IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2023, accepted 26 June 2023, date of publication 3 July 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3291398

RESEARCH ARTICLE

Nighttime Pedestrian Detection Based on a Fusion of Visual Information and Millimeter-Wave Radar

WEI ZHAO, TINGTING WANG, AO TAN, AND CONGCONG REN
College of Vehicle and Traffic Engineering, Henan University of Science and Technology, Luoyang, Henan 471000, China
Corresponding author: Wei Zhao (zhaowei@haust.edu.cn)

ABSTRACT A target detection algorithm based on the fusion of vision sensors and millimeter wave radar data can effectively improve the safety of self-driving vehicles. However, a single sensor cannot obtain comprehensive category status information on the target in the nighttime environment. To improve pedestrian detection in nighttime traffic scenarios, this paper proposes a nighttime pedestrian detection method based on the fusion of infrared vision information and millimeter wave (MMW) radar data. The lateral localization and category features of the target are obtained using the improved YOLOv5 deep learning algorithm, the distance and velocity information of the target is obtained by preprocessing the MMW radar acquisition data, the pedestrian target is tracked by using the extended Kalman filter for MMW radar detection, the projection of the valid radar target point on the Infrared Radiatio (IR) image is completed according to the spatiotemporal fusion, and then the correlation gate method is used to correlate the data, and the visual information is inherited to the radar points to get the target multimodal information, and the success associated valid target sequences are weighted to obtain the accurate target position. Finally, a decision-level fusion algorithm framework is proposed to complete the output of pedestrian multimodal information in nighttime traffic scenes. Theoretical analysis and experimental results show that the accuracy and robustness of this method for nighttime pedestrian recognition are better than those of a single sensor.

INDEX TERMS Extended Kalman filtering, millimeter wave radar, sensor fusion, target detection, YOLOv5 algorithm.

I. INTRODUCTION

At present, the research related to automotive self-driving technology is getting more and more in-depth, and the perception of the surrounding environment in complex traffic scenes is an important part of self-driving cars. The environment perception system is equipped with sensors such as LIDAR, MMW radar, and cameras to obtain information on the category, location, and speed of surrounding traffic targets, which can effectively reduce the risk of collision and improve the safety performance of self-driving cars.

Target detection is the process of extracting object classes and locations based on the observation information from different sensors. As a key step in an autonomous driving environment sensing system, target detection techniques based on deep learning have received extensive attention and research. Among them, single-stage target detection algorithms include YOLO [1], [2], [3], [4], [5], SSD [6], and RetinaNet [7]. The two-stage target detection algorithms are mainly based on R-CNN [8] and Faster R-CNN [9] as the main research. In recent years, the YOLO series algorithms have good detection accuracy and detection speed in target detection. However, the current target detection task mainly focuses on visible images, but visible images are greatly affected by the environment, especially in extremely bad weather and nighttime conditions, and visible images lose some details, leading to degradation of target detection performance.

Radar, as a common sensor, can obtain information such as distance and speed of the detected object through

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

electromagnetic signal processing and Doppler effect measurements [10] but is hardly applicable to classification tasks. The camera is an excellent sensor suitable for object detection and classification. For scenarios such as bad weather, night scenes, and obstacle occlusion, a single sensor cannot obtain comprehensive position status information about the target. Sensor fusion can make full use of the advantages of different sensors to make more reasonable control decisions. Therefore, sensor fusion has become a popular direction for self-driving vehicle research.

Current research on target detection and recognition based on MMW radar and vision fusion has the following problems and shortcomings: for night and low light scenes, the visible camera detection effect will be greatly affected, and the vision sensor cannot provide valid information for sensor fusion. If the region of interest is defined using the target location and state information detected by radar, the valid target may be affected by redundant targets. If different sensor information is fused directly, the fusion detection performance is weak and biased.

To solve the above problems and improve the performance of nighttime pedestrian detection, we use infrared cameras and MMW radar as sensors for nighttime target detection. The infrared camera can acquire target category information in the nighttime environment, which can make up for the false detection caused by the inaccurate location of the radar detection target, but there will be a situation of missed false detection in the complex background environment. The radar can dynamically capture the target ahead, which can compensate for the performance deficiency of the infrared camera to a certain extent.

In this paper, we propose a nighttime pedestrian target detection method based on the fusion of an infrared camera and MMW radar sensors, taking full advantage of the performance of the infrared camera and MMW radar. The method uses infrared cameras and MMW radar to simultaneously acquire forward target information. isual information processing is improved based on the YOLOv5 model, the CVC infrared dataset is trained using a deep migration learning approach, and pre-training weights are used to perform pedestrian detection on the acquired nighttime infrared images. The MMW radar acquisition data is preprocessed to filter out valid radar targets. For the target loss problem caused by radar jitter and target occlusion, the valid radar target is tracked using extended Kalman filtering, and the filtered radar target information is projected onto the IR image to generate a radar detection target frame centered on the projected point. Finally, a decision-level fusion of the two sensor detection targets is performed to achieve nighttime pedestrian detection.

This paper is distinguished by the following main contributions:

- By adding the Ghost module and Squeeze Excitation (SE) module, improving the Spatial Pyramid Poolin (SPP) module, improving the network structure of the YOLOv5 algorithm model, and verifying the good performance of this improved algorithm in infrared image target detection through target detection algorithm comparison experiments.
- A data association method based on visual information is proposed for multimodal information fusion of visual information and radar detection points based on the spatiotemporal calibration of images.
- A decision-level fusion strategy is designed for target detection in the nighttime environment, and it is experimentally verified that the strategy reduces the missed false detection rate of a single sensor and improves the accuracy of target detection.

II. RELATED WORK

With the rapid development of deep learning-based target detection algorithms, more and more deep learning algorithms are being used to process visual information. Liu et al. [11] proposed a new method for vehicle detection in infrared images based on convolutional neural networks. Mahmood et al. [12] used an infrared technique combined with the YOLO algorithm for vehicle and pedestrian detection in complex traffic scenes. Liu et al. [13] inserted an optimized FSAF module into the YOLOv3 detector to complete the detection of infrared vehicles and pedestrians. Zhu et al. [14] proposed a parallel fusion network-based infrared vehicle small target detection method, using parallel residual blocks as the base structure of the backbone network, combined with the improved YOLOv3 algorithm to complete the target detection. To improve the image's small target feature extraction capability, the [15], [16], [17] Research on the YOLOv5 algorithm was started based on the YOLOv3 algorithm. In [18] Jin et al. used the YOLOv5 algorithm to solve the pedestrian detection problem during the driving process of self-driving cars. Kasper-Eulaers et al. [19] applied migration learning to the YOLOv5 algorithm in winter conditions to detect vehicles captured by infrared cameras to predict parking space occupancy.

Sensor fusion can improve target detection and tracking performance to some extent, and in recent years, multi-sensor fusion-based target detection schemes have been increasingly investigated. Kim et al. [20] simultaneously extracts camera view and radar bird's eye view features and then performs object fusion based on the features. Cao et al. [21] investigated a spatial fusion technique for MMV radar and vision sensors, studied the relationship between the coordinate systems covered by MM radar and vision sensors, and implemented the transformation between coordinate systems using matrix transformation methods. Grimm et al. [22] used a minimally distortable function to warp the radar tensor to the camera image and then trained using the camera label information. Lim et al. [23], [24] proposed a new deep learning framework for radar information and camera-captured images to test, train, and extract the radar and camera, respectively. Heinzler et al. [25], [26], [27] demonstrated that different sensors all respond differently in a given situation. Zhu et al. [28] proposed screening the

effective targets of MMW radar to match with the pedestrian targets detected by monocular vision cameras, followed by a data fusion detection method. Li et al. [29] increased the width of the radar point cloud to enhance the spatial information and also designed a method for cross-modal interaction fusion of two modalities using disparate feature attention fusion. Muresan et al. [30] proposed two raw data association methods for sensor fusion and tracking and a fusion framework based on the unscented Kalman filter and data-driven methods to steadily acquire target locations from four types of complementary sensors and verify the semantic classes of super-sensor targets using a fuzzy logic approach. Long et al. [31] proposed a sensor fusion system combining a RGB depth sensor fusion system with vision sensors and MMW radar, using the Mean Shift algorithm to acquire depth information of obstacle contour features and the particle filtering algorithm to fuse color images, depth images, and radar data for output to achieve forward obstacle detection.

