entropy

MDPI

Article

Real and Pseudo Pedestrian Detection Method with CA-YOLOv5s Based on Stereo Image Fusion

Xiaowei Song [1,2,*], Gaoyang Li [1], Lei Yang [1,*], Luxiao Zhu [1], Chunping Hou [3] and Zixiang Xiong [4]

[1] School of Electronic and Information, Zhongyuan University of Technology, Zhengzhou 450007, China
[2] Dongjing Avenue Campus, Kaifeng University, Kaifeng 475004, China
[3] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[4] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
* Correspondence: sxw@zut.edu.cn (X.S.); yanglei@zut.edu.cn (L.Y.)

Abstract: With the development of convolutional neural networks, the effect of pedestrian detection has been greatly improved by deep learning models. However, the presence of pseudo pedestrians will lead to accuracy reduction in pedestrian detection. To solve the problem that the existing pedestrian detection algorithms cannot distinguish pseudo pedestrians from real pedestrians, a real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion is proposed in this paper. Firstly, the two-view images of the pedestrian are captured by a binocular stereo camera. Then, a proposed CA-YOLOv5s pedestrian detection algorithm is used for the left-view and right-view images, respectively, to detect the respective pedestrian regions. Afterwards, the detected left-view and right-view pedestrian regions are matched to obtain the feature point set, and the 3D spatial coordinates of the feature point set are calculated with Zhengyou Zhang's calibration method. Finally, the RANSAC plane-fitting algorithm is adopted to extract the 3D features of the feature point set, and the real and pseudo pedestrian detection is achieved by the trained SVM. The proposed real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion effectively solves the pseudo pedestrian detection problem and efficiently improves the accuracy. Experimental results also show that for the dataset with real and pseudo pedestrians, the proposed method significantly outperforms other existing pedestrian detection algorithms in terms of accuracy and precision.

Keywords: stereo image fusion; pedestrian detection; CA; YOLOv5s; pseudo pedestrian

Citation: Song, X.; Li, G.; Yang, L.; Zhu, L.; Hou, C.; Xiong, Z. Real and Pseudo Pedestrian Detection Method with CA-YOLOv5s Based on Stereo Image Fusion. *Entropy* **2022**, *24*, 1091. https://doi.org/10.3390/e24081091

Academic Editors: Jiayi Ma, Yu Liu, Junjun Jiang, Zheng Wang and Han Xu

Received: 6 July 2022
Accepted: 4 August 2022
Published: 8 August 2022

1. Introduction

Pedestrian detection is an important branch of object detection, having received wide attention in the past two decades [1]. The purpose of pedestrian detection is to find all possible pedestrians in the input image and output the location of pedestrian in the image. Pedestrian detection can be widely used in areas such as safety monitoring and automatic driving, where the accuracy of pedestrian detection is crucial [2].

Pedestrian detection technology has developed from traditional hand-assisted feature detection [3–5] to modern deep learning-based feature detection [6–9]. Traditional pedestrian detection algorithms require the manual design of filters and features, such as Gabor filter, gradient-based feature, channel feature, etc., according to statistical or prior knowledge of the designer. Cheng et al. proposed a pedestrian detection method using a sparse Gabor filter which is designed according to the learned texture features from some manually selected typical images of pedestrian [10]. Dalal proposed a pedestrian detection method using edge features extracted by a histogram of oriented gradient (HOG), which is obtained by the calculation and statistics of HOG in some manually selected local image areas [5]. Dollar et al. proposed a pedestrian detection method using channel features extracted by the integral of some manually selected registered image channels [11]. These

traditional pedestrian detection algorithms are time consuming and laborious due to the manual intervention, with relatively low detection accuracy and efficiency.

