electronics

Article

Vehicle Detection Based on Information Fusion of mmWave Radar and Monocular Vision

Guizhong Cai [1], Xianpeng Wang [1,*], Jinmei Shi [2], Xiang Lan [1], Ting Su [1] and Yuehao Guo [1]

[1] School of Information and Communication Engineering, Hainan University, Haikou 570228, China; gyhao1996@hainanu.edu.cn (Y.G.)
[2] College of Information Engineering, Hainan Vocational University of Science and Technology, Haikou 571158, China
* Correspondence: wxpeng2016@hainanu.edu.cn

Abstract: Single sensors often fail to meet the needs of practical applications due to their lack of robustness and poor detection accuracy in harsh weather and complex environments. A vehicle detection method based on the fusion of millimeter wave (mmWave) radar and monocular vision was proposed to solve this problem in this paper. The method successfully combines the benefits of mmWave radar for measuring distance and speed with the vision for classifying objects. Firstly, the raw point cloud data of mmWave radar can be processed by the proposed data pre-processing algorithm to obtain 3D detection points with higher confidence. Next, the density-based spatial clustering of applications with noise (DBSCAN) clustering fusion algorithm and the nearest neighbor algorithm were also used to correlate the same frame data and adjacent frame data, respectively. Then, the effective targets from mmWave radar and vision were matched under temporal-spatio alignment. In addition, the successfully matched targets were output by using the Kalman weighted fusion algorithm. Targets that were not successfully matched were marked as new targets for tracking and handled in a valid cycle. Finally, experiments demonstrated that the proposed method can improve target localization and detection accuracy, reduce missed detection occurrences, and efficiently fuse the data from the two sensors.

Keywords: vehicle detection; mmWave radar; monocular vision; fusion

Citation: Cai, G.; Wang, X.; Shi, J.; Lan, X.; Su, T.; Guo, Y. Vehicle Detection Based on Information Fusion of mmWave Radar and Monocular Vision. *Electronics* **2023**, *12*, 2840. https://doi.org/10.3390/electronics12132840

Academic Editor: Wojciech M. Zabołotny

Received: 25 May 2023
Revised: 17 June 2023
Accepted: 19 June 2023
Published: 27 June 2023

1. Introduction

Self-driving technology is a significant development direction in the field of intelligent transportation [1]. Achieving self-driving technology on roads requires vehicles to perform autonomous operations and behavioral decisions [2,3]. Self-driving vehicles operate autonomously and make behavioral decisions mainly through sensor perception of the environment. The primary sensors currently used for self-driving environment sensing are Lidar, millimeter wave (mmWave) radar, ultrasonic radar, and cameras [4]. Each sensor has its advantages and disadvantages. Lidar has the advantage of obtaining accurate three-dimensional information about an object but is susceptible to small particles in the air [5] and is more expensive. mmWave radar has a deeper detection range and higher distance and velocity resolution. It can work in all kinds of weather but it is unable to recognize the sort of target and frequently outputs target information with a lot of random noise [6–8]. Ultrasonic radar has the advantages of high penetrating power, simple distance measurement method, and low cost [9,10]. However, it also has the obvious disadvantage that the transmission speed of ultrasound varies in different weather conditions is slower. The advantages of the camera are low cost, rich information, and ease of perception and classification, but it is significantly impacted by light and it is difficult to obtain three-dimensional information about the target.

It can be concluded that single-sensor target detection often has many drawbacks and finds it hard to satisfy the demands of applications in complex road environments. To detect

obstacles on the road more comprehensively, multi-sensor information fusion, including mmWave radar, Lidar, and cameras, are often used [11–15]. Indeed, mmWave radar can detect the location and speed of vehicles and pedestrians, while cameras can provide additional visual information. The fusion of information from these two sensors allows for automatic vehicle control, including technologies such as smart braking, automatic parking, and driverlessness. It provides more accurate and comprehensive information about the vehicle's environment, effectively improving the safety and reliability of self-driving vehicles. On the other hand, both mmWave radar and cameras are inexpensive. Therefore, their fusion is the most widely studied among multi-sensor fusion techniques [16–19].

In the study of information fusion of mmWave radar and vision, as early as 2002, Masaki et al. used the target points obtained by radar projected into the image pixel coordinate system to expand the generated target hypothesis region, then used the traditional visual target detection method for target class identification [20], it can successfully combine data from mmWave radar and cameras. However, the mmWave radar at that time could only detect a few targets, the detection distance was very close, and there was a severe problem of missed detection [21]. To solve the issue of mmWave radar miss detection, some scholars use the same fusion to perform vehicle detection by simultaneously extracting the target hypothesis region from the under-vehicle shadow and mmWave radar points. Compared with [20], in [22], after the mmWave radar points are converted to the pixel coordinate system to acquire the target hypothesis region, then the target hypothesis region is directly fed into the RPN for target classification, which dramatically improves the detection efficiency of the network.

Nowadays, deep learning algorithms for visual object detection have improved in efficiency and detection accuracy as a result of advances in deep learning technology [23,24]. The categories of inspections have grown greatly, and the reliance on the environment has shrunk [25]. So more and more scholars have replaced the traditional visual target detection methods with deep learning visual target detection methods. In [26], a target detection scheme is proposed based on information fusion between mmWave radar and the camera. Where the visual object detection method uses the improved YOLOv5s algorithm, and this scheme is applied to the detection of multi-target obstacles in farmland, which has been proven to significantly improve detection efficiency and detection accuracy.

Based on the above analysis, although all these multi-sensor fusion target detection methods are faster or better than single-sensor target detection methods. Nevertheless, detection accuracy, system robustness, and real-time performance still have many drawbacks and shortcomings. Therefore, a vehicle detection method based on the fusion of mmWave radar and monocular vision is proposed in this paper. This method combines the advantages of mmWave radar and vision and uses the data pre-processing algorithm to filter the mmWave radar data at the detection scene, which can effectively eliminate the influence of interference targets. Then the density-based spatial clustering of applications with noise (DBSCAN) clustering fusion algorithm [27] and the nearest neighbor algorithm are used to correlate the same frame data and adjacent frame data, respectively. This effectively reduces the miss of targets temporarily caused by bumps. Finally, the matched targets are processed by Kalman weighted fusion output algorithm. For targets that are not successfully matched, they are tracked and then output [28]. Experiments show that this method not only achieves accurate detection of vehicle targets but also effectively fuses information from mmWave radar and monocular cameras and acquires accurate and comprehensive target information. The following is a summary of the significant contributions made by the proposed method.

(1) More optimal use of mmWave radar data. The use of data pre-processing algorithms to filter out point cloud data from interfering targets in advance to reduce false detections and missed detections when mmWave radar detects a target;

(2) The use of decision-level fusion can decrease the amount of data processing needed by the system, increase target placement accuracy, and improve system robustness;

(3) The speed of the data processing is enhanced using a one-stage visual target detection technique, and target images of real urban road scenes are used as the training set so that the system can be fully utilized on real roads.