III. NIGHTTIME PEDESTRIAN DETECTION BASED ON VISUAL INFORMATION

A. DATA SET PREPARATION

This paper uses the infrared image open-source data set CVC-09, which consists of images taken by a vehicle infrared camera in a driving environment, including 5999 frames of daytime images and 5081 frames of nighttime images. ach sequence is divided into a training set and a test set. The images captured in this dat set are all infrared images, and the image background is complex, containing single-target as well as multi-target images with mutual occlusion, etc., which is suitable as a data set for infrared pedestrian detection. In this study, 2,454 nighttime infrared images are selected from the CVC-09 dat set as the training set, and 573 images are selected as the test set, with the category label ''people, cars'' and the target label information including the width and height of the target box and the coordinates of the center point.

B. IMPROVEMENT OF THE YOLOv5 ALGORITHM

The YOLOv5 algorithm has four target detection network models, 5s, 5m, 5l, and 5x, which are classified according to the size of the memory, but the principles are the same. Among them, YOLOv5-5s is the network with the smallest depth of the network model and the smallest width of the feature map, which has the feature of being lightweight, and this paper adopts YOLOv5-5s as the basic network structure for target detection. The network structure of YOLOv5 can be divided into input side, Backbone (backbone feature extraction network), FPN (enhanced feature extraction network), and YoloHead (classifier and regressor of Yolov5). The Backbone network uses the Focus structure and the Cross Stage Partia (CS) structure. The Focus structure improves the network speed by segmenting the input image. YOLOv5 uses two CSP structures [32]: CSP1_X and CSP2_X. CSP1_X is used for down-sampling in the backbone feature extraction network, and CSP2_X is used to enhance the feature extraction networ. he CSP structure improves the learning ability of the network and reduces operations while ensuring accuracy.

The improvement of neural networks is an important area of neural network learning applications [33], [34]. To improve the performance of nighttime pedestrian detection, the following improvements were made in this study based on the YOLOv5 neural network structure:

1) END LATERAL NEURAL NETWORK (GHOSTNET)

In neural network models, rich or even redundant information in the feature maps is very important to ensure a comprehensive analysis of the input data. Redundant information in feature maps is an important feature of a successful neural network model, but redundancy in feature maps has rarely been considered in neural network model design [35]. The Ghost module can generate a larger number of feature maps using fewer parameters. In the Ghost module, the Ghost feature maps are generated by simple linear operations, which can extract the information behind the intrinsic features quickly and comprehensively. Clearly, unlike traditional convolution, the Ghost module requires fewer parameters and has low computational complexity. Similar to data augmentation, the feature map is also augmented by Ghost convolution. The structure of Ghost convolution is shown in Fig. 1.

2) SIMCSPSPPF MODULE

The SPP module in the last Conv+BN+SiL (CBS) of the YOLOv5 backbone feature network consists of three Maxpool layers of 5×5, 9×9 and 13×13 in parallel, and the Spatial Pyramid Poolin-Fast (SPPF) module consists of three Maxpool layers of 5×5 size seriall. The SPPF module is faster and more efficient in detection compared with the SPP module and can solve the multi-scale problem to some extent. In this paper, the Simplified-CSPSPPF (SimCSPSPPF) module similar to the CSP structure is used instead of the SPP module, and the image feature layer is divided into two parts, one of which is subjected to regular convolution operation and the other to SPPF after convolution, and finally, the two parts are combined, which enhances the network characterization capability and effectively improves the network detection capability with negligible speed degradation. As Fig.2 and Fig.3 are shown.

3) SE ATTENTION MECHANISMS (SQUEEZE-AND-EXCITATION NETWORKS)

Video images captured by infrared cameras usually have low resolution and blurred details, and the network is prone to feature loss when extracting pedestrian features. Adding an attention mechanism to the network structure can effectively enhance the extraction of key information for pedestrian detection by the network. In this paper, we introduce the channel attention mechanism SE module, which can enhance the feature information of important feature layer and weaken the feature information of non-important feature layers to a

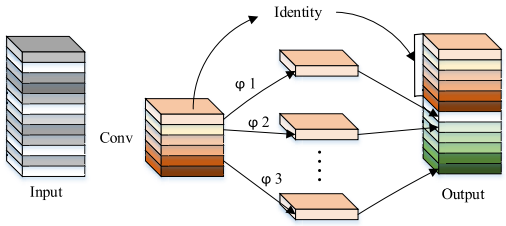


FIGURE 1. Ghost convolution structure.

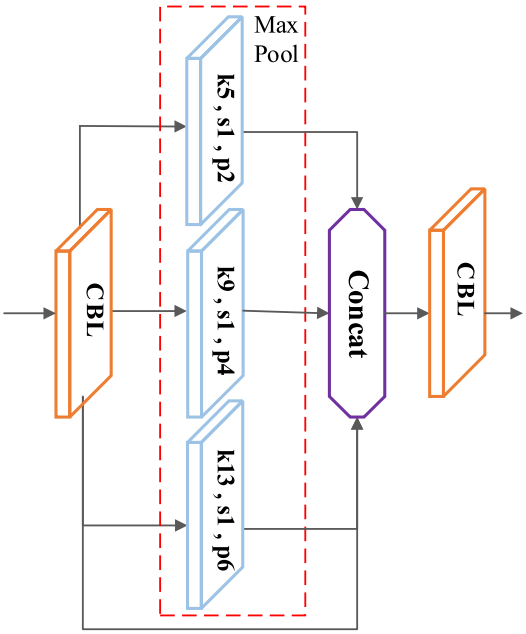


FIGURE 2. SPP module structure diagram.

certain extent to improve the detection capability and accuracy of the network. The SE module is added before the SPPF structure to enhance the detection capability of the network based on the integration of the input image channel features. The structure of the SE layer is shown in Fig.4.

Based on the above improvement strategy, the improved network structure of YOLOv5 is shown in Fig.5.

IV. MMW RADAR TARGET TRACKING

A. RADAR DATA PREPROCESSING

MMW radar can output four categories of valid targets, stationary targets, invalid targets, and null targets simultaneously. According to the radar detection range, combined with the characteristics of interference targets, a reasonable strategy can be developed to reject the interference data and filter out the valid targets. mpty targets and stationary targets can be rejected by using whether the relative angle, relative distance, relative speed, and other values are zero.

Affected by the vehicle bump and swa, air temperature and humidity, electromagnetic wave propagation in the medium, and other factors, radar in the process of work will appear some unstable interference signal called an invalid target, whose main characteristic is the existence of short time, poor

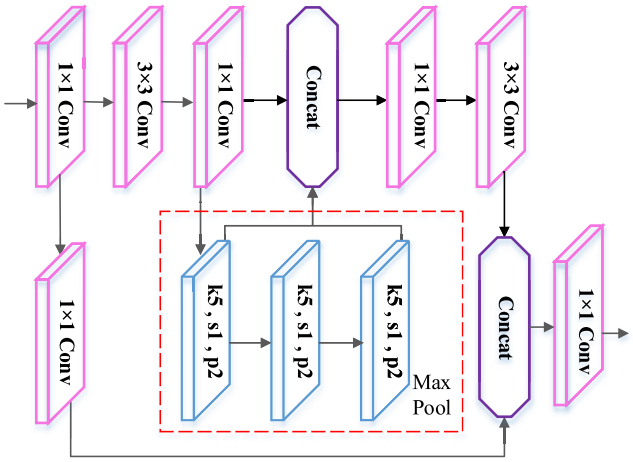


FIGURE 3. SimCSPSPPF module structure diagram.