With the development of convolutional neural networks, the effect of pedestrian detection has been pushed to an unprecedentedly high level by the modern deep learning-based pedestrian detection algorithms [12,13]. Modern pedestrian detection algorithms based on deep learning can autonomously learn and extract features of pedestrian, with high detection accuracy and efficiency. Many challenging problems have been well solved [14]. For instance, Zhang et al. solved the problem of small-scale pedestrian detection with asymmetric multi-stage CNNs [15]. Xu et al. solved the efficiency problem of pedestrian detection through the model reconstruction and pruning of YOLOv3 network [16]. Lin et al. solved the robustness problem of obscured pedestrian detection with multi-grained deep feature learning [17]. Li et al. solved the effectiveness problem of pedestrian detection in hazy weather with a weighted combination layer, which combines multi-scale feature maps with a squeeze and excitation block [18]. However, the elimination problem of false positive samples in pedestrian detection has not been solved yet.

The false positive samples include trash cans, traffic lights, trees and people printed on flat surfaces. Since these false positive samples have similar characteristics to pedestrians, they are always incorrectly detected as pedestrian by most pedestrian detection algorithms [19]. The incorrect detection of false positive samples, such as trash cans, traffic lights and trees has been solved through network improvement [20–22]. However, the incorrect detection of people printed on flat surfaces has not been well solved because printed people have almost exactly the same characteristics as pedestrians. There are mainly two types of pedestrians printed on flat surfaces: pseudo pedestrian in a 2D plane with background (PPWB) and pseudo pedestrian in a 2D plane with no background (PPWNB), which are collectively called pseudo pedestrians in this paper.

There is almost no difference between real and pseudo pedestrians in 2D features, so it is necessary to take advantage of 3D features to distinguish them. There have been some attempts to detect pedestrians with 3D information. Shakeri et al. collected 3D information contained in the left-view and right-view images by a binocular stereo camera, enhanced the image quality of the pedestrian area of interest by 3D information fusion, and thus improved the accuracy of pedestrian detection [23]. However, only 2D information is used in pedestrian detection, which cannot realize real and pseudo pedestrian detection. Wei et al. also captured 3D information included in the left-view and right-view images by a binocular stereo camera, took advantage of the complementary information of the left-view and right-view images, and solved the problem of obscured pedestrian detection [24]. Nevertheless, similar to Ref. [23], only 2D information is used in pedestrian detection, which cannot complete real and pseudo pedestrian detection as well. Zhao et al. acquired 3D information contained in the 2D image and depth map by a light field camera, and performed pedestrian detection according to the 3D information, including 2D information and depth information [25]. PPWB at the same depth as the background can be distinguished from the real pedestrian, while PPWNB not at the same depth as the background can still not be distinguished from the real pedestrian. Therefore, it is necessary to further solve the problem of pedestrian detection involving both PPWB and PPWNB.

In this paper, a real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion is proposed. The proposed method is designed according to the constructed real and pseudo pedestrian detection bionic model based on human stereo vision. A binocular stereo camera is adopted to capture the left-view and right-view images of the pedestrian. The two-view images are respectively detected by the improved CA-YOLOv5s pedestrian detection algorithm to obtain the respective pedestrian regions. The detected pedestrian regions are stereo matched to obtain a feature point set, and the 3D spatial coordinates of the feature point set are calculated with Zhengyou Zhang's calibration method. The mismatched feature points are eliminated, and a matched feature point set is reserved. The 3D features of the matched feature point set are extracted by random sample consensus (RANSAC) plane fitting, and the real and pseudo pedestrian detection

is completed by the trained support vector machine (SVM) model. The proposed method can effectively solve the problem of pseudo pedestrian detection, and increase the accuracy as well.

The rest of the paper is organized as follow. In Section 2, we review some related works on the principle of human stereo vision and attention mechanism. In Section 3, we construct a real and pseudo pedestrian detection bionic model based on human stereo vision and propose a real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion. In Section 4, we report the experimental results. In Section 5, we make a conclusion.

2. Related Works

2.1. Principle of Human Stereo Vision

Human stereo vision can perfectly realize the real and pseudo pedestrian detection, so it is the biological theoretical basis of the proposed method in this paper. In human stereo vision system, as shown in Figure 1, the 3D pedestrian is imaged on the retina through human optical components such as lens, and the photoreceptor cells on the retina convert optical signals into bioelectrical signals which are transmitted to the optic chiasma through the optic nerve. The optic chiasma rearranges the signals and transmits them to the lateral geniculate nucleus (LGN), and the processed signals are sent to the visual center of the occipital lobe through optic radiation. In the visual center of the occipital lobe, the region of interest is extracted by the receptive field division, the binocular single vision is formed through fusion, the stereo vision is achieved through spatial position perception, and the real and pseudo pedestrian judgment is made accordingly.