The remaining sections of this article are structured as follows. Section 2 presents an examination of the data fusion model and a description of the issue. In Section 3, some specific steps and algorithms for information fusion are described. The experimental findings of this study are displayed in Section 4. In Section 5, conclusions are made at the end.

2. Data Fusion Model

The data fusion model based on mmWave radar and monocular vision is shown in Figure 1.

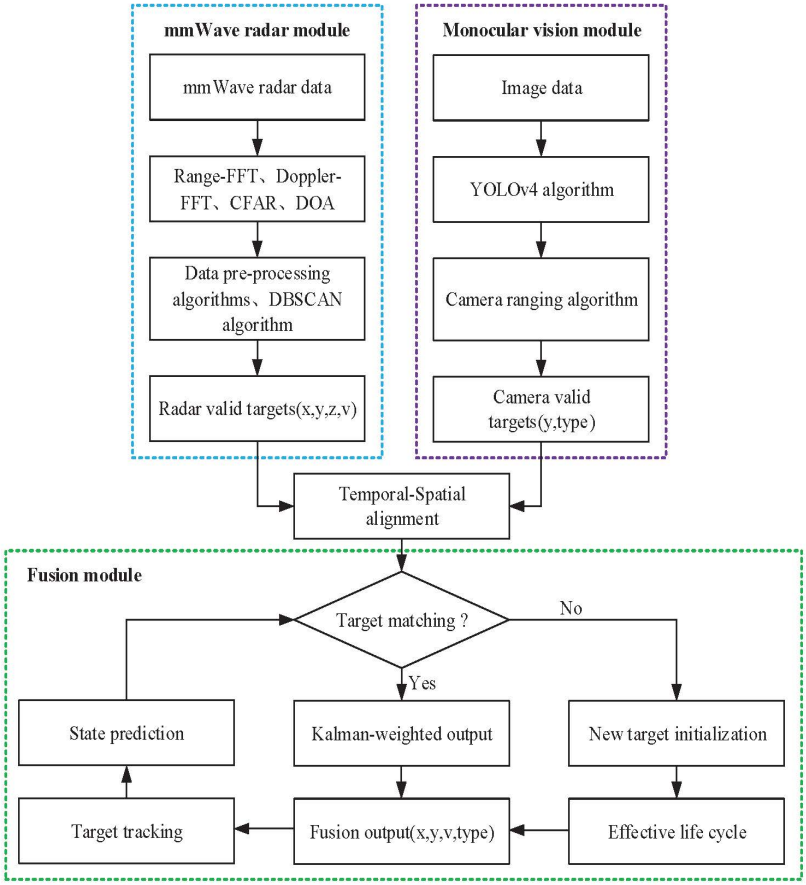

Figure 1. Data fusion model based on mmWave radar and monocular vision.

The data fusion model is mainly divided into the mmWave radar module, the monocular vision module, and the fusion module. In the mmWave radar module, some 3D detection points can be obtained after processing by range-FFT, doppler-FFT, constant false alarm rate detector (CFAR), and direction of arrival (DOA) algorithms. The 3D detection points are processed by the data pre-processing and the DBSCAN algorithms, and the sequence of valid radar targets (x, y, z, v) is output. The image data are processed by the YOLOv4 and monocular camera ranging algorithms, and the sequence of valid monocular vision targets (y, $type$) is output, where x, y, and z are the location information of valid targets. v and $type$ are the velocity and type information of the valid target, which are used as supplementary information. The valid targets detected by the two sensors are then output to the fusion module for target matching. For the successfully matched target, the information of the target is output using Kalman-weighted fusion. The output information

includes the target coordinates, *type*, and velocity. For the successfully matched target, they will be marked as new targets for tracking and handled in a valid cycle.

3. Realization of Fusion

3.1. Temporal Alignment

The timely synchronization of data acquisition for the mmWave radar and the monocular camera is essential to ensuring the correctness of the fusion. The sampling period of the mmWave radar is 40 ms. The sampling period of the monocular camera is about 33 ms. Based on the maximum speed specified for urban roads in China, calculations show that the difference in the images captured by the camera within the 33 ms acquisition period is slight. The change in the longitudinal distance of the target motion converted to the image pixel coordinate system is also tiny within the error time generated by the time matching. So the time-matching method used in this paper uses the data measurement time of the lower-frequency mmWave radar as a reference to achieve backward compatibility with the higher-frequency camera data, as shown in Figure 2.

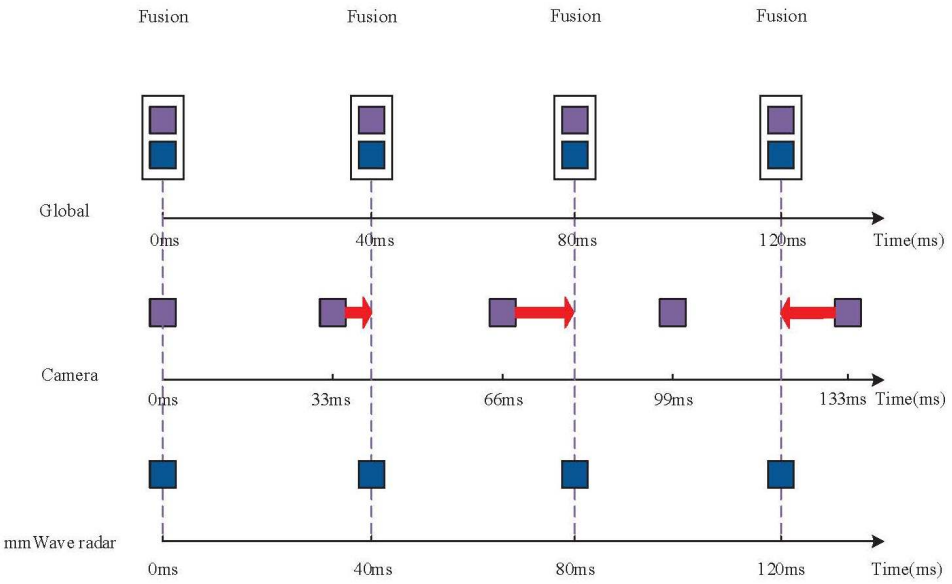

Figure 2. Temporal alignment of sensors data.

3.2. Spatio Calibration

In the case where the mmWave radar data and the monocular camera data have been time matched, in order to accurately convert the mmWave 3D detection points into pixels on the image to achieve the fusion of the two sensors, we must determine how the pixel coordinate system and the mmWave radar coordinate system are related [22,29–33]. Some sort of intermediary coordinate system was required to aid in the transformation, as shown in Figure 3.

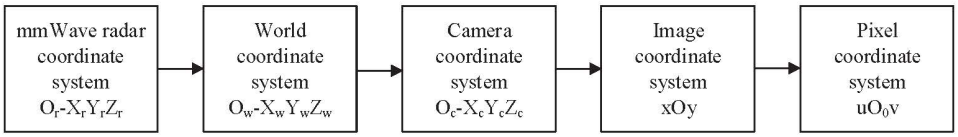

Figure 3. Transformation between different coordinate systems.