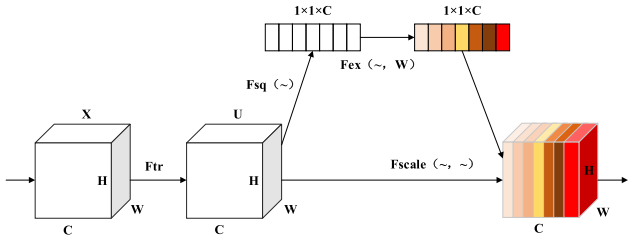


FIGURE 4. SE module structure diagram.

continuity, etc...Therefore, the target point whose life cycle $_{Lt}$ >4 and velocity zVt is not 0 can be set as the real target point. Otherwise, it is rejected as a false target point. The data preprocessing results are shown in Fig.6.

B. EXTENDED KALMAN FILTERING

The radar data after preprocessing still has some interfering target information, this paper uses the Kalman filtering algorithm to filter out non-valid targets to reduce jumping errors, reduce the false alarm rate of radar detection, and continuously track the detected valid targets to obtain more accurate pedestrian movement information.

Extended Kalma Filter (EKF) [36], [37] unlike the conventional Kalman filter, which can handle both linear and nonlinear motion problems, the EKF uses Jacobi matrices to linearize the nonlinear functions. The MMW radar selected in this paper is the German Continental ARS404-21, which uses Frequency Modulated Continuous Wave (FMCW), the nonlinearity of the radar comes from the Doppler effect generated by the high frequency continuous wave emitted during operation, and the Doppler nonlinearity is a first-order approximation of the radar measurement range. The traceless Kalman filter (UKF) [38] is an approximation of the posterior probability density using a deterministic sample approximation state rather than a nonlinear function, and when using MMW radar for speed measurement, tracking the target using the UKF may be unstable and the detection results may

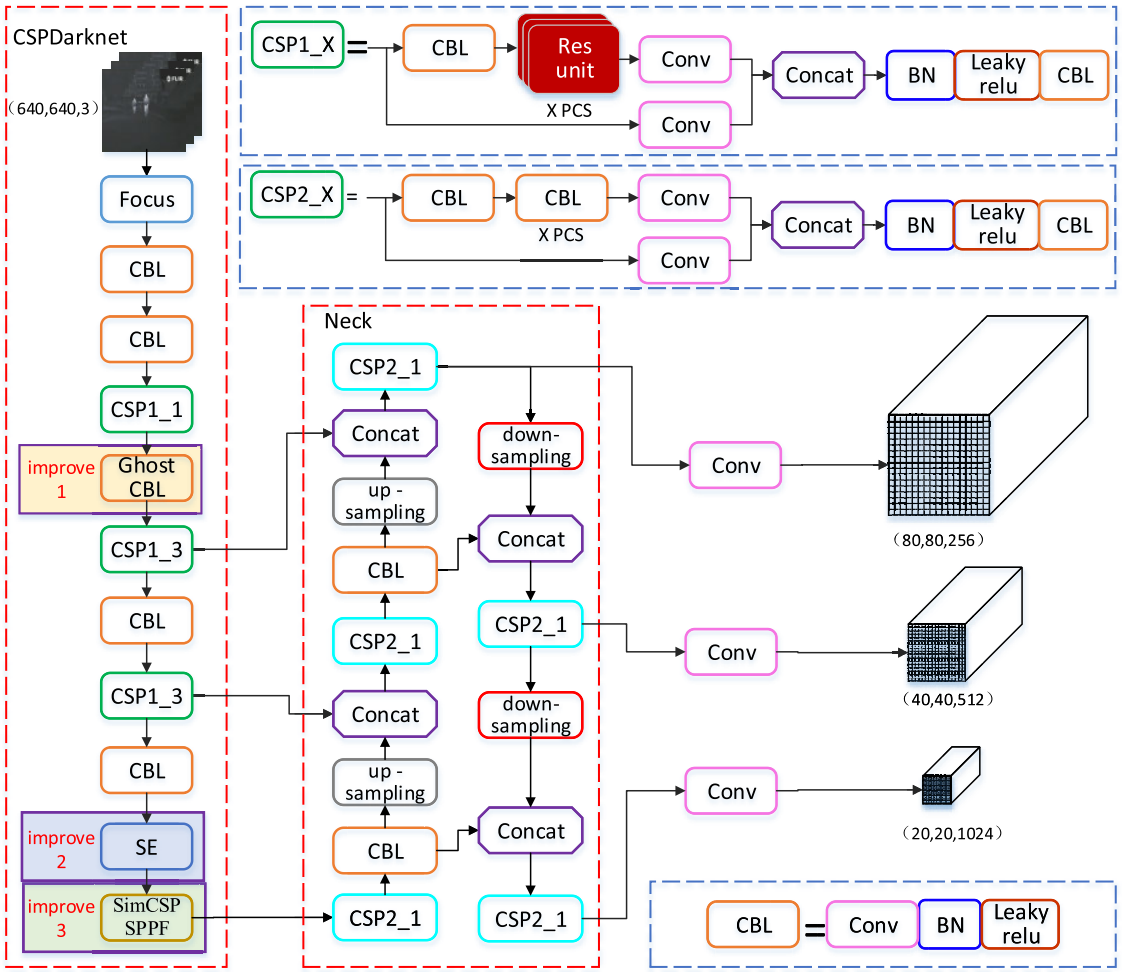


FIGURE 5. YOLOv5 improved network.

be biased. Therefore, in this paper, EKF is chosen to track the radar target based on the uniformly accelerated motion model.

The state vector of the uniformly accelerated moving target can be expressed as follows:

$$X = \begin{bmatrix} R_x & R_y & v_x & v_y & a_x & a_y \end{bmatrix}^T \tag{1}$$

The state transfer matrix is as follows:

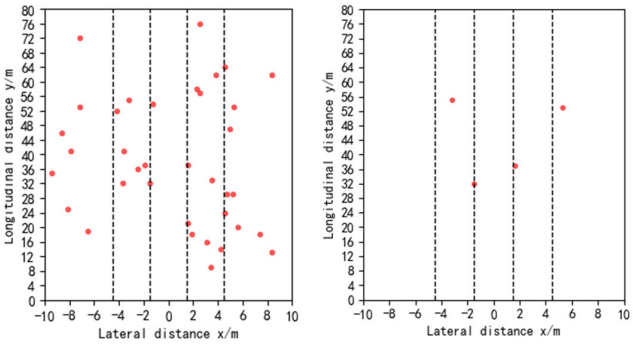
$$F = \begin{bmatrix} 1 & 0 & dt & 0 & \frac{dt^2}{2} & 0 \\ 0 & 1 & 0 & dt & 0 & \frac{dt^2}{2} \\ 0 & 0 & 1 & 0 & dt & 0 \\ 0 & 0 & 0 & 1 & 0 & dt \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The steps of the prediction phase are as follows

$$X_t = FX_{t-1} + W_{t-1} \tag{3}$$

$$P_t = FP_{t-1}F^T + Q_{t\text{-}1} \tag{4}$$

$$Z_t = HX_{t-1} + V_{t-1} \tag{5}$$



(a) Before data preprocessing (b) After data preprocessing

FIGURE 6. Radar data preprocessing. (a) shows the presence of many false invalid radar targets before preprocessing, (b) shows the valid radar targets after preprocessing.

where W_{t-1} is the system process noise; $Q_{t\text{-}1}$ is the covariance of the process nois; Z_t is the measured value at the time of t; H is the observation matrix for the conversion of state quantities into observations; V_{t-1} is the measurement noise, i.e., sensor noise.