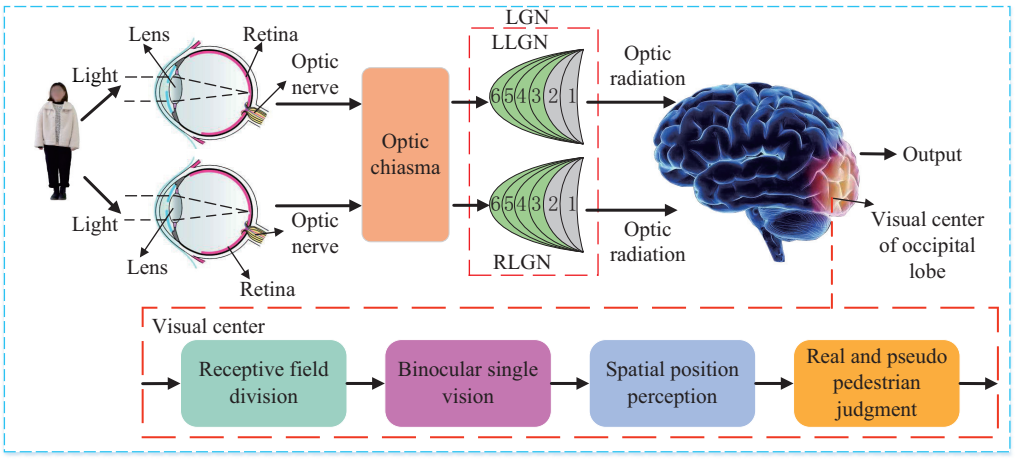


Figure 1. Pedestrian detection principle diagram of human stereo vision.

When viewing an object, the optic chiasma rearranges the signals from the right visual field of the left eye and the right visual field of the right eye and transmits them to the left LGN (LLGN), and rearranges the signals from the left visual field of the left eye and the left visual field of the right eye and transmits them to the right LGN (RLGN) [26]. For LLGN, the light intensity I_L^r of the optical signal perceived at the right visual field of the left retina (x_L^r, y_L^r) from the right visual field of the left eye at time t can be expressed by Equation (1), while the light intensity I_R^r of the optical signal perceived at the right visual field of the right retina (x_R^r, y_R^r) from the right visual field of the right eye at time t can be expressed by Equation (2). For RLGN, the light intensity I_L^l of the optical signal perceived at the left visual field of the left retina $\left(x_L^l, y_L^l\right)$ from the left visual field of the left eye at time t can be expressed by Equation (3), while the light intensity I_R^l of the optical signal perceived at the

left visual field of the right retina $\left(x_R^l, y_R^l\right)$ from the left visual field of the right eye at time t can be expressed by Equation (4).

$$I_L^r(x_L^r, y_L^r, t) = k_L \int_{\lambda_l}^{\lambda_h} P_L(x_L^r, y_L^r, \lambda, t) V_L(\lambda) d\lambda \tag{1}$$

$$I_R^r(x_R^r, y_R^r, t) = k_R \int_{\lambda_l}^{\lambda_h} P_R(x_R^r, y_R^r, \lambda, t) V_R(\lambda) d\lambda \tag{2}$$

$$I_L^l\left(x_L^l, y_L^l, t\right) = k_L \int_{\lambda_l}^{\lambda_h} P_L\left(x_L^l, y_L^l, \lambda, t\right) V_L(\lambda) d\lambda \tag{3}$$

$$I_R^l\left(x_R^l, y_R^l, t\right) = k_R \int_{\lambda_l}^{\lambda_h} P_R\left(x_R^l, y_R^l, \lambda, t\right) V_R(\lambda) d\lambda \tag{4}$$