The conversion process between the mmWave radar coordinate system and the pixel coordinate system can be completed in two primary steps. The mmWave radar coordinate system must first be converted to the world coordinate system. The second step is to construct a link between the coordinate systems of the world and pixels. As shown in Figure 4, the mmWave radar coordinate system $O_r - X_rO_rZ_r$ can be translated or rotated to obtain the world coordinate system $O_w - X_wO_wZ_w$. The coordinate transformation relationship satisfies:

$$X_w = Rcos\alpha cos\theta \tag{1}$$

$$Y_w = Rsin\theta - H \tag{2}$$

$$Z_w = Z_0 + Rcos\alpha sin\theta, \tag{3}$$

where H is the distance between $X_rO_rZ_r$ plane and $X_wO_wZ_w$ plane. The point P is a point in the mmWave radar detection range, and the O_rZ_r axis is in the $Y_wO_wZ_w$ plane. The O_rX_r axis is separated from the $X_wO_wY_w$ plane by Z_0.

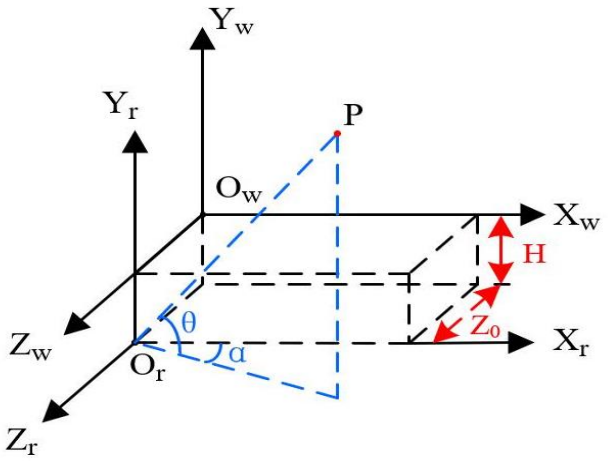

Figure 4. mmWave Radar detection plane and world coordinate system.

As shown in Figure 5, the following coordinate system connection is the fundamental method for establishing the link between the world coordinate system and the pixel coordinate system [34,35].

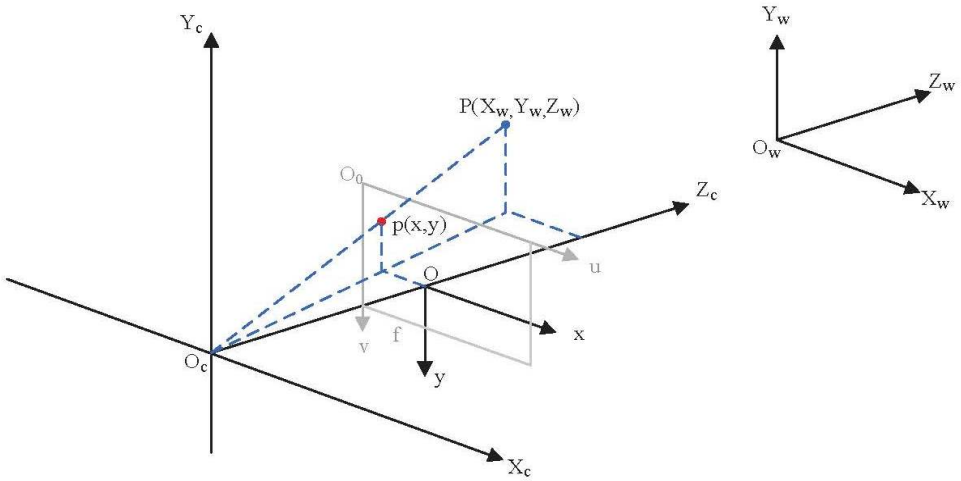

Figure 5. Schematic of coordinate system relationships.

The point $P(X_w, Y_w, Z_w)$ in the world coordinate system $O_w - X_w O_w Z_w$ corresponds to the point $p(x, y)$ in the image coordinate system xOy. To simplify the transformation process, the world coordinate system $O_w - X_w O_w Z_w$ and the camera coordinate system $O_c - X_c O_c Z_c$ are chosen to be the same coordinate system. Then the coordinates of the point P are also (X_c, Y_c, Z_c).

The camera coordinate system $O_c - X_c O_c Z_c$ can be converted to the image coordinate system xOy by perspective projection. Both the pixel coordinate system and the image coordinate system lie on the plane, so only the origin and measurement units need to be converted. As shown in Figure 5, according to the similar triangle principle, we can get:

$$\frac{x}{X_c} = \frac{f}{Z_c} \tag{4}$$

$$\frac{y}{Y_c} = \frac{f}{Z_c}. \tag{5}$$

Thus, the matrix relationship of the transformation of points in the camera coordinate system $O_c - X_c O_c Z_c$ to the image coordinate system xOy can be shown as:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{Z_c} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c. \end{bmatrix} \tag{6}$$

Further, the relationship of points transformed in the image coordinate system xOy to the pixel coordinate system uO_0v can be expressed as:

$$u = x * k + u_0 \tag{7}$$

$$v = y * l + v_0, \tag{8}$$

where k and l are the pixel densities in the x and y directions of the image, respectively, and u_0 and v_0 are the initial pixel offsets in the x and y directions, respectively. Therefore, it is possible to determine the connection between the world coordinate system and the pixel coordinate system by:

$$\begin{aligned} Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} &= \begin{bmatrix} k & 0 & u_0 \\ 0 & l & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ \overrightarrow{0} & 1 \end{bmatrix} \begin{bmatrix} X_\omega \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} kf & 0 & u_0 & 0 \\ 0 & lf & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ \overrightarrow{0} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_1 M_2 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1, \end{bmatrix} \end{aligned} \tag{9}$$

where R and T are the rotation matrix and the translation matrix between the camera coordinate system and the world coordinate system, respectively. Z_c is the Z-axis coordinate of the observation point in the camera coordinate system. f is the focal length of the camera. M_1 and M_2 are the internal parameter matrix and the external parameter matrix of the camera, respectively.

In order to obtain the parameters of the model, the camera needs to be calibrated with a checkerboard grid [36]. The relative position of each corner point in the checkerboard grid was known. A 10 × 7 square grid calibration board with a single square size of 39 mm × 39 mm was selected, and 16 grid images were obtained with a fixed camera position at various angles and distances, as shown in Figure 6.

