu.cn).

Digital Object Identifier 10.1109/JSEN.2022.3167251

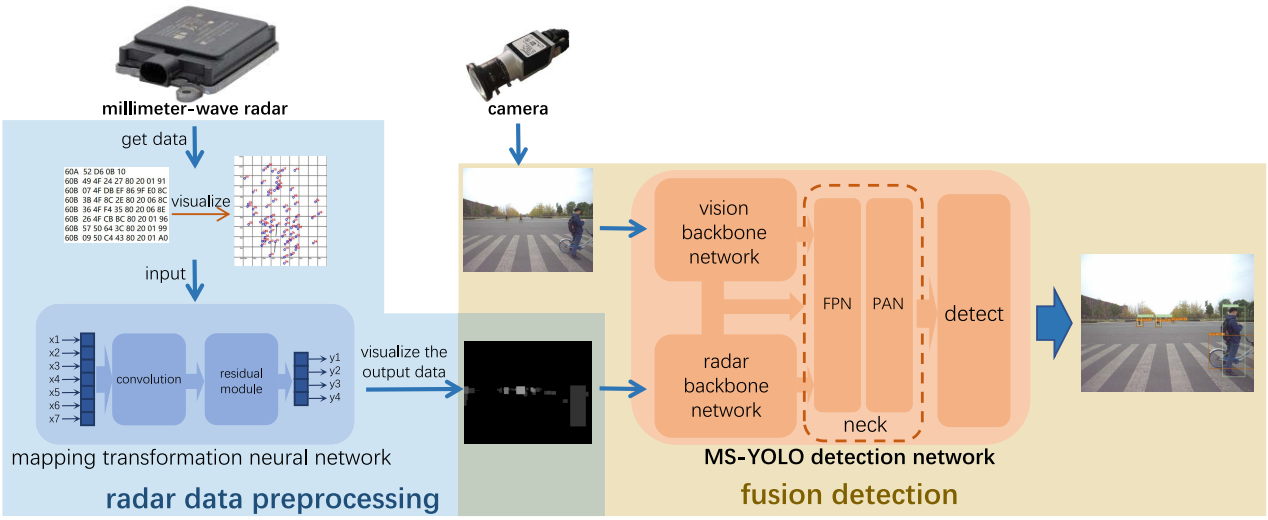


Fig. 1. The entire process of multi-sensor information fusion object detection. First, the radar data are processed to form the corresponding mask map, the mask map and the original camera image are input to MS-YOLO at the same time, and MS-YOLO conducts feature extraction on the two types of data. Then, feature fusion is conducted, and the final detection is conducted.

can generate the mask map that contains object information and unify the information of the two sensors. (2) A multi-data source deep learning object detection network (MS-YOLO) based on millimeter-wave radar and vision fusion was constructed: optimized on the basis of YOLOv5 (the fifth version of the You Only Look Once) network, double backbone network was used to extract the information features of two sensors respectively, and the radar feature fusion path was appropriately added. The network is trained and tested using homemade dataset (MSDataset). (3) The small, medium and large models of MS-YOLO were used for training and testing and compared with the corresponding size of the original YOLOv5 network model.

The experimental results show that the performance of the two sensor data fusion networks proposed in this paper is higher than that of the original image detection network. Fig. 1 shows the entire process of multi-sensor information fusion object detection proposed in this paper.

The remainder of this paper is organized as follows. Section II gives the related work on object detection based on images and object detection based on radar and image fusion. The radar data processing is described in Section III. Section IV mainly introduces the proposed multi-sensor data fusion network, including the network architecture, datasets and some training details. The experimental results are given in Section V, and the work of this paper is summarized in Section VI.

II. Related Work

A. Object Detection Based On Images

A camera records rich original information and has a strong object classification ability. The traditional machine visual object detection algorithm [1], [2] can achieve a good effect under certain conditions. In 2020, Chiman Kwan *et al.* [3] combined flow methods with contrast enhancement, connected component analysis, and object association to detect small moving objects in long-range infrared videos. In 2021, Chiman Kwan *et al.* [4] proposed two unsupervised approaches using change detection algorithms for small moving object detection in IR videos. The Result shows that the module can detect objects quite effectively. His research group [5] proposed a high-performance approach to detecting small objects in long-range and low-quality infrared videos, he used a system consist of a video resolution enhancement module, a proven small object detector based on local intensity and gradient (LIG), a connected component (CC) analysis module, and a track association module. However, traditional algorithm has a large amount of computation, poor robustness and is prone to redundancy. After 2012, the field entered the stage of deep learning object detection [6]. In recent years, image processing algorithms have become increasingly mature, and machine vision has been widely used in various fields. Deep learning-based object detection algorithms can be divided into two-stage detection and single-stage detection. Among them, the two-stage detection is represented by the R-CNN [7] and SPP-Net [8]. The idea is to first detect the position and then classify it; however, the real-time detection is poor. Furthermore, single-stage detection is represented by YOLO [9] and SSD [10]. The idea is to directly take an entire image as input and directly return the object category and position, which has obvious real-time advantages.

With the development of the YOLO network, YOLO has become as accurate as some two-stage detection methods. The YOLO (you only look once) [9] algorithm was first proposed by Joseph Redmon *et al.* in 2016. Then, the improved YOLOv1 algorithm, YOLOv2 [11] and YOLOv3 [12] were proposed. With the follow-up of more researchers, in April 2020, the YOLOv4 algorithm was proposed by Alexey Bochkovskiy's team [13], and the effect was significantly improved. A month later, Glenn Jocher released the YOLOv5 algorithm via open source access. YOLOv5 has four networks with different depths and widths. The larger the model is, the slower its speed, and the greater its accuracy. The mAP and reasoning speed of YOLOv5 on the COCO dataset show that YOLOv5 has excellent performance and has an obvious speed advantage while ensuring certain accuracy. Mohammad Shahab Uddin et.al [14] propsed attention GAN model to

generate more stable IR images and get a better result when they used YOLO algorithm to detect the objects. Chiman Kwan proposed [15] a deep learning approach that directly performs object tracking and classification in the compressive measurement domain without any frame reconstruction. This methond using YOLO can be developed on processing and control element cameras. However, images obtained by cameras are easily affected by the environment, and it is difficult to obtain monocular depth information; therefore, this paper selects the YOLOv5 network for corresponding research and improvement.