Wherein $\left(x_L^r, y_L^r\right)$ and $\left(x_R^r, y_R^r\right)$ are the coordinates of the corresponding imaging points in the right visual field of the left and right retina, respectively; $\left(x_L^l, y_L^l\right)$ and $\left(x_R^l, y_R^l\right)$ are the coordinates of the corresponding imaging points in the left visual field of the left and right retina respectively; k_l and k_r are the adjustable coefficients of the left and right eye respectively; $P_L\left(x_L^r, y_{L,}^r \lambda, t\right)$ and $P_R\left(x_R^r, y_R^r, \lambda, t\right)$ are the radiation power of light with wavelength λ received at $\left(x_L^r, y_L^r\right)$ and $\left(x_R^r, y_R^r\right)$ respectively; $P_L\left(x_L^l, y_L^l, \lambda, t\right)$ and $P_R\left(x_R^l, y_R^l, \lambda, t\right)$ are the radiation power of light with wavelength λ received at $\left(x_L^l, y_L^l\right)$ and $\left(x_R^l, y_R^l\right)$, respectively; $V_L(\lambda)$ and $V_R(\lambda)$ are the spectral response functions of the left and right eye, respectively; λ_h and λ_l are the upper and lower wavelength limits of human eye perception.

The optical signal causes ion exchange in the Na^+-K^+ ion pumps in the photoreceptor cells of the retina, resulting in a change in the electric potential, which is voltage [27]. Thus, the optical signals I_L^r and I_R^r at the right visual field of the left and right retina are converted into the bioelectrical signals U_L^r and U_R^r in the right visual field by photoelectric conversion (PEC), as expressed by Equations (5) and (6). The optical signals I_L^l and I_R^l at the right left field of the left and right retina are converted into the bioelectrical signals U_L^l and U_R^l in the left visual field by PEC, as expressed by Equations (7) and (8).

$$U_L^r(x_L^r, y_L^r, t) = PEC(I_L^r(x_L^r, y_L^r, t)) \tag{5}$$

$$U_R^r(x_R^r, y_R^r, t) = PEC(I_R^r(x_R^r, y_R^r, t)) \tag{6}$$

$$U_L^l\left(x_L^l, y_L^l, t\right) = PEC\left(I_L^l\left(x_L^l, y_L^l, t\right)\right) \tag{7}$$

$$U_R^l\left(x_R^l, y_R^l, t\right) = PEC\left(I_R^l\left(x_R^l, y_R^l, t\right)\right) \tag{8}$$

The bioelectrical signals U_L^r and U_R^r in the right visual field are transmitted to the optic chiasma (OC) through the optic nerve, where they are rearranged and sent to the LLGN. The bioelectrical signals received by the LLGN can be expressed by Equation (9). The bioelectrical signals U_L^l and U_R^l in the left visual field are transmitted to the optic chiasma through the optic nerve, where they are rearranged and sent to the RLGN. The bioelectrical signals received by the RLGN can be expressed by Equation (10).

$$U_{\mathrm{LLGN}}(x^r, y^r, t) = OC(U_L^r(x_L^r, y_L^r, t) \cup U_R^r(x_R^r, y_R^r, t)) \tag{9}$$

$$U_{\mathrm{RLGN}}\left(x^l, y^l, t\right) = OC\left(U_L^l\left(x_L^l, y_L^l, t\right) \cup U_R^l\left(x_R^l, y_R^l, t\right)\right) \tag{10}$$

The bioelectrical signals U_{LLGN} in the LLGN are sent to the left brain through optic radiation (OR). The bioelectrical signals U_{LB} received by the left brain can be expressed by Equation (11), which represents the right visual field. The bioelectrical signals U_{RLGN}

in the RLGN are sent to the right brain through optic radiation. The bioelectrical signals U_{RB} received by the right brain can be expressed by Equation (12), which represent the left visual field.

$$U_{LB}(x^r, y^r, t) = OR(U_{\mathrm{LLGN}}(x^r, y^r, t)) \tag{11}$$

$$U_{RB}\left(x^l, y^l, t\right) = OR\left(U_{\mathrm{RLGN}}\left(x^l, y^l, t\right)\right) \tag{12}$$

In the visual center of the occipital lobe, the bioelectrical signals U_{LB} and U_{RB} are combined into a bioelectrical signal U_B representing the whole visual field, which can be expressed by Equation (13).