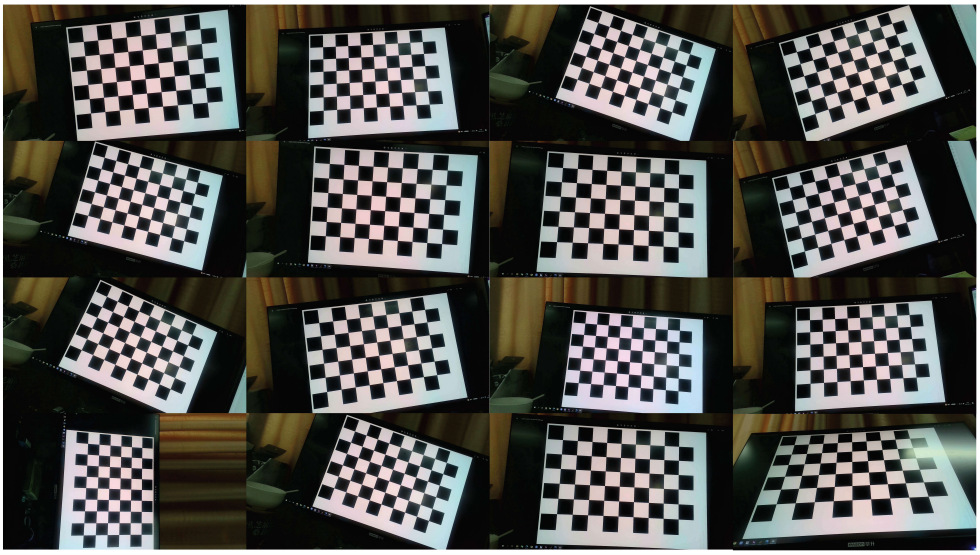

Figure 6. Checkerboard calibration diagram.

Using the camera calibration function integrated into the MATLAB toolbox, the above 16 images were entered and calibrated. Then the actual size of the cells in the checkerboard grid of 39 mm was entered. Finally, the camera's internal parameters, aberration matrix, and external parameters were obtained. The external parameters are related to the installation position of the camera in 3D space and the values are transformed, while the internal parameters and the aberration matrix are unique. After passing the calibration, the internal parameter matrix and the external parameter matrix of the camera were:

$$Internal\ parameter\ matrix: \begin{bmatrix} 2032.1 & 0 & 1338 \\ 0 & 2033.2 & 741.8 \\ 0 & 0 & 1 \end{bmatrix}$$

$$External\ parameter\ matrix: \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 180 \\ 0 & 0 & 1 & 100 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.3. mmWave Radar Data Processing

3.3.1. mmWave Radar Data Pre-Processing Algorithms

mmWave radar data were processed by range-FFT, doppler-FFT, and CFAR algorithms [37]. Then 3D detection points can be obtained through the DOA estimation algorithm [38]. These 3D detection points include the target's 3D detection points of the target and the 3D detection points of the interfering targets, such as roadside trash cans, billboards, guardrails, trees, and other invalid targets generated by other clutter interference, as well as some empty targets. To reduce the amount of post-processing data, it is essential to pre-process the mmWave radar 3D detection points obtained after DOA estimation. For the road area in front of the vehicle, we only focused on the targets in this lane and the left and right lanes so we could set the lateral distance according to the actual lane distance, which can filter out the interference targets on the non-lane next to the road. For the invalid target in the lane, the validity of the five frames of the tracking test was conducted to determine whether it was an invalid target. Whether the target was empty or not, it can be determined by whether its parameter was 0 or not.

3.3.2. Acquisition of Effective mmWave Radar Targets

The reliability of the processed 3D detection points is high. However, there is no correlation between each 3D detection point in the same frame of mmWave radar data and no correlation between 3D detection points in different frames. Therefore, it was necessary to correlate the 3D detection points of mmWave radar to acquire and track each target.

(1) An improved algorithm based on DBSCAN clustering fusion

In the mmWave radar 3D detection points, there may be a case where one target corresponds to multiple 3D detection points. The correlation between mmWave radar data of the same frame was used to determine which 3D detection points are from the same target and gather the target points of the same object together. Currently, the commonly used clustering algorithms in engineering mainly include the K-Means algorithm [39] and the DBSCAN algorithm. The K-Means algorithm is simple in principle but it does not apply to the clustering of mmWave radar 3D detection points because it needs to set the *K* value in advance. That is, the number of targets to be clustered in each frame is given in advance, which does not apply to the actual road situation with unknown targets. The DBSCAN algorithm does not require the number of clustering targets to be set in advance as long as specific parameters are entered. In this paper, a DBSCAN clustering-based Euclidean distance fusion algorithm was designed based on the characteristics of mmWave radar 3D detection points. The specified radius and the minimum point were set according to the actual multiple test data and the specific dimensions of the vehicle. In several data clusters obtained by DBSCAN clustering fusion, the 3D detection points in each cluster were averaged by Euclidean distance to obtain the 3D detection point that can represent the target.

(2) Data association algorithm based on the nearest neighbor method.

Using the association between the data of previous and subsequent frames, it is possible to determine which target data points in the next frame are the continuation of the target data points of the previous frame. The nearest neighbor algorithm [40,41] is one of the most basic methods for data association between previous and subsequent frames. Its essence is to find the point closest to the previous frame data point from the subsequent frame data for the association. Due to its simple principle, high timeliness, and easy implementation, we used the nearest neighbor method to realize the data association between the front and back frames in this paper.

3.4. Monocular Camera Data Processing

3.4.1. Monocular Camera Ranging Algorithm

In this paper, the principle of the monocular camera ranging algorithm was as follows Figure 7.

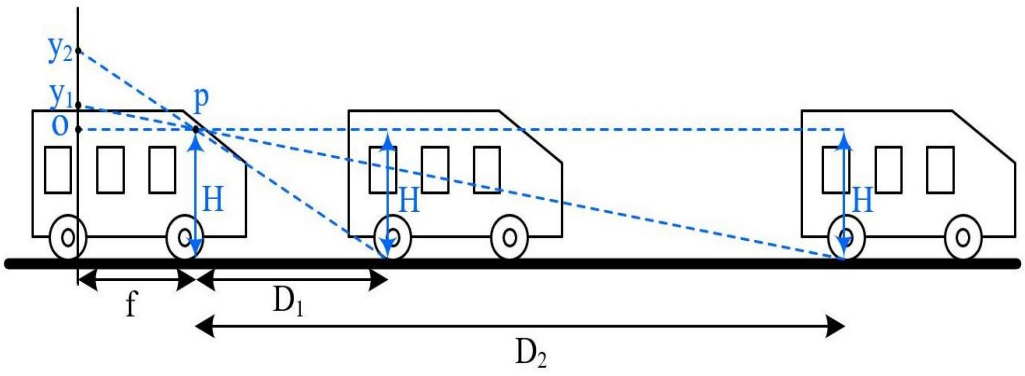

Figure 7. Monocular camera ranging schematic.

p is the camera focus, f is the camera focal length, H is the camera installation height, y_1 is the y-image coordinate of the rear wheel of the B target vehicle and the ground grounding point, and y_2 is the y-image coordinate of the rear wheel of the C target vehicle and the ground grounding point. D_1 and D_2 are the distances from the rear wheels of the targets B and C to the camera, respectively. Using the principle of triangle similarity, we can get the following:

$$\frac{y_1}{f} = \frac{H}{D_1} \tag{10}$$

$$\frac{y_2}{f} = \frac{H}{D_2}, \tag{11}$$

where y_1 and y_2 are the values in the image coordinate system; the unit is mm. The pixel coordinate value in the image also needs to be converted to the image coordinates, and the transformation relation as:

$$y = \frac{Y - dy}{l}, \tag{12}$$

where dy is the offset between the camera axis and the image plane y, and l is the pixel density in the y-direction of the image. The distance D between the target and the camera can be estimated from the pixel height coordinate value of the target. It is expressed as:

$$D = \frac{Hfl}{Y - dy}. \tag{13}$$

3.4.2. Target Detection of the Camera

YOLOv4 [42], as a classical one-stage detection network model, has high detection accuracy and efficiency. YOLOv4's average precision (AP) value is nearly 15% higher than that of the SSD, and the detection speed is almost the same. Compared with the two-stage fast-RCNN algorithm, the detection speed of YOLOv4 is dozens of times faster. In comparison to the previous generation YOLOv3, YOLOv4 improves detection accuracy by approximately 10% while keeping the detection speed unchanged. Among many detection networks, the YOLOv4 combines detection efficiency and accuracy with a relatively high-cost performance. In this paper, the YOLOv4 network was used as a visual target detection algorithm, vehicles in road traffic environments were used as detection objects, 28,000 real vehicle images were used as the training set, and 2000 real vehicle images were used as the test set for the network training of YOLOv4.