B. Object Detection Based on Radar and Image Fusion

Intelligent vehicle object detection based on multi-sensor fusion has important theoretical value and practical significance. Using camera and millimeter-wave radar fusion, the two sensors complement each other to obtain the positioning information and relative speed to classify the detected objects. The dual sensors can also improve reliability. Domestic and foreign scholars have performed a considerable amount of research on object detection based on visual and millimeter-wave radar fusion. In 2000, Lakshmanan *et al.* [16] used radar and visual information to detect vehicles and lanes. The basic method was traditional and possessed great limitations. The Fade proposed by Steux's [17] team can detect vehicles in the left and right lanes in real time and predict vehicle behavior by detecting turn signals, but it needs to be improved to detect other objects such as pedestrians. The fusion algorithm proposed by Bombini [18] in 2006 uses radar data to locate the region of interest, and the visual system is used to verify and improve the accuracy. This algorithm cannot detect and distinguish between multiple similar vehicles. Wang *et al.* [19] proposed mapping radar data to images and comparing the image detection results, combined with radar data and tracking algorithms, to improve the vehicle detection accuracy in bad weather. However, it is still challenging to accurately detect vehicles when the camera is heavily jammed. In 2019, John [20] proposed a deep learning-based sensor fusion framework RV-Net, which effectively integrates image and radar features and can detect obstacles in real time. In 2020, he [21] proposed an RV-Net extended learning framework called SO-Net, which added a semantic segmentation branch to reduce the computational complexity and realize spatial segmentation and vehicle detection. Nobis *et al.* [22] proposed CRF-Net that used a logistic regression method to automatically locate the optimal fusion layer. The results indicated that the detection effect of the fusion network was better than that of the image network only, and the authors concluded that further exploration of the optimized network structure was needed. In the 2020, YOdar proposed by Kowol *et al.* [23], two networks are used to process image and radar data, and the results are used for joint prediction, which significantly improves the detection performance. In the process, the features of the data of the two sensors are not fused. Dong *et al.* [24] feed radar detection results with original images for fusion detection after transforming them to the same scale and designed the loss caused by labeling error; however, the network structure is relatively simple.

III. Radar Data Preprocessing

The millimeter-wave radar used in this paper is ARS 408-21 manufactured by German Continental. ARS 408-21 has a pseudorandom coding design, which can avoid the interference caused by multiple radars working at the same time and is very suitable for the environmental sensing sensors of intelligent vehicles. The ARS sensor uses radar radiation to analyze its surroundings. After the reflected signals are processed by the algorithm of the radar module, they become available and output through CAN(Controller Area Network). The CAN protocol is used to parse the radar data and obtain information about each object detected. In this paper, the radar is configured to generate the position, speed, length, width, and possible category of each object in each frame.

Since the data collected by cameras and millimeter-wave radar belong to different coordinate systems, to achieve information fusion, it is necessary to ensure the spatial unity of the two types of information. That is, the conversion relationship between the two kinds of information needs to be determined.

Common spatial transformation methods are as follows: (1) The traditional coordinate transformation [25], [26] uses each parameters and derives the radar coordinate and pixel coordinate conversion formula through the transformation of each coordinate system. (2) To calibrate multiple groups of data, the least square method is used to fit the transformation process [27]. (3) Training a neural network to simulate the spatial transformation process. While the first two methods can only achieve different point-to-point mapping, the last method can extract advanced features and directly obtain the radar information mapped object box. In order to strengthen the fusion effect, this paper adopts the third method, construct the corresponding datasets and design the neural network to map the millimeter-wave radar information into the image to obtain the corresponding mask to conduct the subsequent data fusion work.

A. Mapping Transformation Neural Network Description

Neural networks are widely used in all types of fitting problems and prediction problems. In this paper, the corresponding dataset was constructed and optimized based on the typical BP (Back Propagation) neural network: convolution and residual modules are added. The constructed network processes the millimeter-wave radar data and converts the objects information detected by the radar into the pixel reference frame.

The overall design of the mapping transformation neural network is shown in Fig. 2. Seven kinds of object state information can be obtained by parsing CAN information output by the radar module. Namely, longitudinal distance, transverse distance, longitudinal velocity, transverse velocity, category, length and width, therefore, the input layer of the network is set to 7. The network output needs to obtain the bounding box of the object in the image corresponding to the object detected by radar, therefore, the output layer is set to 4, which is the pixel coordinates of the upper left corner and the lower right corner of the bounding box. The network structure is optimized based on the traditional BP neural network: (1) The first step is to add the convolution operation. The one-dimensional convolution operation usually involves convolution along a certain

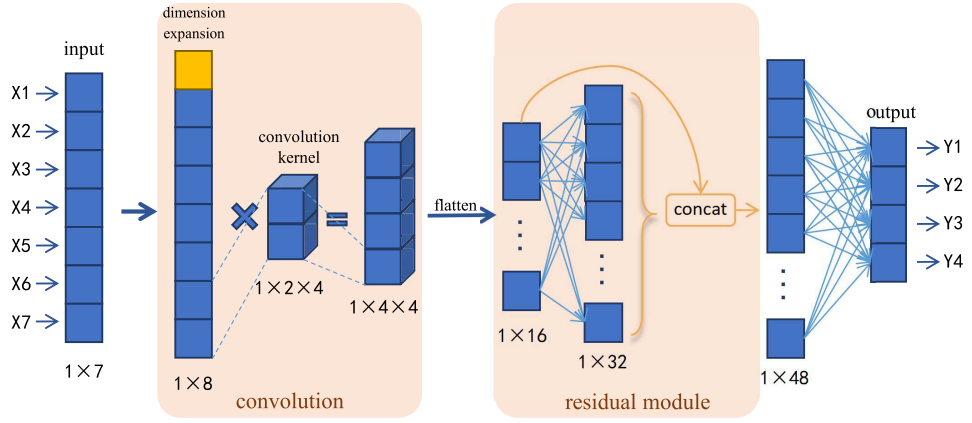


Fig. 2. Mapping transformation neural network structure.

direction, that is, extracting features in a certain direction. The network input is a one-dimensional vector with a length of 7 to represent the millimeter-wave radar data, and the data are extended above to ensure normal convolution. The added convolution is a one-dimensional convolution with 1 input channel, 4 output channels, a 1 × 2 convolution kernel and a step size of 2. (2) The second step is to design the residual module [28]. By adding residual module into the network, the deep-level features can be integrated with the shallow-level features. Therefore, the deep-level feature map and the shallow-level feature map are superimposed so that the number of parameters can be reduced whereas the network performance does not degrade as the network depth increases.

Since the network output is two coordinate points, the loss calculation adopts the mean square error, which represents the distance and is more suitable for this paper compared with other losses. The loss function can be expressed as:

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where E is the average error of this batch of data, n is the number of this batch of data, y_i is the correct value of the ith data in the batch, $\hat{y}_i$ is the predicted value given by the neural network.

The initial learning rate of the network is set to 0.000002, and setting a small learning rate for the sake of avoiding gradient explosion. The Adam optimizer [29] is adopted to reduce the loss quickly, which can make the network find the optimal solution quickly. In this paper, the PyTorch network framework is used to build the neural network, the linear module is used to build each layer in the network, and LeakyReLU is used as the activation function.

B. Datasets

The data obtained by millimeter-wave radar after processing by the underlying algorithm communication protocol parsing are the seven types of state information of the objects detected in each frame, including longitudinal distance, transverse distance, longitudinal velocity, transverse velocity, possible category, length and width. Objects with obvious features can be labeled with categories, including people, car, bicycle and bus. Visualizing the radar information and comparing it with the optical image, there are obvious corresponding objects. Labeling the position of the object bounding box in the image. The radar data of the object and the pixel coordinates of the bounding box together form a data point. A total of 800 objects are labeled in this homemade dataset for training the mapping transformation neural network. Of these objects, 700 data points are used as the training set and 100 data points are used as the test set.