The radar is known to collect the distance and velocity in the x-axis and y-axis directions of a target in uniformly accelerated motion, the observations are as follows

$$Z_t = \begin{bmatrix} L_x & L_y & V_x & V_y \end{bmatrix}^T \tag{6}$$

he observation matrix is as follows:

$$H = \begin{bmatrix} 1\,0\,0 & 0\,0\,0 \\ 0\,1\,0 & 0\,0\,0 \\ 0\,0\,1 & 0\,0\,0 \\ 0\,0\,0 & 1\,0\,0 \end{bmatrix} \tag{7}$$

The steps of the update phase are as follows

$$K_t = P_t H^T (H P_t H^T + R_t)^{-1} \tag{8}$$

$$X_t' = X_t + K_t(Z_t - H X_t) \tag{9}$$

$$P_t' = (I - K_t H)\, P_t \tag{10}$$

where, R_t is the variance of the observed noise; K_t is the Kalman gain; X_t' is the optimal estimate at time t; P_t' is the posterior estimated covariance matrix; I is the unit matrix.

Valid targets for radar detection are screened using a life-cycle algorithm, and the current frame of radar observations of the valid targets is fed into the Kalman filter with the algorithmic estimates of the previous frame for state estimation. The MMW radar detection and tracking of dynamic targets for current nighttime traffic scenes is achieved.

V. SENSOR FUSION

The MMW radar and infrared camera have different installation positions, acquisition information forms, and sampling frequencies on the vehicle. It is necessary to establish an accurate coordinate system conversion relationship to realize the spatial fusion of MMW radar and infrared camera, then project the radar detection data points at the same moment onto the infrared image to make a data correlation of the two sensors' information, and finally make a decision and output of multimodal information about pedestrian targets.

A. SPATIAL INTEGRATION

The spatial fusion of MMW radar and the infrared camera is the conversion of measurements from different sensor coordinate systems into the same coordinated system. According to the internal and external parameters of the camera and the sensor installation position, the world coordinate system can be converted to the pixel coordinate system through the camera coordinate system and the image coordinate system, and the conversion relationship between them is shown in Fig.7. The point $P(X_W, Y_W, Z_W)$ in the world coordinate system corresponds to the point $P(x, y)$ in the image coordinate system, and the Z_C axis in the camera coordinate system coincides with the Z axis of the image coordinate system, with the intersection point O of the Z_C axis of the camera coordinate system and the imaging plane as the coordinate origin of the image coordinate system, and the plane formed by the x and y axes of the image coordinate system is parallel to the camera plane $X_C O_C Y_C X_C O_C Y_C$. The O_o point of the pixel coordinate system uO_ov is the upper left vertex of the imaging plane, and the u and v axes are parallel to the x and y axes of the image coordinate system, respectively.

From this, the conversion relationship equation between the radar coordinate system and pixel coordinate system is obtained as follows.

$$\begin{bmatrix} r * sin\theta \\ -H \\ Z_0 + r * cos\theta \\ 1 \end{bmatrix} = \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{11}$$

$$\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{Z_c}{f} & 0 & 0 \\ 0 & \frac{Z_c}{f} & 0 \\ 0 & 0 & Z_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_x & 0 & -p_0 d_x \\ 0 & d_y & -q_0 d_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p \\ q \\ 1 \end{bmatrix} \tag{12}$$

where the radar detects the position of the target is (r, θ), in the world coordinate system (X_W, Y_W, Z_W), R is the rotation matrix between the camera coordinate system and the world coordinate system, T is the translation matrix between the camera coordinate system and the world coordinate system, the Z_c is the Z axis coordinate of the radar observation point in the camera coordinate system, and f is the focal length of the camera, d_x and d_y denote the physical size of each pixel on the x axes and y axes, and (p_0, q_0) is the position of the origin of the image coordinate system in the pixel coordinate system.

B. TIME FUSION

The essence of time fusion is to achieve temporal coherence between two or even more sensors [39]. The sampling frequency of the MMW radar used in this study is 20Hz (sampling period of 50ms) and the sampling frequency of the camera is 30Hz (sampling period of 33.33ms). To ensure sensor time synchronization, the minimum common multiple of 100ms of the sampling period of the two sensors is chosen as the sampling period of the fusion system, and the MMW radar and the camera complete the time fusion by interval sampling.

C. TARGET MATCHING

The radar sensor output is a sparse point cloud containing position, velocity, and other information, and the detection frame is a two-dimensional border in the pixel coordinate system with prediction categories and confidence levels. After completing the spatiotemporal calibration of the IR camera and the MM radar sensor, a key step to reaching fusion is to correlate the data from both sensors to improve the accuracy of target detection [40].

Common data association methods are Global Nearest Neighbor (GNN), Probabilistic Data Association (PDA), and Joint Probabilistic Data Association (JPDA). The core of either association method is the association gate. The association gate is a multidimensional space centered on the predicted state points. In this paper, we choose a rectangular

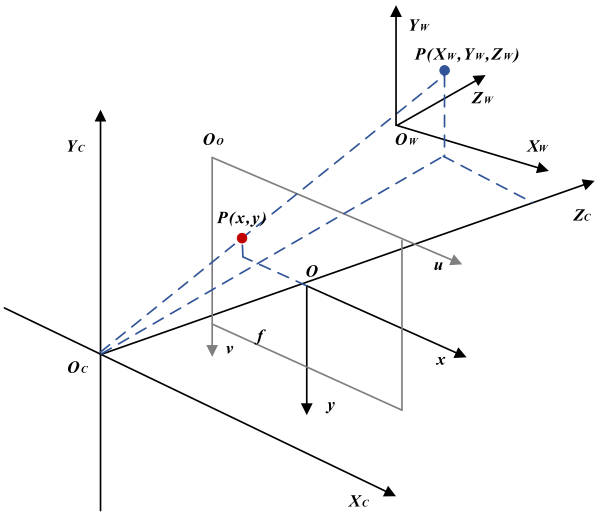


FIGURE 7. Relationship between the world coordinate system, camera coordinate system, and pixel coordinate system.

association gate to realize data association. The association gate is shown in Fig.8, the point is the center of the visual detection frame, which is the center of the association gate, Δx and Δy are the length and width of the association gate, which are usually taken according to the actual driving scene. Combining the characteristics of the experimental object of this study, the radar performance, and the accuracy of detection for small targets, the length and width of the association gate are initially set as follows:

$$\begin{cases} \Delta x = x + \frac{1}{2}x \\ \Delta y = y + \frac{1}{2}y \end{cases} \tag{13}$$

where x and y are the length and width of the visual inspection frame, respectively.

If a radar point falls within the association gate, the radar point is considered to be associated with the visual detection target corresponding to the association gate, and the relationship is shown in the following equation:

$$\begin{cases} \left|X_k - \hat{X}_k\right| \le \dfrac{\Delta x}{2} \\ \left|Y_k - \hat{Y}_k\right| \le \dfrac{\Delta y}{2} \end{cases} \tag{14}$$

where X_k and Y_k are the positions of the radar point in the kth frame in the world coordinate system, and $\hat{X}_k$ and $\hat{Y}_k$ are the associated gate centers of the target in the kth frame image.

$$\Delta d_k = \sqrt{(X_k - \hat{X}_k)^2 + (Y_k - \hat{Y}_k)^2} \tag{15}$$

where Δd_k is the distance between the MMW radar and the infrared camera detection target.

The radar target corresponding to $\min(\Delta d_k)$ is selected as the target that is successfully matched with the visual detection target. The radar point can inherit the attributes of the corresponding visual detection frame according to the above association relationship and complete the target matching between radar and infrared camera to realize the data information fusion of different sensors.