3.5. Fusion of Sensors Data

3.5.1. Target Matching

Due to the error caused by testing and conversion and external interference, the target information detected by mmWave radar and camera may differ. Therefore, the valid targets detected by the two sensors must be matched to determine whether the measurements from the different sensors belong to the same target. The matching condition in this paper was mainly the Euclidean distance between the pixel point on the image corresponding to the mmWave radar point cloud data and the midpoint of the visual target detection recognition box, as well as the longitudinal distance of the target calculated by the monocular ranging model and the longitudinal distance of the target measured by the mmWave radar. It can be defined as:

$$\Delta Ed = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \tag{14}$$

$$\Delta d = |d_i - d_j|, \tag{15}$$

where ΔEd is the Euclidean distance between the pixel point on the image corresponding to the mmWave radar 3D detection points and the midpoint of the visual target detection recognition box, u_i is the horizontal coordinate of the image pixel point corresponding

to the mmWave radar 3D detection points, and v_i is the vertical coordinate of the image pixel point corresponding to the mmWave radar 3D detection points. u_j is the horizontal coordinate of the pixel at the center of the rectangular box identified by the target vision, and v_j is the vertical coordinate of the pixel at the center of the rectangular box identified by the target vision. d_i is the longitudinal distance of the target measured by the mmWave radar, and d_j is the longitudinal distance of the target calculated by the camera range measurement algorithm. Then the result obtained according to Equations (14) and (15) are compared with the Euclidean distance threshold and the longitudinal distance threshold. It may be described using the formula below:

$$\Delta Ed < Ed_{th} \tag{16}$$

$$\Delta d < d_{th}, \tag{17}$$

where Ed_{th} is the threshold for pixel Euclidean distance and d_{th} is the threshold for longitudinal distance. If the relationship between the measured values of the two sensors satisfies the Equations (16) and (17), then the two targets are considered to be successfully matched and output using the Kalman-weighted fusion algorithm. Conversely, if the relationship between the measured values of the two sensors does not satisfy any of the Equations (16) and (17), then the match between these two targets is considered unsuccessful.

3.5.2. Target Fusion

When mmWave radar effective targets and effective camera targets are processed by corresponding target matching algorithms, there may be two situations: successful matching and unsuccessful matching. If the target data are successfully matched, the target distance parameters are weighted using a Kalman weighted fusion algorithm to output the distance parameters measured by the two sensors. The following is the computation process:

$$\hat{Z} = (1 - K)Z_1 + K * Z_2, \tag{18}$$

where K is the Kalman gain; it ranges from 0 to 1, i.e., [0, 1]. Z_1 is the measured value of the parameter of the mmWave radar, and Z_2 is the measured value of the parameter of the camera. We need to find the value of K that makes the least standard deviation of the estimated value. Based on the relationship between variance and standard deviation, it can be deduced as follows:

$$\sigma_z^2 = Var((1 - K)Z_1 + K * Z_2). \tag{19}$$

Since the measured values Z_1 of the mmWave and Z_2 of the monocular camera are independent of each other, the standard deviation σ_z^2 can be further obtained as:

$$\sigma_z^2 = (1 - K)^2 * \sigma_1^2 + K^2 * \sigma_2^2. \tag{20}$$

To obtain the minimum value of the variance of the estimated value, it is necessary to find the derivative of K and then make it equal to 0 to obtain the extreme value. After the derivation, the Kalman gain K can be further expressed as:

$$K = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \tag{21}$$

where σ_z^2 is the standard deviation of the estimated value $\hat{Z}$, σ_1^2 is the standard deviation of the measured value Z_1 and σ_2^2 is the standard deviation of the measured value Z_2.

When the effective targets of the mmWave radar and the monocular camera cannot be successfully matched, there are two main scenarios.

(1) Target is detected by the monocular camera but not by the mmWave radar.

The presence of a target was determined by the confidence level obtained by the YOLOv4 algorithm. If the confidence level was greater than 0.8, the target was considered present and the output parameters were temporarily used as measured by the monocular camera ranging algorithm. The data were then matched to the radar data frames in the next few frames.

(2) Target is detected by the mmWave radar but not by the camera.

The mmWave radar data were processed and tracked for five frames before and after. If the target was always present, the output directly used the target parameters detected by the mmWave radar. If the target was lost, the target was considered lost due to bumps and uneven road surfaces, and the target data were discarded directly.

4. Experiments and Results

To verify the superiority of the proposed method for vehicle detection based on the fusion of mmWave radar and monocular vision, the mmWave radar and monocular camera were mounted on the same tripod for a series of experiments. As shown in Figure 8, in the mmWave radar module, a Texas Instruments (TI) cascaded MMW radar system consisting of MMWAVCAS-RF-EVM and MMWAVCAS-DSP-EVM was used for follow-up work [43]. The system relies on the multiple-input multiple-output (MIMO) regime for target detection by transmitting frequency-modulated continuous waves (FMCW). This radar system has 16 receiving and 12 transmitting elements. So, the MMW radar can form a sizeable virtual array, providing higher detection accuracy for target detection. In the monocular vision module, a 1080p resolution Hewlett-Packard (HP) monocular camera was used for follow-up work. It transmits data to the computer via the universal serial bus(USB). The center of the camera establishes the camera coordinate system and the mmWave radar coordinate system is established by the center point of the mmWave radar. At the road intersection of the school, the hardware system was used to collect data for offline verification to evaluate the feasibility of the vehicle detection method based on the fusion of mmWave radar and monocular vision.

Figure 8. Hardware system diagram of mmWave radar and monocular camera.

4.1. Validation of mmWave Radar Algorithms

The built hardware system platform was used to detect vehicle targets at school intersections and to verify the relevant algorithms of the mmWave radar proposed in Section 3.3. The actual distance and speed of vehicle targets are shown in Figure 9. The distance and speed in the picture were manually measured and calibrated.

Figure 9. The actual distance of the target.