$$U_B(x, y, t) = U_{RB}\left(x^l, y^l, t\right) \cup U_{LB}(x^r, y^r, t) \tag{13}$$

Visual cortex cells only respond significantly to the bioelectrical signal U_{B_RF} in their receptive field (RF), as expressed by Equation (14).

$$U_{B\_RF}(x_{RF}, y_{RF}, t) = RF(U_B(x, y, t)) \tag{14}$$

The bioelectrical signal U_{B_RF} has a hierarchical structure, in which different layers correspond to the different bioelectrical signals from the left and right eyes, respectively. The visual cortex of the brain fuses the layered bioelectrical signals U_{B_RF} in the receptive field to form a single object image, that is, binocular single vision, then the spatial position perception is realized, as expressed by Equation (15).

$$I_P(X, Y, Z, t) = F\left(U_{B\_RF}(x_{RF}, y_{RF}, t)\right) \tag{15}$$

Finally, the real and pseudo pedestrian judgment is made by the brain according to the perceived stereo vision information I_P and the judgment result is output, as expressed by Equation (16).

$$\text{Output } = J(I_P(X, Y, Z, t)) \tag{16}$$

With the above process, the real and pseudo pedestrian judgment is completed by the human stereo vision system.

2.2. Attention Mechanism

In the pedestrian detection network, more weight can be allocated to the pedestrian area and less weight to the background area through the focusing effect of the attention mechanism, so as to improve the accuracy of pedestrian detection and reduce the network model parameters.

According to its processing mechanism, the attention module can be divided into three types: spatial attention module, channel attention module and mixed attention module [28–32]. The spatial attention module carries out average pooling and maximum pooling in the channel direction at the same time using the spatial weight matrix. The spatial attention matrix is obtained by convolution, and a 2D spatial attention map is generated by the activation function, thus the spatial position that needs to be focused on is determined. Moreover, the attention mechanism has also been used in multimodal image fusion [33–35] to enhance the pedestrian detection, and has achieved promising results.

Typical channel attention module includes squeeze-and-excitation (SE) and efficient channel attention (ECA). SE samples the input image by global average pooling, learns the dependence to each channel by the shared multilayer perceptron (MLP), and generates the channel attention map by the activation function [28]. ECA improves the shared MLP part of SE, focusing on the interaction of each channel and its k neighborhood channels, and greatly reduces the network parameters [29]. The mixed attention module combines different kinds of attention. The convolutional block attention module (CBAM) and coordinate attention (CA) are the typical representatives. CBAM connects the channel attention module with the spatial attention module through convolution, and can obtain the spatial attention and

channel attention joint optimized features [30]. CA embeds the location information into the channel attention module, and decomposes the channel attention module into two 1D feature coding processes, aggregating features along two spatial directions. The network can quickly focus on the region of interest, and the performance of the pedestrian detection network can be effectively improved [31].

3. Proposed Method

A real and pseudo pedestrian detection bionic model based on human stereo vision is designed in this paper, as shown in Figure 2. In the bionic model, the human eyes are imitated by a binocular stereo camera, which captures external visual information. The photoreceptor cells on the retina are imitated by the charge coupled device (CCD) in the camera, which converts the optical signal into an electrical signal. The electrical signal is transmitted to the processor through the signal line, and the visual center of the occipital lobe is imitated by the processor. In the processor, the pedestrian region in the image is firstly extracted by the 2D pedestrian detection network, the fusing process of binocular single vision is then simulated by the binocular stereo matching, the spatial position perception is next simulated by the binocular stereo ranging, and the real and pseudo pedestrian judgment is finally simulated by the SVM prediction.