C. The Training Results

The standard error between the network output and the label, which is regarded as the number of predicted error pixels, is used as the performance index. The average predicted speed of the trained network was 0.360ms per data, and the standard error of the test was dropped to 13.57 pixels. After the radar data of an object is processed by the network, the coordinates of the corresponding pixel bounding box are obtained. Form a mask according to the coordinates. The gray value of the mask is determined by the speed value of the object through amplitude limiting and normalization so that the speed information is reflected in the mask. A mask map consists of the masks of all detected objects in a frame of radar data. Algorithm 1 is the process of transforming a frame of radar data into a radar mask map.

The mapping transformation neural network obtained by training in this section has certain adaptability and robustness in different scenes. Fig. 3 shows the transformative effect of the mapping transformation neural network model in three scenes of daytime crossroads, daytime curve and night straight road. Among the three groups of images, each group includes the original image collected by the camera, the vector map of radar information in the corresponding frame, visualization of radar data label, and the mask map of radar data. Label the objects that correspond to the image and radar data, which form a visual diagram of the radar data label. The mask map is composed of masks formed by the mapping transformation neural network.

IV. Proposed Network

The deep learning object detection algorithm of multi-data source fusion proposed in this paper is based on the

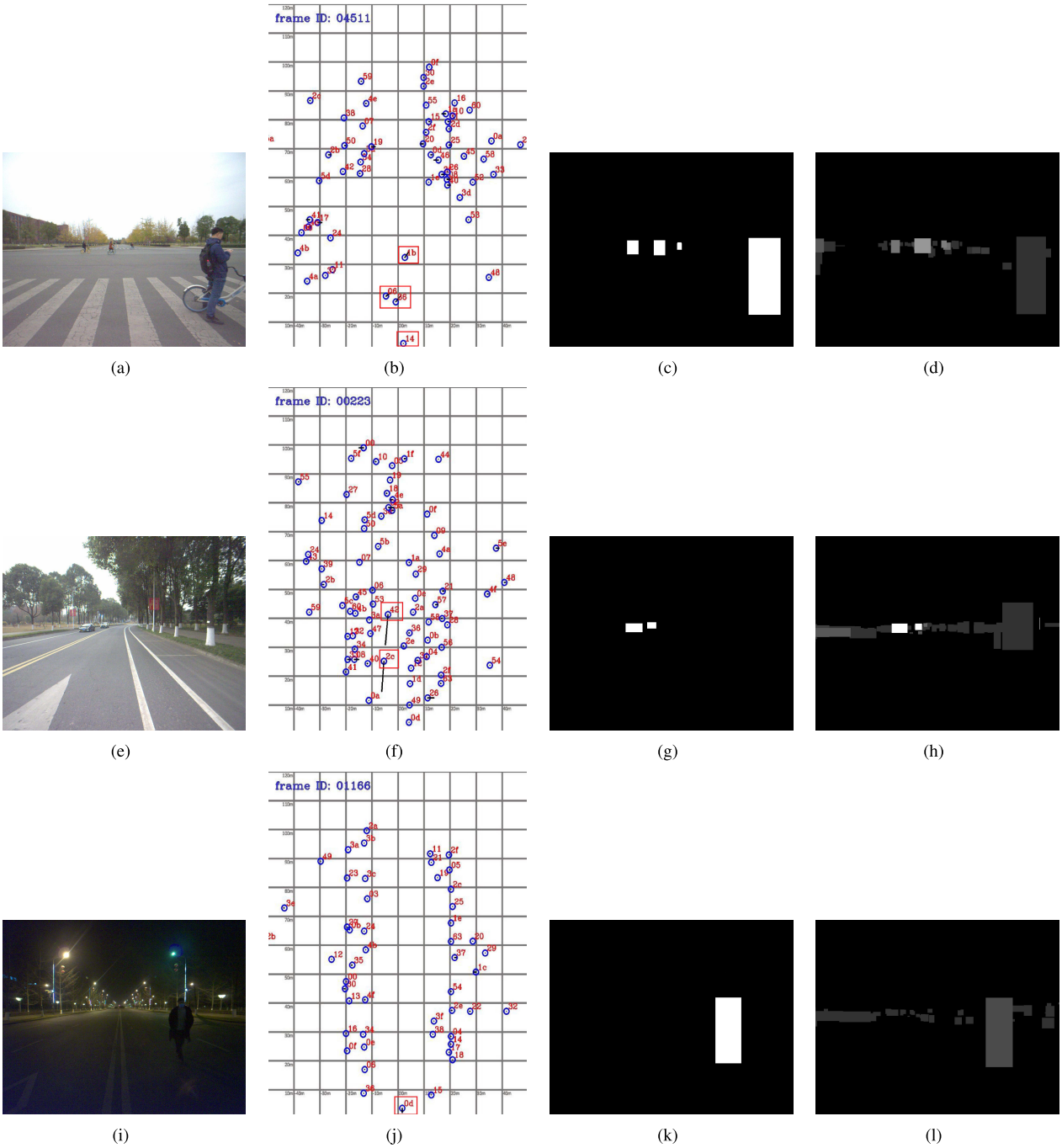


Fig. 3. Effect of the mapping transform neural network model. (a) Daytime crossroad scene original image. (b) Daytime crossroad scene radar vector map. (c) Visualization of radar data label numbered 1b, 06, 36, 14. (d) Daytime crossroad scene radar mask map. (e) Daytime curve scene original image. (f) Daytime curve scene radar vector map. (g) Visualization of radar data label numbered 2c, 42. (h) Daytime curve scene radar mask map. (i) Night straight scene original image. (j) Night straight scene radar vector map. (k) Visualization of radar data label numbered 0d. (l) Night straight scene radar mask map.

multi-source improvement of the YOLOv5 network, so it is named the MS-YOLO network. This section covers the MS-YOLO network structure, datasets, and parameter setting.

The MS-YOLO network designs two backbone networks to extract the features of image and millimeter-wave radar information and fuses the feature map with different depths in the two backbones in the middle layer. Finally, three detection layers are used for detection. The network achieve the fusion of the information of the two different sensors through these steps.

A. Proposed Network Description

The performance of YOLOv5 on the COCO dataset shows that YOLOv5 can guarantee certain accuracy and has an obvious speed advantage. Therefore, this paper selects the YOLOv5 network for the corresponding research and improvement. The entire YOLOv5 network is divided into the backbone network (Backbone), the middle layer (Neck), and the detection layer (Head). YOLOv5 uses CSPDarknet [30] as the backbone network, and the middle layer contains both the FPN (Feature Pyramid Networks) structure and PAN