Figure 10a shows the 3D detection points map of the detection scene in Figure 9, obtained by processing the raw mmWave radar data by range-FFT, doppler-FFT, CFAR, and DOA estimation algorithms. It can be found that the 3D detection points include both the target 3D detection points and the 3D detection points of many other interfering targets. According to the algorithm proposed in Section 3.3, the 3D detection points of the detection scene were processed and the results are shown in Figure 10. By comparing Figure 10a,b, it can be found that the data pre-processing algorithms proposed in Section 3.3.1 can effectively filter most of the 3D detection points with interfering targets, and get the 3D detection points with small amount of data and containing target information. In addition, as shown in Figure 10c,d, the improved algorithm based on DBSCAN clustering fusion proposed in Section 3.3.2 can effectively cluster multiple 3D detection points corresponding to the same target, and then Euclidean distance averaging yields 3D detection points that can represent the actual target location and eliminate some non-target clusters. Compared with the actual measured value of the target in Figure 9, the distance error between the 3D detection points of the stationary target on the left and the measured distance values in the x-direction is about 0.11 m and the distance error in the y-direction is about 0.78 m. The distance error between the 3D detection points, representing the stationary target on the right and the measured distance values in the x-direction is about 0.3 m, and the distance error in the y-direction is about 0.53 m. The velocity of both targets, is roughly 0 m/s, which matches the target situation. On the other hand, it can be found that, after the DBSCAN clustering algorithm and the Euclidean distance averaging process, an interfering target with a distance of -7 m in the x-direction and 16 m in the y-direction was added. This interfering target can be eliminated by the target fusion algorithm proposed in Section 3.5.2. After calculation, it shows that the distance error caused by the conversion to image pixels does not change much, so the obtained 3D detection points can be effectively converted to the pixel coordinate system on behalf of the target for further processing.

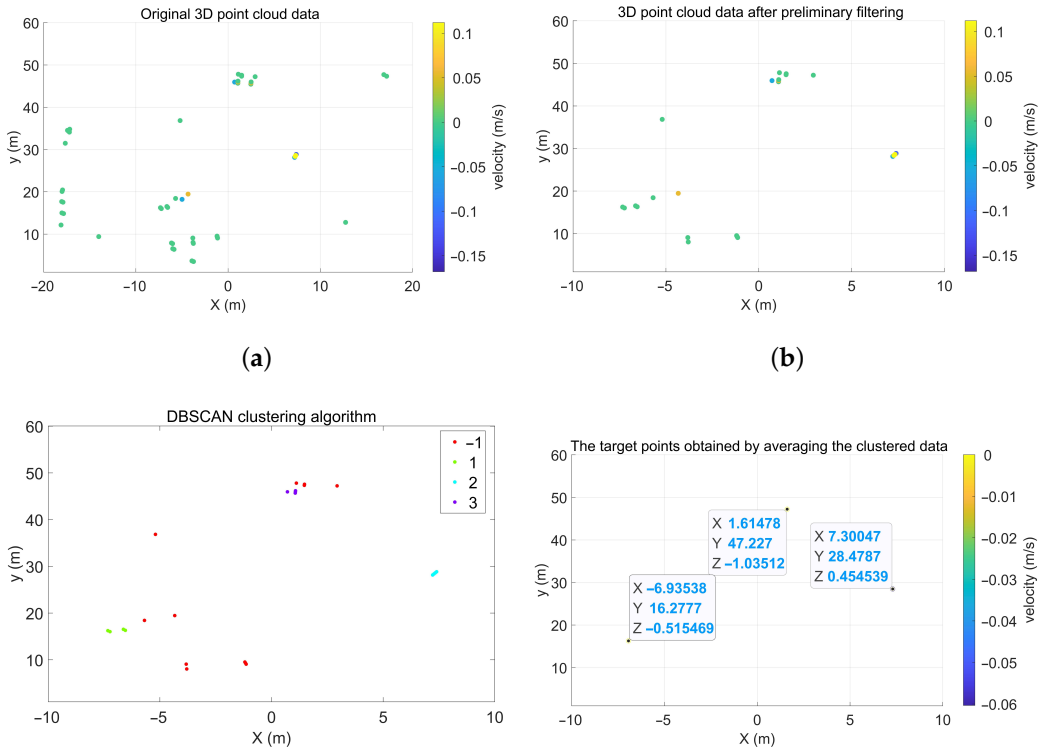

Figure 10. Processing of mmWave radar algorithms. (**a**) Original 3D detection point. (**b**) 3D detection points after data pre-processing. (**c**) DBSCAN clustering algorithm processing results. (**d**) The target points obtained by averaging the clustered data.

4.2. Validation of YOLOv4 Algorithm and Monocular Camera Ranging Algorithm

The collected monocular camera data were processed using the trained YOLOv4 network. The results are shown in Figure 11. The trained YOLOv4 network can detect the target well without occlusion and at short distances.

Figure 11. Detection of target by YOLOv4 algorithm.

The monocular camera was fixed on a 1.5 m high tripod and the silver target car was parked at different distances to collect photos. The closest distance was 10 m and the

farthest was 50 m. The road image data of the target car were collected at nine different distances to validate the camera ranging algorithm, and some of the image data are shown in the following Figure 12. These images were fed into the YOLOv4 target detection model to derive the target frame data of the silver car in the same lane.

Figure 12. Partial silver car target images at different distances.

As shown in Figure 13, the longitudinal distance of this silver target vehicle can be estimated using the pixel value of *ymax* in the target box, which was input into this monocular camera ranging algorithm, i.e., Equation (13). The longitudinal distance of the target at all distances and the statistics of its error were estimated and the results are shown in Table 1.

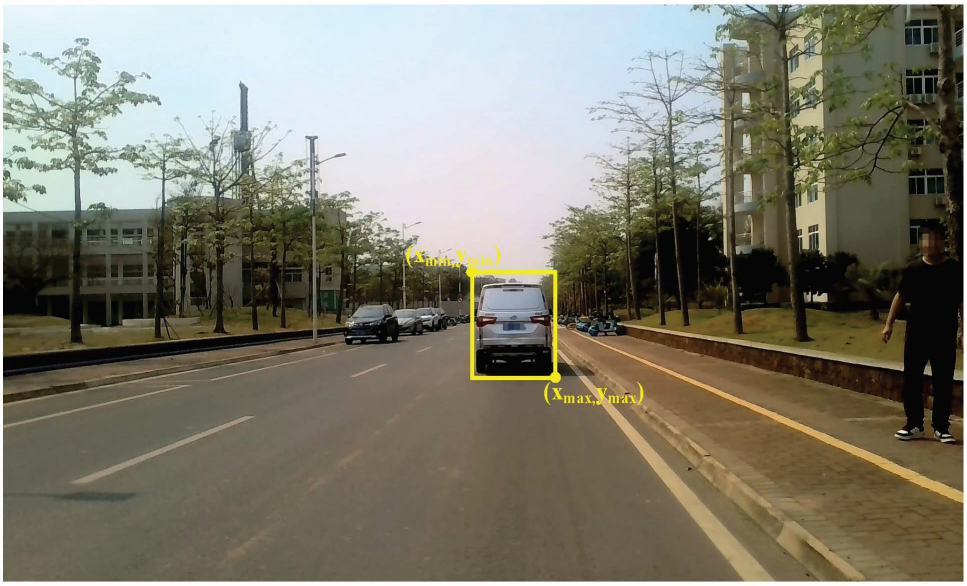

Figure 13. Monocular camera ranging algorithm application schematic.