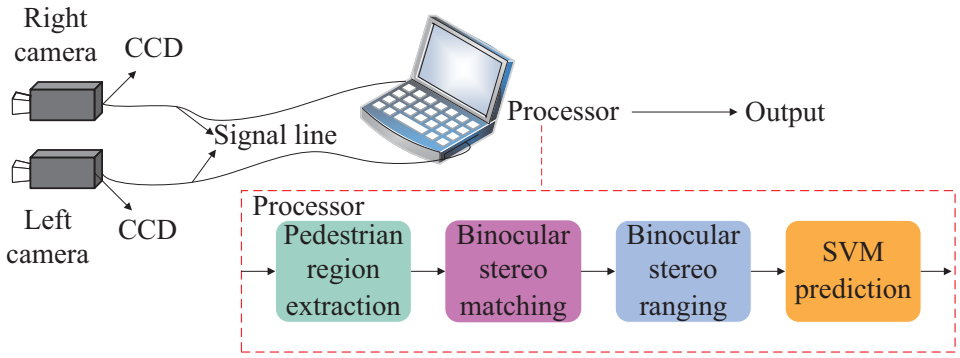


Figure 2. Bionic model diagram for real and pseudo pedestrian detection based on human stereo vision.

To realize the function of the processor in the designed bionic model, a real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion is proposed in this paper. As shown in Figure 3, the proposed method consists of four modules, pedestrian region extraction, binocular stereo matching, binocular stereo ranging and SVM prediction, which correspond to the four processes of the visual center, that is, receptive field division, binocular single vision, spatial position perception and real and pseudo pedestrian judgment. In the pedestrian region extraction module, the dual-view images containing pedestrian are collected by the binocular stereo camera, and the left-view pedestrian regions ROI_L and the right-view pedestrian regions ROI_R are extracted by the improved CA-YOLOv5s pedestrian detection algorithm, respectively. In the binocular stereo matching module, SURF matching [36] is performed on the ROI_L and ROI_R to obtain matched feature point pairs $(\mathbf{p}_{Li}, \mathbf{p}_{Ri}), i = 1, 2, \ldots, N$, and the calibration parameters f_L, f_R, $\mathbf{R}$ and $\mathbf{T}$ of the binocular stereo camera are calculated by Zhengyou Zhang's calibration method [37]. In the binocular stereo ranging module, the feature point set $S = \{\mathbf{P}_i(x_i, y_i, z_i), i = 1, 2, \ldots, N\}$ corresponding to all the matched feature point pairs $(\mathbf{p}_{Li}, \mathbf{p}_{Ri}), i = 1, 2, \ldots, N$ in ROI_L and ROI_R is calculated according to the calibration parameters of the binocular stereo camera. The space distance d_i between each spatial feature point $\mathbf{P}_i$ and the origin of the world coordinate system, namely the optical center of the left-view camera, is calculated, the mean value $\bar{d}$ and standard deviation σ of all d_i are derived, and the absolute difference $|\Delta d_i|$ between each d_i and $\bar{d}$ is computed. The mismatched feature points are eliminated according to the relationship between $|\Delta d_i|$ and σ, and the matched feature point set

$S_{\text{match}} = \{\mathbf{P}_j(x_j, y_j, z_j), j = 1, 2, \ldots, M\}$ is obtained, $M \leq N$. In the SVM prediction module, the mean values in x, y and z directions of all the points in S_{match} are calculated to form a new point $\overline{\mathbf{P}}(\bar{x}_{\text{match}}, \bar{y}_{\text{match}}, \bar{z}_{\text{match}})$, and the space distance $\overline{d}_{\text{match}}$ is calculated to represent the space distance between the pedestrian and the camera. According to the optimal threshold TH_{opt}, fitting is performed on all M points in S_{match} to obtain a fitting plane α_{Fit}. The standard deviation $\sigma_{d_{fit}}$ of the distance d_{fit_j} from each point in S_{match} to the fitting plane α_{Fit} is calculated. The $\overline{d}_{\text{match}}$ and $\sigma_{d_{fit}}$ are input into the pre-trained real and pseudo pedestrian classification model, and real and pseudo pedestrian detection can be achieved. The proposed method solves the problem that the existing pedestrian detection algorithms cannot identify the pseudo pedestrian well, effectively reduces the number of false positive samples, and improves the accuracy of pedestrian detection.