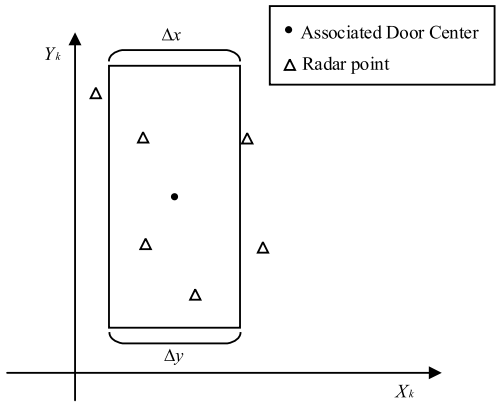


FIGURE 8. Set up an associated gate.

D. GOAL FUSION DECISION

After the training of a deep learning algorithm based on visual information and effective target filtering of millimeter wave radar is completed, the two sensor data sets are target matched to obtain the radar target points inherited from visual information. Due to the low contrast between the infrared image and the surrounding environment and possible occlusion, the visual detection algorithm may produce missed or false detections. To solve this problem, this paper proposes a decision-level information fusion method, which calculates the Intersection over Union (IoU of target detection frames obtained by MMW radar and camera for the decisive judgment of the detection result output.

The *Io* calculation formula is as follows:

$$IoU = \frac{S_r \cap S_R}{S_r \cup S_R} \tag{16}$$

where S_r is the camera detection target frame area and S_R is the MMW radar detection tracking target frame area.

Considering the large lateral distance detection error of MMW radar and the large vertical distance detection error of the IR camera, the weighting strategy to suppress the respective errors of the two sensors to obtain more accurate position information of the target is as follows:

$$\begin{cases} x = \frac{x_r \varepsilon_{cx}^2}{\varepsilon_{rx}^2 + \varepsilon_{cx}^2} + \frac{x_c \varepsilon_{rx}^2}{\varepsilon_{rx}^2 + \varepsilon_{cx}^2} \\ y = \frac{y_r \varepsilon_{cy}^2}{\varepsilon_{ry}^2 + \varepsilon_{cy}^2} + \frac{y_c \varepsilon_{ry}^2}{\varepsilon_{ry}^2 + \varepsilon_{cy}^2} \end{cases} \tag{17}$$

where $\varepsilon_{cx}, \varepsilon_{cy}, \varepsilon_{rx}, \varepsilon_{ry}$, denotes the average error in the x and y directions of the target positions detected by the camera and radar, respectively.

The errors of the MMW radar and IR camera in the x-axis and y-axis directions satisfy the following conditions:

$$\begin{cases} \varepsilon_{rx} < \varepsilon_{cx} \\ \varepsilon_{cy} < \varepsilon_{ry} \end{cases} \tag{18}$$

The algorithmic framework for the fusion of vision sensors with MMW radar sensors is shown in Fig.9.

After completing the target matching based on spatiotemporal fusion and data correlation during pedestrian

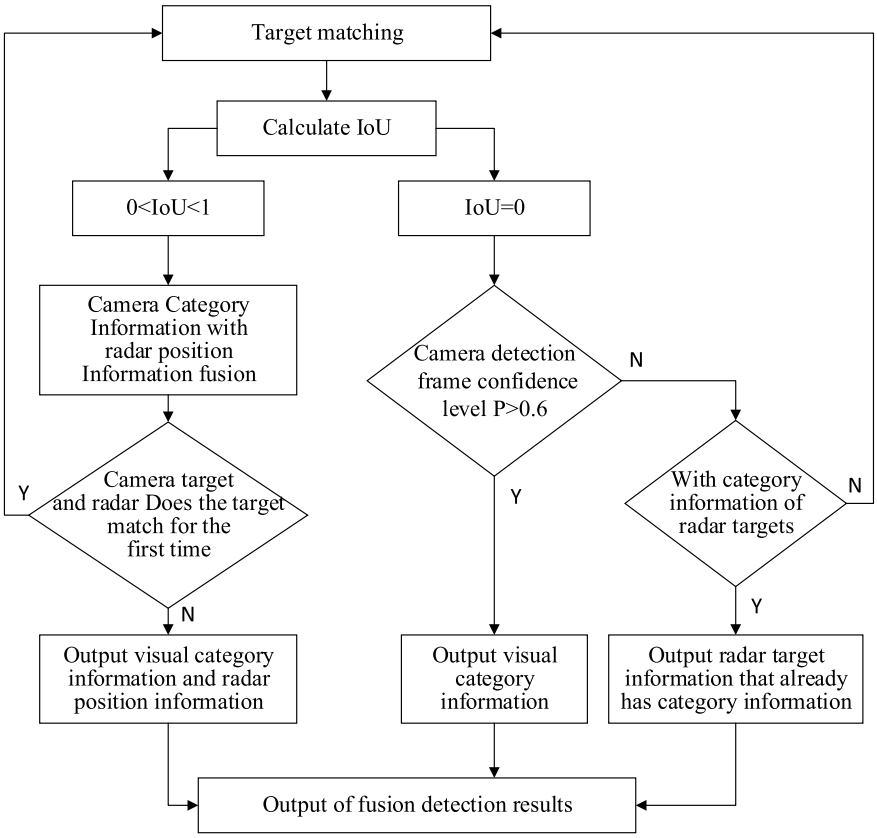


FIGURE 9. Decision-level convergence framework.

TABLE 1. Experimental platform configuration.

Hardware Configuration	CPU	AMD Ryzen 7 5800H
	Memory	16G
	Video Cards	AMD Radeon（TM） Graphics
Software Configuration	Operating System	Windows 10
	CUDA	10.2
	OpenCV	3.4.2
	Python	3.6

TABLE 2. Comparison of test results of different modules for improvement.

Improvements Mode	P (%)	Map (%)	Recall (%)	FPS
Before improvement	88.42	89.28	85.21	29
Ghost Modules	87.60	88.30	84.53	38
SimCSPSPPF Modules	89.06	90.14	85.90	32
SE Module	90.47	89.59	88.36	28

movement, the integrated decision output of different sensor information is performed. The algorithmic framework for the fusion of vision sensors with MMW radar sensors is shown in Fig.9. The visual detection target frame and the radar detection target frame are calculated IoU. When $0 < IoU < 1$, the camera category information is fused with the radar position information, if the camera target and the radar target are matched for the first time then the camera category information is added to the radar detection frame, otherwise it means that the radar and the camera continuously and stably detect the same target, and the visual detection category information is output, containing the target position and status information detected by the MMW radar; when $IoU = 0$ and the camera detection frame confidence $P > 0.6$, make up for the radar missed detection, output the visual category information; when $IoU = 0$ and camera detection frame confidence $P < 0.6$, if the radar target already has category information, then the radar target with category information is not successfully matched with the visual detection frame, make up for the visual detection frame miss detection and output the radar detection target that already has category information, if the radar target does not have category information, then the target matching is performed again.

VI. EXPERIMENTS AND ANALYSIS OF RESULTS

A. EXPERIMENTAL CONFIGURATION

The experiments were conducted using the infrared camera model FLIR PathFindIR and the MMW radar model ARS404-21, with Python as the programming language and training and testing in Pycharm. he experimental platform configuration is shown in Table 1.