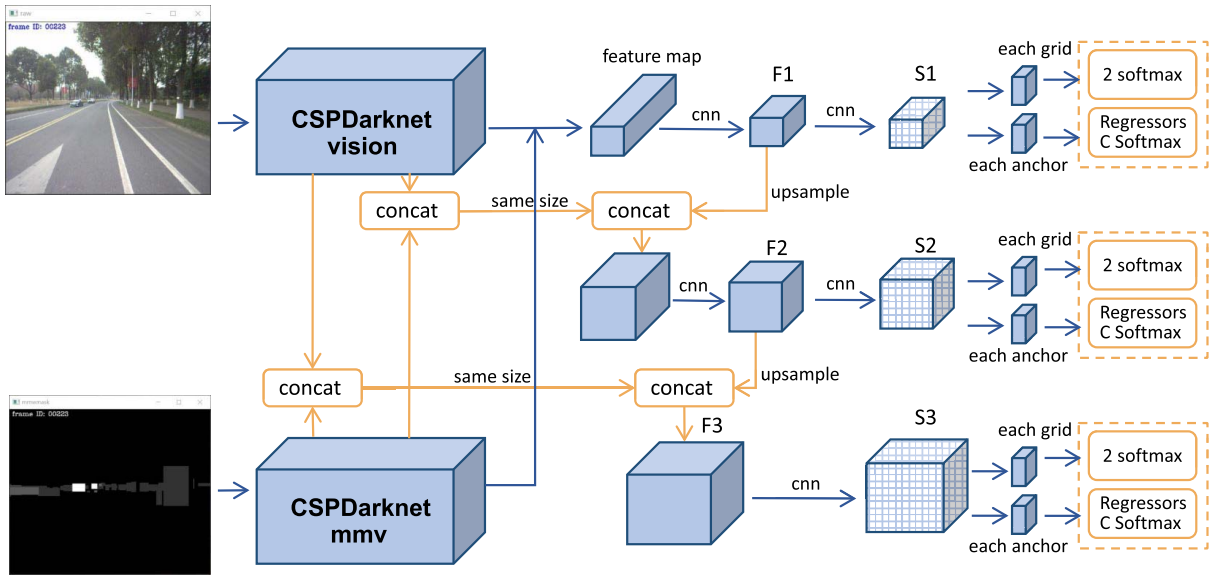


Fig. 4. Schematic diagram of the overall network structure of MS-YOLO.

Algorithm 1 Framework of Forming a Radar Mask Map

Require:

The status information of all detected objects in a frame of radar data, D_n;

The mapping transformation neural network, M;

Ensure:

A radar mask map, Img;

Initialize a 1024×1280 array Img with all 0 values;
for all D_i such that $D_i \in D_n$ **do**
  Calculate the velocity s_i from the velocity component information in D_i;
  $s_i \Leftarrow s_i \times 25$;
  if $s_i < 5$ **then**
    $s_i \Leftarrow 5$;
  else if $s_i > 255$ **then**
    $s_i \Leftarrow 255$;
  end if
  Feed D_i into M and get the bounding box B_i corresponding to the image;
  Set all values of B_i in Img to s_i;
end for
return Img;

(Spatial Pyramid Pooling) structure. Features are obtained from different backbone layers for feature fusion so that the feature information of the detection layer is greatly enhanced.

Referring to the idea of the YOLOv5 model framework, a multi-source object detection network (MS-YOLO) is proposed in this paper. The purpose of the MS-YOLO network is to fuse the information of two sensors, so the data features of the two sensors need to be combined in the specific implementation of the MS-YOLO object detection model. The MS-YOLO network is improved based on YOLOv5 in the two following aspects: (1) a dual backbone network, (2) a feature guidance structure of millimeter-wave radar features in the middle layer. The MS-YOLO network constructs a double backbone structure based on an image feature extraction backbone network and conducts feature extraction for millimeter-wave radar and camera data in the early stage, which is used for later fusion and improves the detection accuracy.

Fig. 4 shows a simple framework of the MS-YOLO network. The inputs of the network are image and the mask map generated by the corresponding frame radar data. The network constructs two backbone networks for feature extraction. CSPDarknet-vision and CSPDarknet-MMW, which extract the shallow features of the image and millimeter-wave radar information, respectively. After the backbone features are extracted, the features of the double backbone are fused to obtain F1, which is input into the first detection layer. The fusion here fuses the advanced features of the radar mask map with the advanced features of the image. F1 is upsampled and then fused with feature maps, which have the same resolution and are in two CSPDarknet networks, and then F2 is obtained through convolution and other operations after fusion. F2 is upsampled and then fused with feature maps, which have the same resolution and are in two CSPDarknet networks, and then F3 is obtained through convolution and other operation. F1, F2 and F3 then obtain S1, S2 and S3, respectively, through convolution and other operations, which serve as the inputs of the three detection layers.

Fig. 5 is the schematic diagram of the network module composition and connection of MS-YOLO. The black connection blocks are all modules of the original YOLOv5 network, the blue connection blocks are the backbone network for processing the radar mask map, and the orange connection blocks are the added radar data feature fusion paths. Focus is a slicing operation, which halves the width and height of the feature map and increases the number of channels by four times. Fig. 6(a) shows the structure of the Focus. SPP (Spatial Pyramid Pooling) [31] can combine local features with global