Table 1. Error of all estimated distances for the monocular camera ranging algorithm.

Actual Distance (m)	Estimated Distance (m)	Error
38	44.38	16.78%
42	45.81	9.07%
46	45.81	0.43%
50	47.24	5.52%
54	51.65	4.35%
58	52.40	9.65%
62	52.83	14.79%

The accuracy of the monocular camera ranging algorithm is within 10% when the target distance is between 42 m and 58 m. However, when the target distance is within 42 m or more than 58 m, the error of the monocular camera ranging algorithm is relatively large—more than 10%.

4.3. Validation of Data Fusion

The fusion algorithm proposed in this paper was used to process and fuse the target data obtained from the hardware system. The fusion result in Figure 14 was compared with the actual manual measurement result of the target in Figure 9, where the red dot represents the position of the target point in the image pixel coordinate system converted by the coordinate system from the three-dimensional detection point of the target's mmWave radar. The results show that the proposed fusion algorithm of mmWave radar and the monocular camera can effectively integrate the distance and velocity information detected by mmWave radar with the distance and category information detected by the monocular camera. Furthermore, the target positioning accuracy has also been significantly improved.

Figure 14. Fusion result of target.

On the other hand, we acquired 467 images and mmWave radar data frames in different scenarios during the same sampling period. There were 1674 targets in the 436 images. The 1680 targets were processed using fusion detection and vision-only detection, respectively. The results of some of these tests are shown in Figure 15. Figure 15a,c,e represent targets that are detected by the vision only in different scenes, while Figure 15b,d,f represent targets that can be detected by fusion in different scenes. It shows that vision-only detection sometimes misses road vehicle targets due to the effects of strong light, foggy days, and occlusion. It greatly affects the accuracy of vision-only detection. Fusion detection

can effectively reduce the number of undetected obstacles. As shown in Table 2, in three different detection scenarios, the fusion detection method has a target detection accuracy of more than 78%, and it can improve the detection accuracy by more than 12% compared to the pure vision detection method. In scene A, fusion detection is particularly effective and can be as high as 90.32% detection accuracy.

(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 15. Detection tests in road scenes. (**a**) Scene A: Detect with vision only. (**b**) Scene A: Fusion detection of monocular vision and mmWave radar. (**c**) Scene B: Detect with vision only. (**d**) Scene B: Fusion detection of monocular vision and mmWave radar. (**e**) Scene C:Detect with vision only. (**f**) Scene C: Fusion detection of monocular vision and mmWave radar.

Table 2. The scenes using fusion detection and vision-only detection to detect targets.

	Scene A	Scene B	Scene C
Number of data frames	186	114	167
Testing environment	night, fog	night, strong light	day, covered
Total targets	744	228	708
Targets by vision-only detection	504	114	483
Accuracy of vision-only detection	67.74%	50.00%	68.22%
Targets by fusion detection	672	178	568
Accuracy by fusion detection	90.32%	78.07%	80.23%

5. Conclusions

For the shortage of single sensor detection target robustness and accuracy, a vehicle detection method based on the fusion of mmWave radar and monocular vision was proposed in this paper. The method combines the benefits of mmWave radar for measuring distance and speed with the vision for classifying objects. It uses the proposed data pre-processing algorithms to process the mmWave radar 3D detection points of the detection scene, effectively eliminating the influence of interference targets such as empty targets and invalid targets. The DBSCAN clustering fusion algorithm and the nearest neighbor algorithm were also used to correlate the same frame data and adjacent frame data, reducing the temporary loss of targets caused by bumps and improving the success rate of target matching. Then, the mmWave radar effective target was matched with the vision effective target obtained by processing with the YOLOv4 algorithm and monocular camera ranging algorithm under temporal-spatio alignment. The successfully matched targets were output using the Kalman weighted fusion algorithm. Targets that were not successfully matched were marked as new targets for tracking and were handled in a valid cycle. Finally, real road data were collected for validation. The experiments showed that the proposed method can not only effectively fuse the data from the two sensors to obtain more comprehensive and accurate information about the target location, speed, and category but also make up for the shortcomings of a single sensor.

Author Contributions: Research design, G.C. and X.W.; data acquisition, J.S., X.L. and Y.G.; writing—original draft preparation, G.C.; writing—review and editing, X.W., X.L. and T.S.; supervision, X.W. and J.S.; funding acquisition, X.W. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Natural Science Foundation of Hainan Province (620RC555), the National Natural Science Foundation of China (Nos. 61961013, 62101088).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fue, K.; Porter, W.; Barnes, E.; Li, C.; Rains, G. Autonomous Navigation of a Center-Articulated and Hydrostatic Transmission Rover using a Modified Pure Pursuit Algorithm in a Cotton Field. *Sensors* **2020**, *20*, 4412. [CrossRef] [PubMed]
2. Dan, P.; Stoican, F.; Stamatescu, G.; Ichim, L.; Dragana, C. Advanced UAV–WSN System for Intelligent Monitoring in Precision Agriculture. *Sensors* **2020**, *20*, 817.