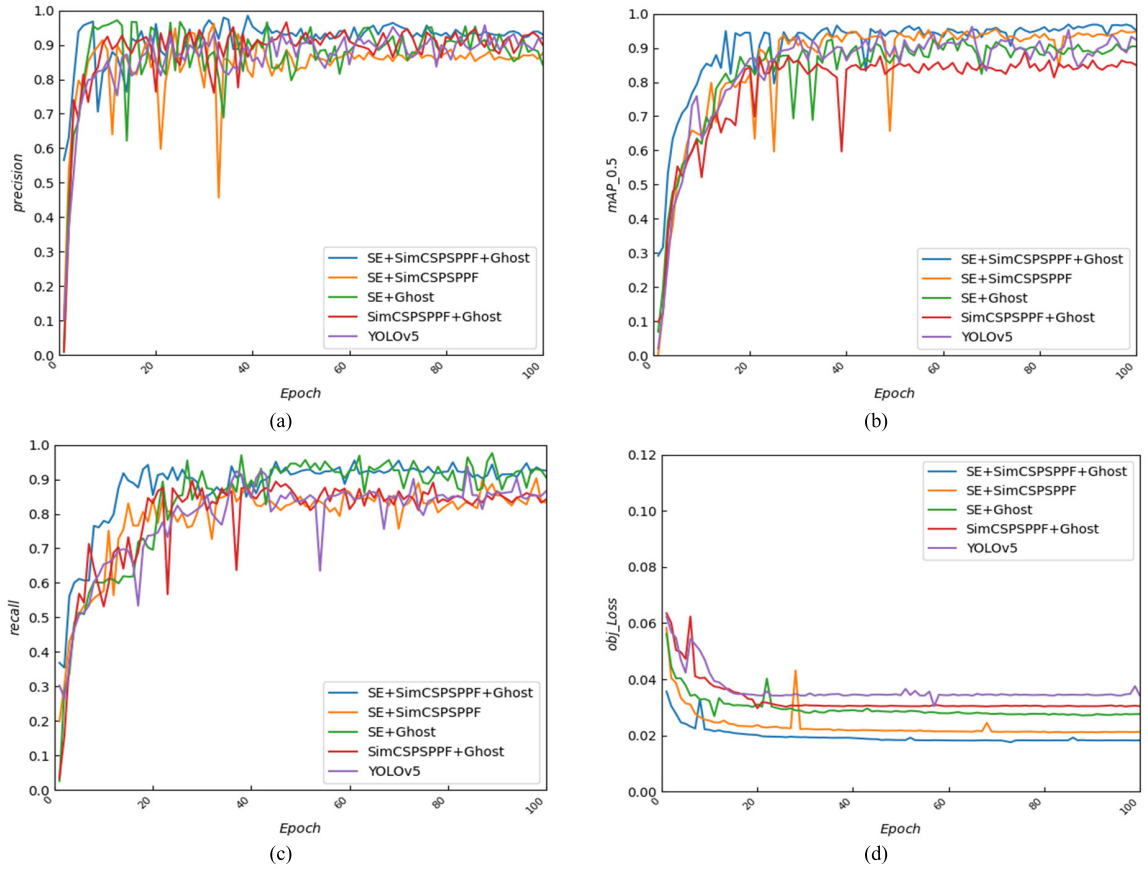


FIGURE 10. Comparison of detection results of improved module combinations.

B. EXPERIMENTS ON NIGHTTIME PEDESTRIAN DETECTION BASED ON VISUAL INFORMATION

In order to compare and verify the improvement effect of the model, individual modules are added to the YOLOv5 algorithm to analyze the detection results before and after the improvement, and a combination of individual modules is added to the YOLOv5 algorithm to evaluate the network quality with precision, Recall, and Map as performance metrics. As Table Table 2 shows.

For Table Table 2, improving Ghost convolution alone, the SimCSPSPPF module, and adding the attention mechanism SE module can affect the detection accuracy to different degrees. Among them, improving the Ghost module has a 0.82% decrease in accuracy but improves the detection speed from 29 FPS to 38 FPS, and improving the SE module has a 2.05% increase in accuracy, but the change in detection speed is not obvious. Although the change in Map value is not obvious when improving the SE module alone, the combination of Map value with other modules has a significant effect on improving Map value. As shown in Fig.10, the Ghost convolution, SimCSPSPPF module, and SE module combination has the lowest target loss value and the highest accuracy. Although the recall rate after the combination is not the highest, it has the smallest fluctuation and is more stable after 50 rounds. The experimental results show that using a combination of three modules effectively improves the YOLOv5 algorithm.

TABLE 3. Comparison of detection results of different algorithm models.

Testing Method	P (%)	Map (%)	Recall (%)	FPS
Faster R-CNN	82.37	80.02	83.19	15
YOLOv3	88.20	86.34	80.30	28
YOLOv5	88.43	84.85	85.28	29
YOLOv5-ours	92.13	93.62	93.24	42

To further verify the detection performance of the improved algorithm YOLOv5-ours in this paper, several typical target detection models, including Faster R-CNN, YOLOv3, YOLOv5, and YOLOv5-ours, are compared and experimented wit, and the test results are shown in Table 3.

It can be seen that the improved algorithm YOLOv5-ours used in this paper has a 3.7% improvement in accuracy compared to the YOLOv5 algorithm, and the detection speed has increased from 29 FPS to 42 FPS. ompared to the classical deep learning target detection algorithm YOLOv5-ours has better detection results and detection speed in pedestrian feature extraction, which applie to the sensor fusion detection part of this paper. The processing part of the visual information in the algorithm.