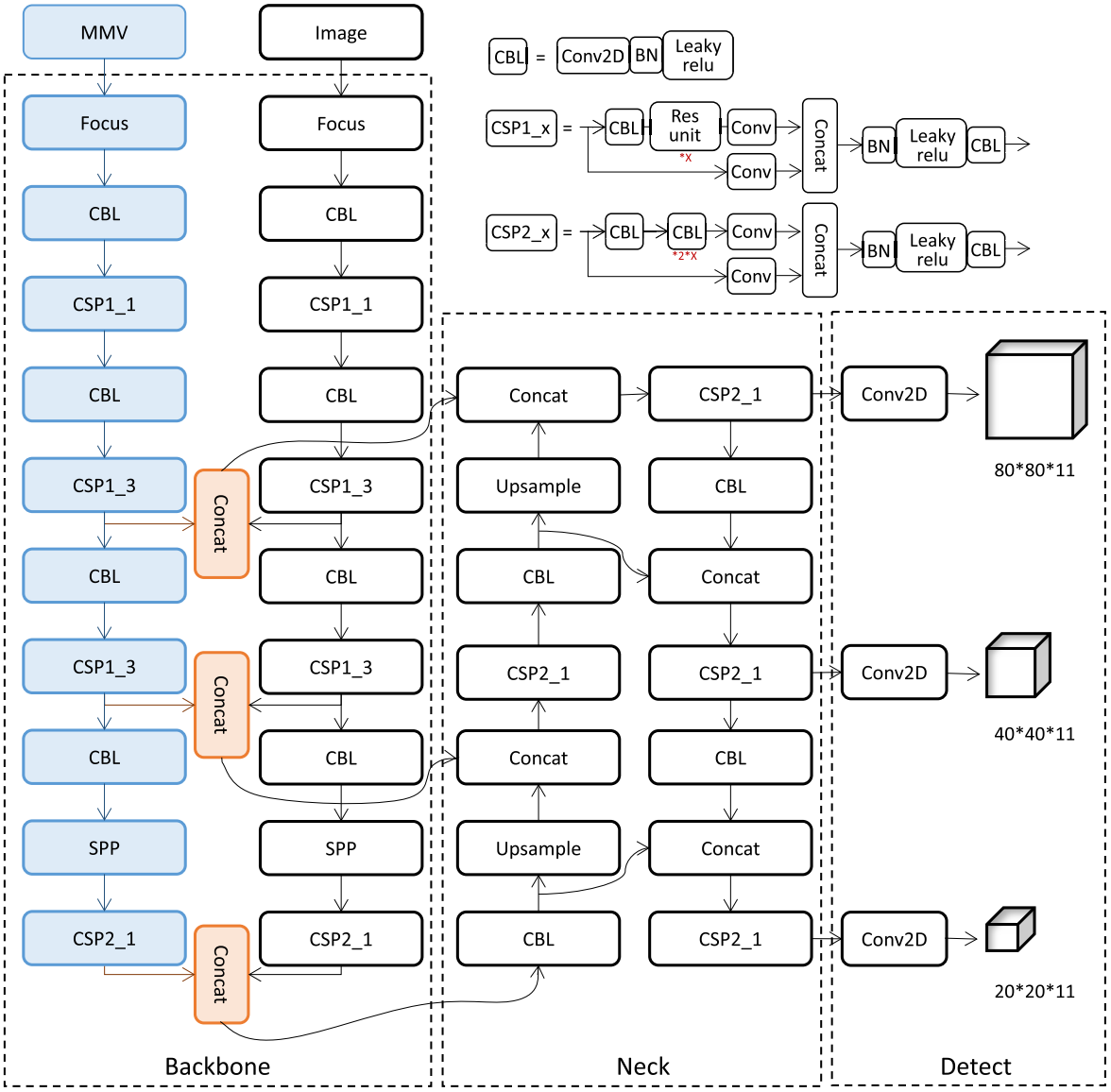


Fig. 5. Schematic diagram of the network module composition and connection of MS-YOLO.

features, which is beneficial to image detection with the large object size differences. Fig. 6(b) shows the structure of the SPP. CBL is the standard convolution layer, consisting of two-dimensional convolution, Batch Normalization [32] and activation function, which here uses LeakyReLU. CSP (Cross Stage Partial Network) consists of several Bottleneck [28] and several standard convolution layers. Fig. 5 shows the structure of the CSP, and the red font under the blocks in the figure represents the number of blocks. The entire network is divided into the dual backbone network (Backbone), middle layer (Neck), and detection layer (Detect). The Backbone is an orderly combination of several modules, such as Focus, CBL, CSP and SPP. The main function of the Backbone is feature extraction. The radar branch MMV and the image branch Image form the Backbone of the network together. The middle layer (Neck) is an ordered combination of CSP, CBL and Upsampling modules; and its main function is feature fusion. The left and right sides of Neck are the FPN structure [33] and PAN structure [34], respectively, which receive fusion features of different depths from the Backbone and then conduct all-round feature fusion and generate three groups of features with different resolutions. The detection layer (Detect) is composed of the convolution block, receives three groups of features from the Neck, and generates three groups of detection results with resolutions of 20×20, 40×40 and 80×80 through the convolution operation. The depth is 6 category numbers plus 4 positional parameters and a confidence level, the sum is 11. Through the above process, millimeter-wave radar and image fusion detection can be conducted.

B. Training Details

1) Loss Function: A good loss function can make the network converge quickly and perform better. The part of the loss function of the MS-YOLO model in this paper is similar to that of YOLOv5, and it includes three parts: the classification loss L_{cls}, the position loss L_{box} and the confidence loss L_{obj}. Each loss function is as follows:

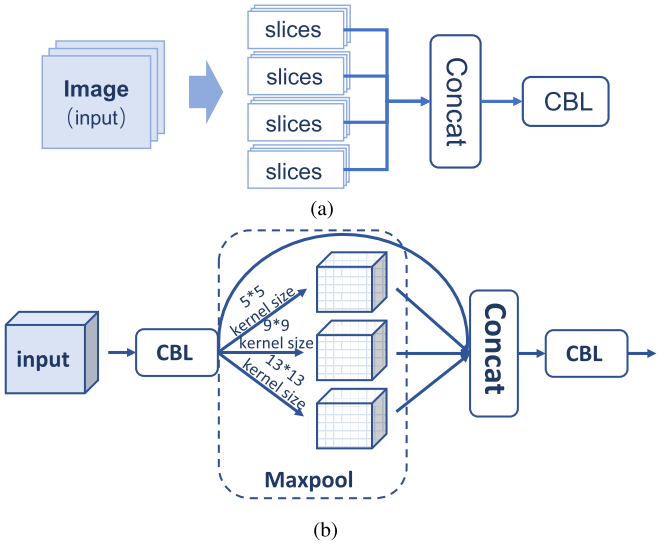


Fig. 6. (a) Structure of Ffocus. (b) Structure of SPP.

The position loss can be expressed as:

$$\begin{aligned} L_{box} &= \lambda_{coord} \sum_{i=1}^{s\times s} \sum_{j=1}^{b} I_{ij}^{obj} (2 - w_i \times h_i) \\ &\times [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\ &+ \lambda_{coord} \sum_{i=1}^{s\times s} \sum_{j=1}^{b} I_{ij}^{obj} (2 - w_i \times h_i) \\ &\times [(w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2] \end{aligned} \tag{2}$$

The confidence loss can be expressed as:

$$\begin{aligned} L_{obj} &= \sum_{i=1}^{s\times s} \sum_{j=1}^{b} I_{ij}^{obj} [\hat{C}_i \log(C_i) \\ &+ (1 - \hat{C}_i)(1 - log(C_i))] \\ &- \lambda_{noobj} \sum_{i=1}^{s\times s} \sum_{j=1}^{b} I_{ij}^{noobj} [\hat{C}_i \log(C_i) \\ &+ (1 - \hat{C}_i)(1 - \log(C_i))] \end{aligned} \tag{3}$$

The classification loss can be expressed as:

$$\begin{aligned} L_{cls} = \sum_{i=1}^{s\times s} \sum_{j=1}^{b} I_{ij}^{noobj} \sum_{c\in classes} &[\hat{p}_i(c) \log(p_i(c)) \\ &+ (1 + \hat{p}_i(c)) \log(1 - p_i(c))] \end{aligned} \tag{4}$$

where $s \times s$ is the partition of the image; b is the number of anchors corresponding to each grid; c is the number of categories; p is the probability of categories; x_i, y_i, w_i, and h_i are the horizontal and vertical coordinates and width and height at the center point of the bounding box in the grid, respectively; λ_{coord} is the weight of the predicted loss of bounding box coordinates; and λ_{noobj} is the weight of the predicted loss of confidence of bounding box without an object; C_i is the confidence of the category, and $\hat{C}_i$ is the predicted confidence of the category. When the object center falls into the grid, the confidence loss and classification loss need to be calculated. The position loss is calculated only when the intersection between the prediction box and the actual box is greater than the specified threshold.

TABLE I
depth_multiple AND *width_multiple* SETTINGS

Net	$depth\_multiple$	$width\_multiple$
MS-YOLOs	0.33	0.5
MS-YOLOm	0.67	0.75
MS-YOLOl	1	1

2) Parameter Configuration: This section describes the key parameters to be configured. The number of object categories nc is determined according to the number of dataset categories and is set to 6. The model depth (*depth_multiple*) represents the number of modules which control the size of the model, and the model width (*width_multiple*) controls the number of channels in the module. Table I shows the values of these two parameters for three different size models. There are 9 anchors in 3 groups. MS-YOLO extensively uses 3×3 convolution kernels and 1×1 convolution kernels. In SPP, four types of max pooling, 1×1, 5×5, 9×9, and 13×13, were used for multi-scale fusion.