3. Ji, Y.; Peng, C.; Li, S.; Chen, B.; Miao, Y.; Zhang, M.; Li, H. Multiple object tracking in farmland based on fusion point cloud data. *Comput. Electron. Agric.* **2022**, *200*, 107259. [CrossRef]
4. Yi, C.; Zhang, K.; Peng, N. A multi-sensor fusion and object tracking algorithm for self-driving vehicles. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2019**, *233*, 2293–2300. [CrossRef]
5. Cho, M.G. A study on the obstacle recognition for autonomous driving RC car using lidar and thermal infrared camera. In Proceedings of the 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Zagreb, Croatia, 2–5 July 2019; pp. 544–546.
6. Zhang, R.; Cao, S. Real-time human motion behavior detection via CNN using mmWave radar. *IEEE Sens. Lett.* **2018**, *3*, 3500104. [CrossRef]
7. Yoneda, K.; Hashimoto, N.; Yanase, R.; Aldibaja, M.; Suganuma, N. Vehicle localization using 76 GHz omnidirectional millimeter-wave radar for winter automated driving. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Suzhou, China, 26–30 June 2018; pp. 971–977.
8. Nabati, R.; Qi, H.C. Center-based Radar and Camera Fusion for 3D Object Detection. *arXiv* **2020**, arXiv:2011.04841.
9. Premachandra, C.; Murakami, M.; Gohara, R.; Ninomiya, T.; Kato, K. Improving landmark detection accuracy for self-localization through baseboard recognition. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 1815–1826. [CrossRef]
10. Cavanini, L.; Benetazzo, F.; Freddi, A.; Longhi, S.; Monteriu, A. SLAM-based autonomous wheelchair navigation system for AAL scenarios. In Proceedings of the 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA), Senigallia, Italy, 10–12 September 2014; pp. 1–5.
11. Ji, Y.; Li, S.; Peng, C.; Xu, H.; Cao, R.; Zhang, M. Obstacle detection and recognition in farmland based on fusion point cloud data. *Comput. Electron. Agric.* **2021**, *189*, 106409. [CrossRef]
12. Chen, X.; Zhang, B.; Luo, L. Multi-feature fusion tree trunk detection and orchard mobile robot localization using camera/ultrasonic sensors. *Comput. Electron. Agric.* **2018**, *147*, 91–108. [CrossRef]
13. Maldaner, L.F.; Molin, J.P.; Canata, T.F.; Martello, M. A system for plant detection using sensor fusion approach based on machine learning model. *Comput. Electron. Agric.* **2021**, *189*, 106382. [CrossRef]
14. Xue, J.; Fan, B.; Yan, J.; Dong, S.; Ding, Q. Trunk detection based on laser radar and vision data fusion. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 20–26. [CrossRef]
15. Fayyad, J.; Jaradat, M.A.; Gruyer, D.; Najjaran, H. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors* **2020**, *20*, 4220. [CrossRef]
16. Steinbaeck, J.; Steger, C.; Brenner, E.; Holweg, G.; Druml, N. Occupancy grid fusion of low-level radar and time-of-flight sensor data. In Proceedings of the 2019 22nd Euromicro Conference on Digital System Design (DSD), Kallithea, Greece, 28–30 August 2019; pp. 200–205.
17. Will, C.; Vaishnav, P.; Chakraborty, A.; Santra, A. Human target detection, tracking, and classification using 24-GHz FMCW radar. *IEEE Sens. J.* **2019**, *19*, 7283–7299. [CrossRef]
18. Chen, B.; Pei, X.; Chen, Z. Research on target detection based on distributed track fusion for intelligent vehicles. *Sensors* **2019**, *20*, 56. [CrossRef]
19. Kim, D.; Kim, S. Extrinsic parameter calibration of 2D radar-camera using point matching and generative optimization. In Proceedings of the 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 15–18 October 2019; pp. 99–103.
20. Fang, Y.; Masaki, I.; Horn, B. Depth-based target segmentation for intelligent vehicles: Fusion of radar and binocular stereo. *IEEE Trans. Intell. Transp. Syst.* **2002**, *3*, 196–202. [CrossRef]
21. Yeong, D.J.; Velasco-Hernandez, G.; Barry, J.; Walsh, J. Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors* **2021**, *21*, 2140. [CrossRef] [PubMed]
22. Zhang, X.; Zhou, M.; Qiu, P.; Huang, Y.; Li, J. Radar and vision fusion for the real-time obstacle detection and identification. *Ind. Robot. Int. J. Robot. Res. Appl.* **2019**, *46*, 391–395. [CrossRef]
23. Cong, J.; Wang, X.; Huang, M.; Wan, L. Robust DOA Estimation Method for MIMO Radar via Deep Neural Networks. *IEEE Sens. J.* **2020**, *21*, 7498–7507. [CrossRef]
24. Cong, J.; Wang, X.; Yan, C.; Yang, L.T.; Dong, M.; Ota, K. CRB Weighted Source Localization Method Based on Deep Neural Networks in Multi-UAV Network. *IEEE Internet Things J.* **2023**, *10*, 5747–5759. [CrossRef]
25. Jiang, W.; Ren, Y.; Liu, Y.; Leng, J. Artificial Neural Networks and Deep Learning Techniques Applied to Radar Target Detection: A Review. *Electronics* **2022**, *11*, 156. [CrossRef]
26. Lv, P.; Wang, B.; Cheng, F.; Xue, J. Multi-Objective Association Detection of Farmland Obstacles Based on Information Fusion of Millimeter Wave Radar and Camera. *Sensors* **2023**, *23*, 230. [CrossRef]
27. Liu, Y.; Zhang, L.; Li, P.; Jia, T.; Du, J.; Liu, Y.; Li, R.; Yang, S.; Tong, J.; Yu, H. Laser Radar Data Registration Algorithm Based on DBSCAN Clustering. *Electronics* **2023**, *12*, 1373. [CrossRef]
28. Pearce, A.; Zhang, J.A.; Xu, R.; Wu, K. Multi-Object Tracking with mmWave Radar: A Review. *Electronics* **2023**, *12*, 308. [CrossRef]
29. Hsu, Y.W.; Lai, Y.H.; Zhong, K.Q.; Yin, T.K.; Perng, J.W. Developing an on-road object detection system using monovision and radar fusion. *Energies* **2019**, *13*, 116. [CrossRef]
30. Jin, F.; Sengupta, A.; Cao, S.; Wu, Y.J. Mmwave radar point cloud segmentation using gmm in multimodal traffic monitoring. In Proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 28–30 April 2020; pp. 732–737.

31. Zhou, T.; Jiang, K.; Xiao, Z.; Yu, C.; Yang, D. Object detection using multi-sensor fusion based on deep learning. In Proceedings of the CICTP 2019, Nanjing, China, 6–8 July 2019; pp. 5770–5782.
32. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
33. Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; Heide, F. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11682–11692.
34. Cao, C.; Gao, J.; Liu, Y.C. Research on space fusion method of millimeter wave radar and vision sensor. *Procedia Comput. Sci.* **2020**, *166*, 68–72. [CrossRef]
35. Neal, R.M. Bayesian methods for machine learning. *NIPS Tutor.* **2004**, *13*.
36. Zhang, Z. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
37. Bhatia, J.; Dayal, A.; Jha, A.; Vishvakarma, S.K.; Cenkeramaddi, L.R. Object Classification Technique for mmWave FMCW Radars using Range-FFT Features. In Proceedings of the 2021 International Conference on COMmunication Systems and NETworkS (COMSNETS), Bangalore, India, 5–9 January 2021.
38. Zhang, X.; Xu, L.; Xu, L.; Xu, D. Direction of Departure (DOD) and Direction of Arrival (DOA) Estimation in MIMO Radar with Reduced-Dimension MUSIC. *IEEE Commun. Lett.* **2010**, *14*, 1161–1163. [CrossRef]
39. Yun, D.J.; Jung, H.; Kang, H.; Yang, W.Y.; Seo, D.W. Acceleration of the Multi-Level Fast Multipole Algorithm Using K-Means Clustering. *Electronics* **2020**, *9*, 1926. [CrossRef]
40. Wu, X.; Ren, J.; Wu, Y.; Shao, J. *Study on Target Tracking Based on Vision and Radar Sensor Fusion*; Technical Report, SAE Technical Paper; SAE: Warrendale, PA, USA, 2018.
41. Gong, P.; Wang, C.; Zhang, L. Mmpoint-gnn: Graph neural network with dynamic edges for human activity recognition through a millimeter-wave radar. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–7.
42. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
43. Cong, J.; Wang, X.; Lan, X.; Huang, M.; Wan, L. Fast Target Localization Method for FMCW MIMO Radar via VDSR Neural Network. *Remote Sens.* **2021**, *13*, 1956. [CrossRef]