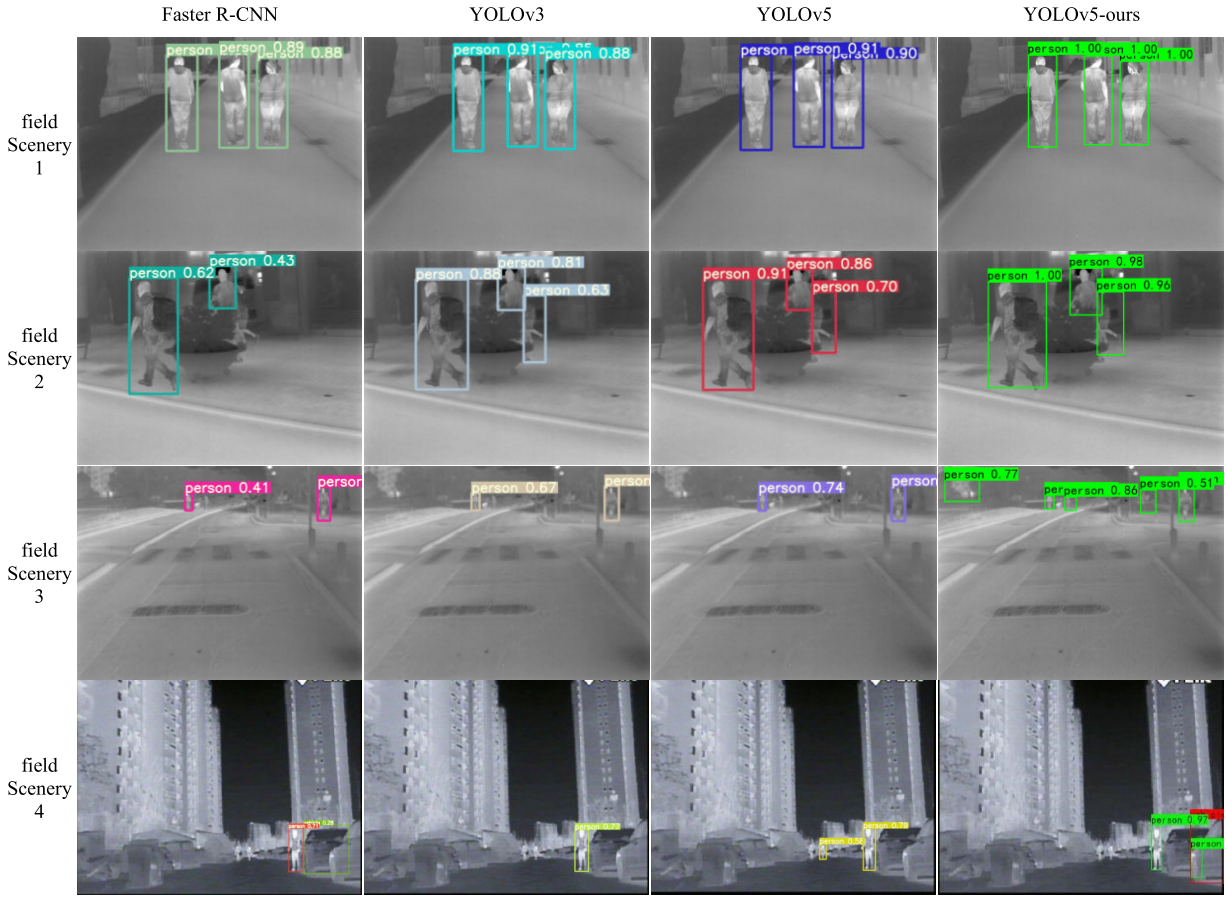


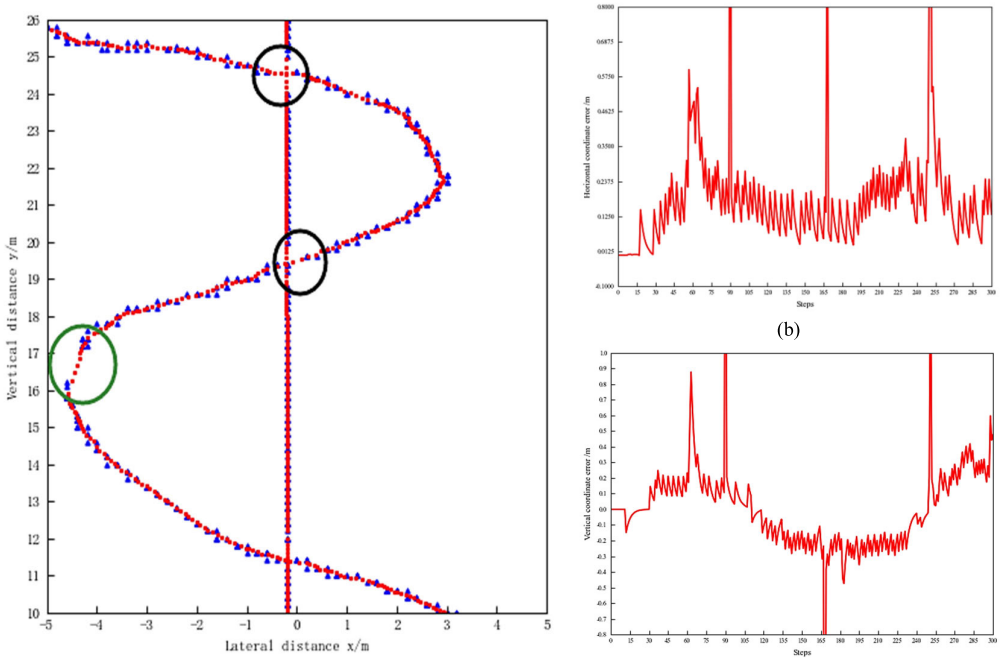
FIGURE 11. Detection results of different algorithms. The road is unobstructed at night in Scene 1. Scenario 2 has an occluded road at night. The contrast of the night environment is low in scene 3. Foggy weather in Scene 4.

In order to be able to compare more clearly the effects of different algorithmic models for nighttime pedestrian detection. Fig.11 shows the detection of pedestrian targets in infrared images in different scenes. From the detection results, it can be seen that the YOLOv5-ours target detection algorithm used in this paper has the highest confidence level and the best detection ffect, especially in scene 2, when obstacles are blocking the pedestrians on the road, YOLOv5-ours can also accurately detect the pedestrian targets. However, in scenario 3, the algorithm in this paper shows weak performance in detecting small targets when the environmental contrast is low. The detection performance of the foggy weather affecting the infrared camera acquisition images in scenario 4 degrades the situation of missed detection and false detection. Since radar information is not easily affected by weather and light, it can be used in the target detection process to assist in improving the robustness and anti-interference capability of target detection.

C. MMW RADAR-BASED TARGET TRACKING

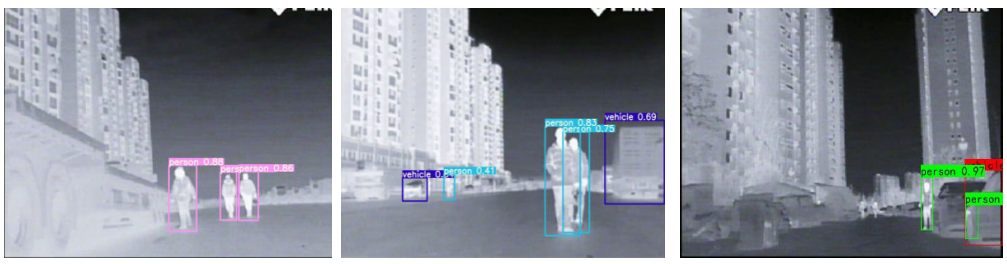
To test the detection and tracking effect of the Kalman filter on the effective radar targets, the uniform linear motion target 1 data and the uniform acceleration curved motion target 2 data were selected from the collected radar data for preprocessing. The initial position and state of target 1 are (10m, 3.2m, 0.25m/s, −1m/s, -0.61m/s^2, 0m/s^2). The initial position and state of target 2 are (10m, -0.2m, −0.12m/s, 0m/s, 0m/s^2, 0m/s^2), with the sudden change of velocity direction added in the moving process and the occlusion of target 1 and target 2 in the motion process.

Inputing the preprocessed radar data into the Kalman-filter, the Kalman filtered tracking trajectory and the horizontal and vertical coordinate errors of the uniformly accelerated curved moving target can be obtained, as shown in Fig.12. From Fig.12 (a), we can see that there is a missing phenomenon for the radar detection of the uniformly accelerated curved motion target, such as the location of the green circle in the figure, and the transverse and longitudinal coordinate errors increase suddenly, as shown in Fig.12 (b) and (c). he phenomenon may be due to the radar detection of the target suddenly turnin or other reasons. When target 1 meets target 2 and there is occlusion, the target will also b lost briefly, as shown in the position of the black circle in Fig.12 (a). The experiments show that the Kalman filter algorithm can compensate for the transient loss of radar detection targets to a certain extent, and accurately predict and track the moving targets.



(a) (b) (c)

FIGURE 12. Experimental diagram of MMW radar filtering. (a) is the Kalman filtered tracking trajectory. The blue points are the radar detection target points, and the red points are the target points after Kalman filtering. (b) is the uniform acceleration curve motion radar target transverse coordinate error. (c) is the longitudinal coordinate error of the radar target with a uniform acceleration curve.



(a) (b) (c)

FIGURE 13. Visual detection results. (a) is a pedestrian unobstructed scene. (b) is a scene with obstructed pedestrians. (c) is a foggy weather scene.

(a) (b) (c)

FIGURE 14. Radar detection results. The red dot is the radar target projection point and the blue box indicates the radar detection target box.

D. SENSOR FUSION DETECTION EXPERIMENTS

To verify the robustness and anti-interference capability of the fusion algorithm target detection, this paper collects urban road pedestrian data and residential pedestrian data to experimentally compare the single-sensor target detection algorithm and the fusion target detection algorithm.

From the comparative analysis of the above target detection experimental results, the visual inspection results in

(a) (b) (c)

FIGURE 15. Detection results of the fusion algorithm. The red box indicates the output with visual detection results, the blue box indicates the output with radar target detection results including visual category information, and the green box indicates the output with radar information and visual information together.

TABLE 4. Comparison of sensor detection results.

Processing Method	Pedestrian Total	Number of positive checks	Number of missed detections	Accuracy rate (%)	Missing detection rate (%)
Radar Testing	1852	1719	133	92.82	7.18
Vision Testing	1852	1695	157	91.52	8.48
Integration Testing	1852	1770	82	95.57	4.43

Fig.13 (c) are affected by weather and lighting, the detection performance is unstable, and false detection can occur on small targets at long distances. In Fig.14 (a), the radar detects a position shift in the target frame, and in Fig.14 (b) the radar has only one target match point for occluded pedestrian targets. In Fig.15 fusion detection results show that all pedestrian targets in the image can be detected without false detection.