The input image size of the network is 640 $times$ 640. The batch size in each training epoch was set to 64, the Adam optimization algorithm was adopted in the training process, and the learning rate momentum was set to 0.843. The weight delay is set to 0.00036. There are a total of 300 epochs of network training, and the learning rate is updated to half of the original after 64 steps of epoch back propagation. The pre-training model is loaded in the network training, and the initial learning rate is set to 0.0032. The trend of the learning rate first linearly rises and then slowly declines. The rising process occurs because the network uses the pre-training rate, which can make the model converge to a local optimum first. The algorithm model is deployed on the PyTorch framework.

V. EXPERIMENTS

The experimental part of the multi-source fusion network includes six experiments in total, including three contrast experiments and three improved network experiments. The contrast experiments are the YOLOv5s experiment, YOLOv5m experiment and YOLOv5l experiment, which are small model, medium model and large model, respectively. The improved network experiments are similar. There are the small model MS-YOLOs, the medium model MS-YOLOm and the large model MS-YOLOl. In the following, three multi-source fusion object detection network experiments are analyzed using the index changes in the training process and the index of the training results and compared with the contrast experiment networks. The indices involved in the experiment are the same as those in YOLOv5, including the precision, recall, average precision (mAP), and F-measure. In this section, indices such as the precision and mAP are expressed in the form of percentages, and the percentage sign is ignored.

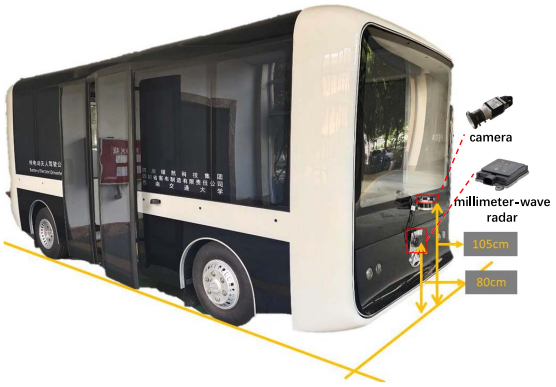


Fig. 7. Millimeter-wave radar and camera installation position diagram.

A. The Experiment Platform

The experimental platform of this paper is an unmanned driving platform based on millimeter-wave radar and visual fusion. The autonomous bus used in the experiment is modified based on an electric bus, as shown in Fig. 7. The car body part is a new electric bus manufactured by a passenger transport company in Sichuan Province and includes a power supply system, actuator, motor and other parts. The central processing unit of the autonomous bus adopts an ECX-1400 PEG series industrial control computer produced by Taiwan Vecow Company, and the underlying control unit is a high-performance STM32F4 series MCU produced by ST company. Corresponding sensors are installed in all directions of the car body to sense the surrounding environment in real time.

The millimeter-wave radar used in this paper is the ARS 408-21 radar of Continental and is installed on the central axis of the autonomous bus at a horizontal height of 80cm. The camera is an acA130-60gc Basler ace GigE high-speed industrial camera produced by Basler Company. It is over the radar and 105cm from the ground. The detection direction of both sensors is horizontal forward. The installation positions are shown in Fig. 7.

Because deep learning requires strong computing power in the training process, the training platform used in this paper is the comprehensive traffic big data and intelligent computing platform of Southwest Jiaotong University, which includes eight NVIDIA RTX TITAN V GPUs. Each RTX TITAN V has 12 GB of independent memory, and the total memory is 96 GB. The equipment used for testing is the ECX-1400 PEG computing platform for an autonomous bus industrial computer, which is equipped with an Intel Core i7 8700H CPU and an NVIDIA GeForce series RTX2060-SUPER GPU to provide the hardware foundation for real-time object detection.

B. Dataset

The Dataset used in this paper is the MSDataset made by the author and is used to train and test the MS-YOLO network and verify network performance. It contains 6 types of common road objects. The data collection includes the original data of millimeter-wave radar and camera and the road surface data of different sections under different lighting conditions, which are all video sequences and radar data recorded from the autonomous bus. There are 7000 images in total, including

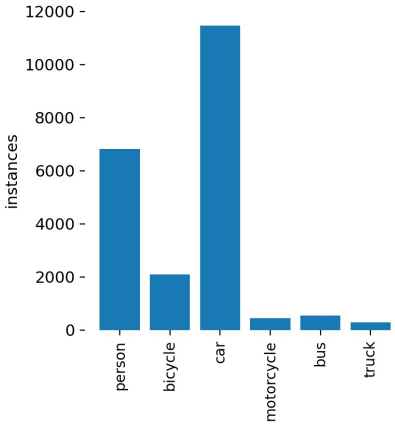


Fig. 8. Distribution of various categories.

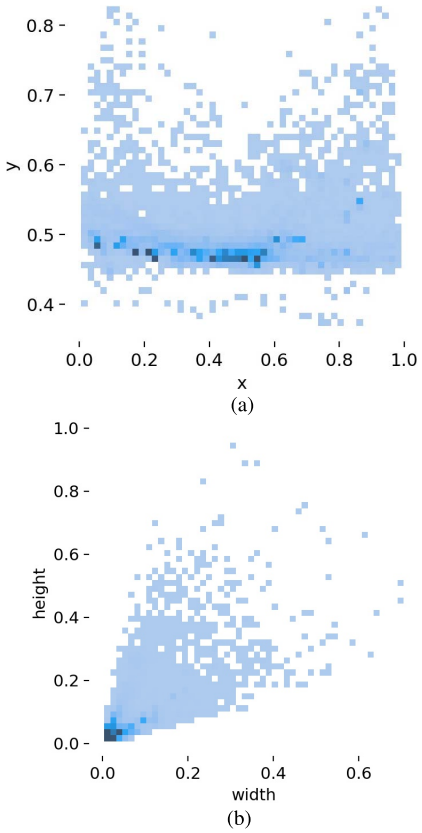


Fig. 9. Distribution of (a) objects center. (b) objects size.

5600 images of the training set and 1400 images of the test set; and the size of the collected images is 1280×1024. Each frame of radar corresponds to the image one by one. Since the application scene does not require a large range of data, only the data within 40 meters and 100 meters forward are retained.

Fig. 8 shows the category distribution of detected objects. The dataset includes 6 common categories including cars, people, bicycles, motorcycles, buses and trucks. People and cars account for most of the detected objects. The main reason is that the data source of this dataset is the environment of the campus scene, with fewer motorcycles, trucks and buses, so most of the objects are people and cars.

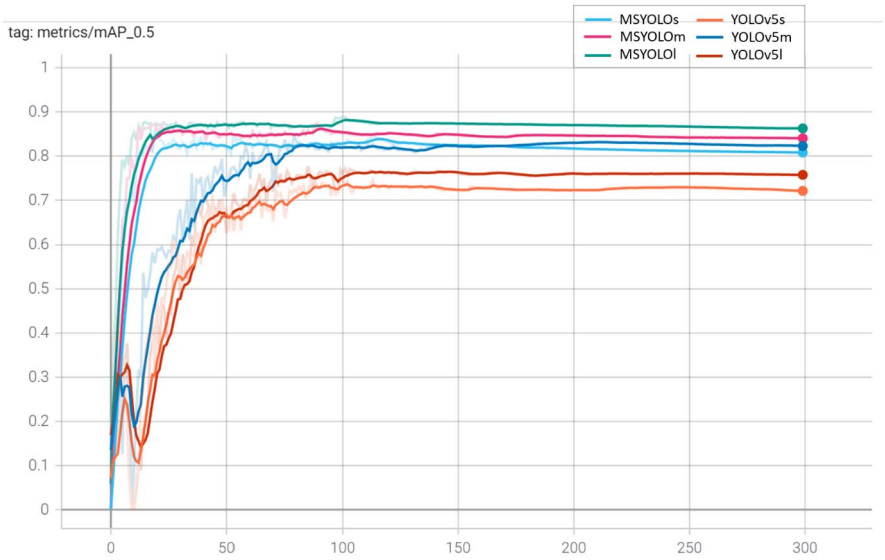


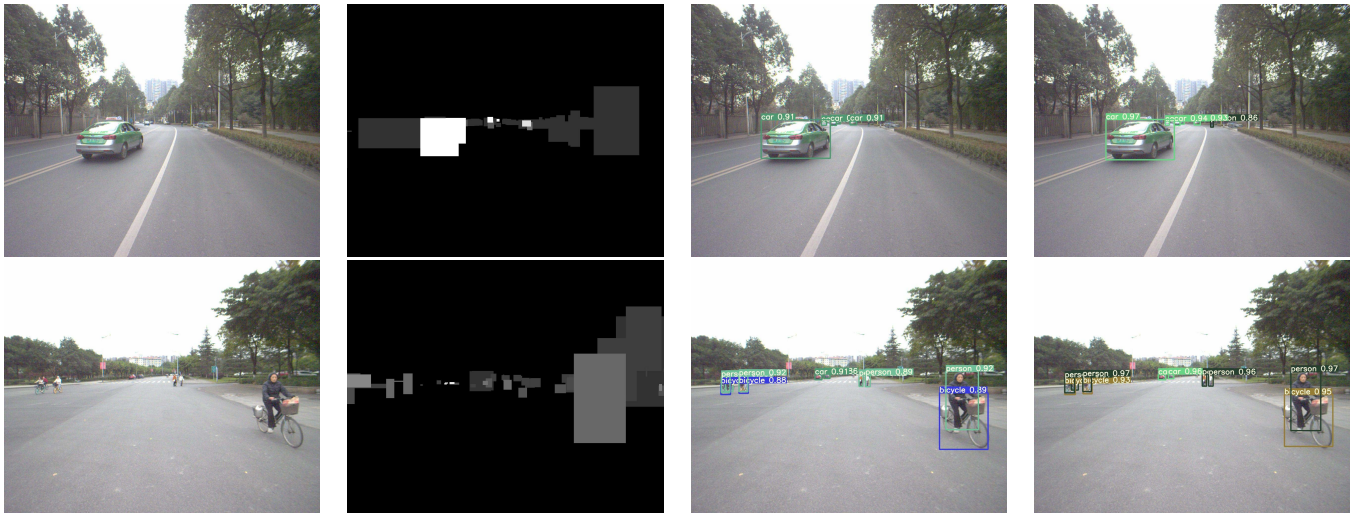
Fig. 10. Comparison of mAP trends.

TABLE II
TRAINING PARAMETERS OF MS-YOLO AND YOLOv5 MODELS

Net	Dataset	depth multiple	width multiple	optimizer	batch size	weight decay	momentum	parameters
YOLOv5s	image part of the MSDataset	0.33	0.5	Adam	64	0.00036	0.843	7.4M
YOLOv5m	image part of the MSDataset	0.67	0.75	Adam	64	0.00036	0.843	21.8M
YOLOv5l	image part of the MSDataset	1	1	Adam	64	0.00036	0.843	47.8M
MS-YOLOs	MSDataset	0.33	0.5	Adam	64	0.00036	0.843	11.5M
MS-YOLOm	MSDataset	0.67	0.75	Adam	64	0.00036	0.843	34.0M
MS-YOLOl	MSDataset	1	1	Adam	64	0.00036	0.843	74.8M

TABLE III
PERFORMANCE COMPARISON OF MS-YOLO AND YOLOv5 MODELS

Net	Precision	Recall	F1	mAP(0.5)	The contrast of mAP	Speed(ms)/Frame rate(fps)
YOLOv5s	80.1	68.5	73.8	74.5	- base	9.8/102
YOLOv5m	85.1	79.3	82.1	82.9	↑ 8.4	21.5/46
YOLOv5l	84.0	71.4	77.2	76.0	↑ 1.5	37.6/26
MS-YOLOs	82.5	80.2	81.3	**84.1**	↑ **9.6**	**15.4/65**
MS-YOLOm	88.3	83.3	85.7	87.4	↑ 12.9	36.1/27
MS-YOLOl	87.4	86.1	86.7	**88.8**	↑ 14.3	62.8/16



(a) (b) (c) (d)

Fig. 11. Daytime scene detection effect. (a) Original camera image. (b) Radar mask map. (c) YOLOv5*s* test results. (d) MS-YOLO*s* test results.

Fig. 9 is the distribution of MSDataset. In Fig. 9(a), x and y represent the center location of the object, which can represent the center location distribution of the objects of the dataset. Detected objects are mostly distributed in the lower part of the field of vision because the dense distribution is on the road surface collected by the camera. The width and height in Fig. 9(b)

Fig. 12. Night scene detection effect. (a) Original camera image. (b) Radar mask map. (c) YOLOv5*s* test results. (d) MS-YOLO*s* test results.

represent the proportion of the width and height of objects to the width and height of the whole image, respectively, which can represent the size distribution of the images of objects in this dataset. Detected objects in this dataset have multi-scale characteristics, and small objects occupy the majority. Because the road information collected is complex, and the scene contains more small objects at a long distance than large objects at a short distance.

C. Comparison and Analysis

Fig. 10 is the comparison of mAP changes during the training process between the three MS-YOLO models and the three YOLOv5 models. Besides, the experimental conditions of the two types of networks are consistent except for the data source (MS-YOLO uses both radar and image data sources, YOLOv5 uses a single image data source), and the pre-training model is used in all tests. Table II shows the training parameters of these models. The increase of the mAP of the MS-YOLO network is faster than that of the YOLOv5 network, and the increase of the mAP of the MS-YOLO series models is stable after the fusion of the data of the two sensors. Furthermore, local instability occurs in the training of the YOLOv5 series models with a single data source.

Table III shows, among the six models, the MS-YOLO*l* model with millimeter-wave radar has the best performance, and the mAP reaches 0.888. The best network for both accuracy and speed is the MS-YOLO*s* model, which achieves an mAP of 0.841 while maintaining a high frame rate of 65fps.

D. The Experimental Results

It shows from detection results in Fig. 11 that the MS-YOLO*s* model has a similar effect as the YOLOv5*s* model in the detection of daytime scenes. Sometimes the detection of small objects at a distance with the YOLOv5*s* model may be missed. The possible reason is that the radar data is fused. Since the radar data contains objects detected from a distance, the mask map generated by the radar data makes the network increase object confidence for distant objects during object detection.