To compare the difference in detection effect between a single sensor and a fusion algorithm, 500 sets of infrared pedestrian images were selected to detect pedestrians using radar detection, visual detection, and the fusion algorithm respectively. The detection results of the three methods are shown in Table 4.

From Table 4, It can be seen that, compared with the single sensor detection algorithm, the fusion algorithm proposed in this paper combines the advantages of MMW radar and vision sensors to make up for the lack of single sensor performance, and the target detection accuracy reaches 95.57%, while effectively reducing the leakage rate and achieving better detection results.

VII. CONCLUSION

To solve the problem of low accuracy of pedestrian target detection by a single sensor at night when the road environment is complex and the contrast of the surrounding environment is low, this paper proposes a nighttime pedestrian detection algorithm based on the fusion of visual sensors and MMW radar. The YOLOv5-ours algorithm is used to detect infrared images of pedestrians at night, which improves the detection accuracy by 3.7% compared to the original YOLOv5 algorithm while increasing the detection speed from 29 FPS to 42 FPS. MMW radar can detect target location and state information, and the EKF tracking algorithm is used to continuously estimate and correct filter noise online for MMW radar observation data. statistical features to detect and track the radar detection targets to improve the reliability and stability of radar data. Then a decision-level fusion detection algorithm is proposed to compensate for the false detection misses of infrared cameras and millimeter-wave radar. This paper designs target detection experiments under different nighttime traffic road scenarios, and the experimental results show that the accuracy of the fusion algorithm for nighttime pedestrian detection can reach 95.57% under nighttime traffic scenarios such as small targets at a long distance, the presence of pedestrian occlusion and foggy weather, which effectively reduces the false detection rate of single sensor target detection and to a certain extent solves the fusion algorithm target detection performance The problem of weak target detection performance of the fusion algorithm is solved to a certain extent, and the accuracy and robustness of nighttime pedestrian detection are improved. The current study cannot be directly applied to self-driving cars. In the next step, we will integrate the sensors into the controller of self-driving cars through embedded devices to achieve dynamic real-time target detection of pedestrians in front of self-driving cars in night scenarios, enriching the application scenarios of this study.

REFERENCES

[1] 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.
[2] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[5] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, 2022.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, Alexander, and C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.
[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[8] 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. (CVPR)*, Jun. 2014, pp. 580–587.
[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[10] S. M. Patole, M. Torlak, D. Wang, and M. Ali, "Automotive radars: A review of signal processing techniques," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 22–35, Mar. 2017.
[11] X. Liu, T. Yang, and J. Li, "Real-time ground vehicle detection in aerial infrared imagery based on convolutional neural network," *Electronics*, vol. 7, no. 6, p. 78, May 2018.
[12] M. T. Mahmood, S. R. A. Ahmed, and M. R. A. Ahmed, "Detection of vehicle with infrared images in road traffic using YOLO computational mechanism," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, 2020, pp. 1–9.
[13] Y. Liu, H. Su, C. Zeng, and X. Li, "A robust thermal infrared vehicle and pedestrian detection method in complex scenes," *Sensors*, vol. 21, no. 4, p. 1240, Feb. 2021.
[14] Z. J. Zhu et al., "A parallel fusion network-based detection method for aerial infrared vehicles with small targets," *J. Photon.*, vol. 51, no. 2, pp. 182–194, 2022.
[15] T. Wu, T. Wang, and Y. Liu, "Real-time vehicle and distance detection based on improved YOLO v5 network," in *Proc. 3rd World Symp. Artif. Intell. (WSAI)*, Jun. 2021, pp. 24–28.
[16] H.-K. Jung and G.-S. Choi, "Improved YOLOv5: Efficient object detection using drone images under various conditions," *Appl. Sci.*, vol. 12, no. 14, p. 7255, Jul. 2022.
[17] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.
[18] X. Jin, Z. Li, and H. Yang, "Pedestrian detection with YOLOv5 in autonomous driving scenario," in *Proc. 5th CAA Int. Conf. Veh. Control Intell. (CVCI)*, Oct. 2021, pp. 1–5.
[19] M. Kasper-Eulaers, N. Hahn, S. Berger, T. Sebulonsen, Ø. Myrland, and P. E. Kummervold, "Short communication: Detecting heavy goods vehicles in rest areas in winter conditions using YOLOv5," *Algorithms*, vol. 14, no. 4, p. 114, Mar. 2021.
[20] J. Kim, Y. Kim, and D. Kum, "Low-level sensor fusion network for 3D vehicle detection using radar range-azimuth heatmap and monocular image," in *Proc. Asian. Conf. Comput. Vis.*, 2020, pp. 1–16.
[21] C. Cao, J. Gao, and Y. C. Liu, "Research on space fusion method of millimeter wave radar and vision sensor," *Proc. Comput. Sci.*, vol. 166, pp. 68–72, Jan. 2020.
[22] C. Grimm, T. Fei, E. Warsitz, R. Farhoud, T. Breddermann, and R. Haeb-Umbach, "Warping of radar data into camera image for cross-modal supervision in automotive applications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9435–9449, Sep. 2022.
[23] T. Y. Lim, A. Ansari, B. Major, D. Fontijne, M. Hamilton, R. Gowaikar, and S. Subramanian, "Radar and camera early fusion for vehicle detection in advanced driver assistance systems," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 1–11.
[24] J. Gao, Y. Zhu, and K. Lu, "Object detection method based on radar and camera fusion," *J. Comput. Appl.*, vol. 41, no. 11, p. 3242, 2021.
[25] R. Heinzler, P. Schindler, J. Seekircher, W. Ritter, and W. Stork, "Weather influence and classification with automotive LiDAR sensors," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1527–1534.
[26] R. Heinzler, F. Piewak, P. Schindler, and W. Stork, "CNN-based LiDAR point cloud de-noising in adverse weather," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2514–2521, Apr. 2020.
[27] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2805–2814.
[28] B. Zhu et al., "A pedestrian detection method based on neural network and data fusion," *Automot. Eng.*, vol. 42, no. 11, pp. 1482–1489, 2020.
[29] Y. Li, T. Shen, and K. Zeng, "3D target detection based on millimeter wave radar point cloud and visual information disparity feature attention fusion," *Laser Optoelectron. Prog.*, vol. 34, no. 1, pp. 26–33, 2023.
[30] 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.
[31] N. Long, K. Wang, R. Cheng, W. Hu, and K. Yang, "Unifying obstacle detection, recognition, and fusion based on millimeter wave radar and RGB-depth sensors for the visually impaired," *Rev. Sci. Instrum.*, vol. 90, no. 4, Apr. 2019, Art. no. 044102.
[32] 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. 390–391.
[33] B. Jiang, X. Ma, Y. Lu, Y. Li, L. Feng, and Z. Shi, "Ship detection in spaceborne infrared images based on convolutional neural networks and synthetic targets," *Infr. Phys. Technol.*, vol. 97, pp. 229–234, Mar. 2019.
[34] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1469–1483, Aug. 2020.
[35] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.
[36] S. Haykin, *Kalman Filtering and Neural Networks*. Hoboken, NJ, USA: Wiley, 2004.
[37] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *IEEE Trans. Autom. Control*, vol. AC-24, no. 1, pp. 36–50, Feb. 1979.
[38] L. Wang, X. H. Cheng, and S. X. Li, "Gaussian and high-order traceless Kalman filtering algorithms," *J. Electron.*, vol. 45, no. 2, pp. 424–430, 2017.
[39] B. Zhang, "Research on obstacle recognition method with millimeter wave radar and machine vision fusion," North China Univ. Sci. Technol., 2021, doi: 10.27108/d.cnki.ghelu.2021.000387.
[40] Y. Zhou, Y. Dong, F. Hou, and J. Wu, "Review on millimeter-wave radar and camera fusion technology," *Sustainability*, vol. 14, no. 9, p. 5114, Apr. 2022.