According to the comparison of night scenes in Fig. 12, it is found that the YOLOv5*s* model using only image information has a higher missed detection rate in dark light, especially for fast objects. In the same scene, the MS-YOLO*s* model

using image and radar information fusion can accurately detect objects regardless of whether the light is good or bad. The possible reason is that after the location and speed information are added, the network increases object confidence for the object at the corresponding position. The car often runs at night, and the image shooting will be blurred due to the fast running of itself or the detection object, so obviously MS-YOLO is more in line with the requirements of intelligent car driving.

VI. Conclusion

In this paper, millimeter-wave radar and cameras are taken as the research objects and combined with deep learning, and a multi-data source fusion deep learning object detection network called MS-YOLO is proposed for millimeter-wave radar and visual information fusion. MS-YOLO mainly adds input channels and feature extraction and fusion channels based on the YOLOv5 network. The radar mask map and original image data are input of the network. First, two backbone networks are established to extract features, then feature fusion is conducted in the middle layer, and finally detection is conducted. In addition, to realize the fusion of millimeter-wave radar information and visual information, the mapping transformation neural network is constructed to transform the millimeter-wave radar information so that the millimeter-wave radar information can be mapped into a mask map, which is unified to the same scale as the visual information. Using network models with different sizes (MS-YOLO*s*, MS-YOLO*m*, and MS-YOLO*l*) and compared with the different sized YOLOv5 network models using a single data source, the object detection performance is significantly improved, especially in scenes with blurred or dim light. MS-YOLO greatly improves the detection accuracy on the premise of ensuring the speed and simultaneously solves the problem of partial information feature waste when using multiple sensors to detect the surrounding environment in automatic driving.

In terms of the environmental perception of unmanned driving, this paper believes that multi-sensor fusion will be the future development direction. In this paper, deep learning is mainly used to achieve the fusion of radar data and image data for object detection, which is used for an intelligent vehicle to sense the surrounding environment. However, the algorithm proposed in this paper does not fully reflect all the effective information obtained by millimeter-wave radar, such as accurately obtaining the specific speed of each object. In addition, in the design of radar data mapping transformation neural network, it is not sure how many layers are set in the network and how many neurons are set in each layer, the performance is better. In the future, we will try to obtain the radar data corresponding to an object to determine the specific position and speed of each object and use the automated tuning method in the design of the network architecture to design a model more in line with the use scenario.

References

[1] M. Jones and P. Viola, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, p. 87, 2001.

[2] 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.

[3] C. Kwan and B. Budavari, "Enhancing small moving target detection performance in low-quality and long-range infrared videos using optical flow techniques," *Remote Sens.*, vol. 12, no. 24, p. 4024, Dec. 2020.

[4] C. Kwan and J. Larkin, "Detection of small moving objects in long range infrared videos from a change detection perspective," *Photonics*, vol. 8, no. 9, p. 394, Sep. 2021.

[5] C. Kwan and B. Budavari, "A high-performance approach to detecting small targets in long-range low-quality infrared videos," *Signal, Image Video Process.*, vol. 16, no. 1, pp. 93–101, Feb. 2022.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[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] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[9] 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.

[10] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[12] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? Modeling dog behavior from visual data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4051–4060.

[13] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.

[14] M. S. Uddin, R. Hoque, K. A. Islam, C. Kwan, D. Gribben, and J. Li, "Converting optical videos to infrared videos using attention GAN and its impact on target detection and classification performance," *Remote Sens.*, vol. 13, no. 16, p. 3257, Aug. 2021.

[15] C. Kwan, B. Chou, J. Yang, and T. Tran, "Deep learning based target tracking and classification for infrared videos using compressive measurements," *J. Signal Inf. Process.*, vol. 10, no. 4, pp. 167–199, 2019.

[16] M. Beauvais and S. Lakshmanan, "CLARK: A heterogeneous sensor fusion method for finding lanes and obstacles," *Image Vis. Comput.*, vol. 18, no. 5, pp. 397–413, Apr. 2000.

[17] B. Steux, C. Laurgeau, L. Salesse, and D. Wautier, "Fade: A vehicle detection and tracking system featuring monocular color vision and radar data fusion," in *Proc. Intell. Vehicle Symp.*, Jun. 2002, pp. 632–639.

[18] L. Bombini, P. Cerri, P. Medici, and G. Aless, "Radar-vision fusion for vehicle detection," in *Proc. Int. Workshop Intell. Transp.*, vol. 3, 2006, pp. 65–70.

[19] J.-G. Wang, S. J. Chen, L.-B. Zhou, K.-W. Wan, and W.-Y. Yau, "Vehicle detection and width estimation in rain by fusing radar and vision," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1063–1068.

[20] V. John and S. Mita, "RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments," in *Image and Video Technology*, C. Lee, Z. Su, and A. Sugimoto, Eds. Cham, Switzerland: Springer, 2019, pp. 351–364.

[21] V. John, M. K. Nithilan, S. Mita, H. Tehrani, R. S. Sudheesh, and P. P. Lalu, "SO-Net: Joint semantic segmentation and obstacle detection using deep fusion of monocular camera and radar," in *Image and Video Technology*, J. J. Dabrowski, A. Rahman, and M. Paul, Eds. Cham, Switzerland: Springer, 2020.

[22] 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.

[23] K. Kowol, M. Rottmann, S. Bracke, and H. Gottschalk, "YOdar: Uncertainty-based sensor fusion for vehicle detection with camera and radar sensors," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 177–186.

[24] X. Dong, B. Zhuang, Y. Mao, and L. Liu, "Radar camera fusion via representation learning in autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1672–1681.
[25] J. Kim, D. S. Han, and B. Senouci, "Radar and vision sensor fusion for object detection in autonomous vehicle surroundings," in *Proc. 10th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2018, pp. 76–78.
[26] Q. Feng, S. Qi, J. Li, and B. Dai, "Radar-vision fusion for correcting the position of target vehicles," in *Proc. 10th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, vol. 2, Aug. 2018, pp. 352–355.
[27] T. Wang, N. Zheng, J. Xin, and Z. Ma, "Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications," *Sensors*, vol. 11, no. 9, pp. 8992–9008, Sep. 2011.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.
[30] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision—ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 346–361.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 448–456.
[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[34] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

Yunyun Song received the B.S. degree in automation from Southwest Jiaotong University, China, in 2020, where she is currently pursuing the M.S. degree with the Department of Control Theory and Control Engineering. Her current research interests include object detection based on millimeter-radar and camera fusion for autonomous vehicles.

Zhengyu Xie received the M.S. degree in control engineering from Southwest Jiaotong University, China, in 2021. He currently works in development at Tencent.

Xinwei Wang received the B.S. degree in network engineering from Southwest Jiaotong University, China, in 2021, where he is pursuing the master's degree. His research focuses on the object detection of the autonomous driving.

Yingquan Zou received the Ph.D. degree in signal and communications engineering from Southeast University in 2012. He is currently an Associate Professor with Southwest Jiaotong University. His current research interests include intelligent information acquisition and control, autonomous driving technology, and new energy technology. His team has long cooperated with enterprises in product development and design. His several scientific achievements have been got the promotion and application by enterprises, creating better economic, and social value.