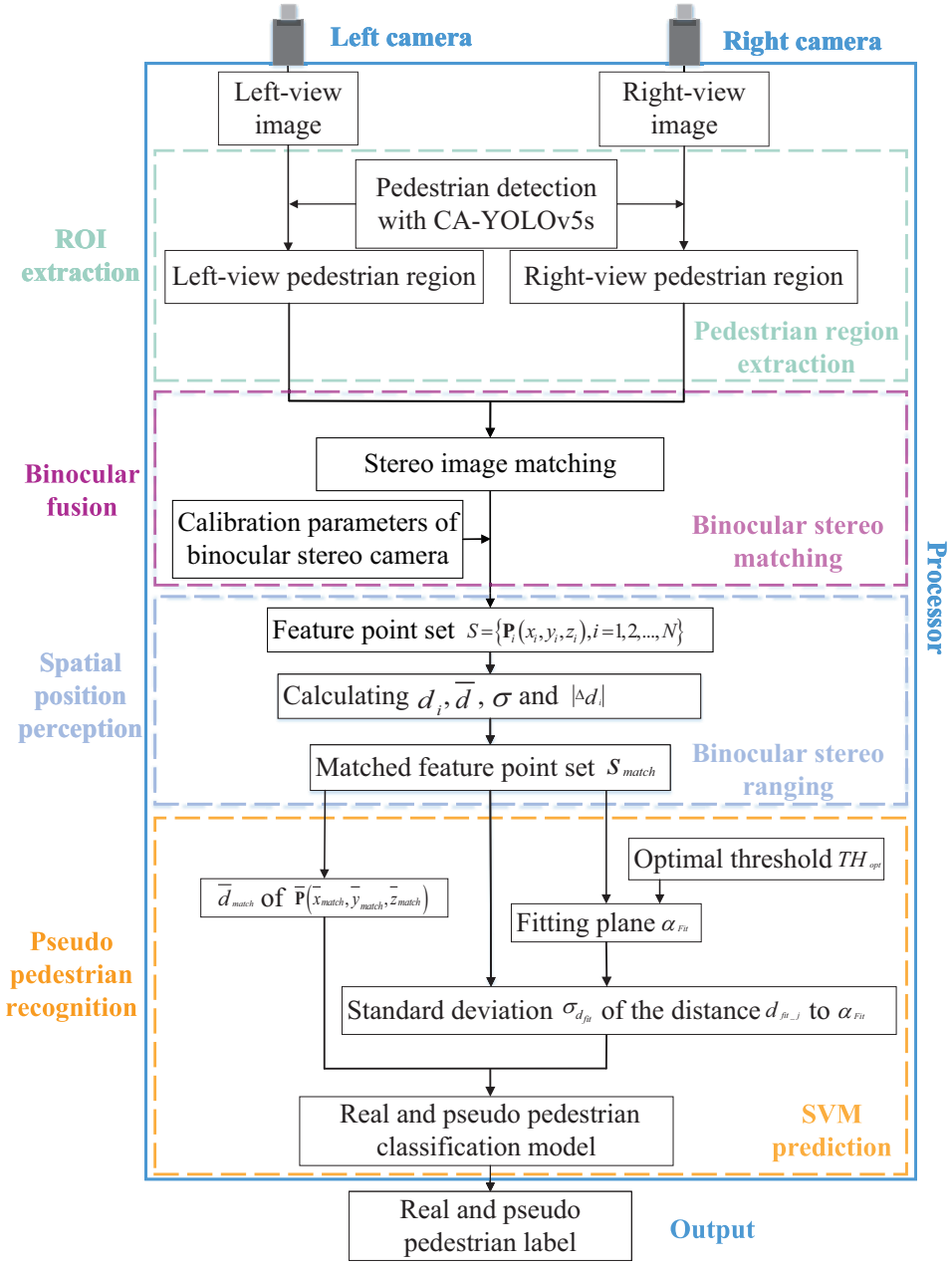


Figure 3. Block diagram of the real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion.

3.1. Pedestrian Region Extraction

Modern deep learning-based pedestrian detection algorithm can be divided into a two-stage pedestrian detection algorithm and single-stage pedestrian detection algorithm [38]. The most representative two-stage pedestrian detection algorithm is R-CNN series [39], including Fast R-CNN [40], Faster R-CNN [7], Cascade R-CNN [41], etc., with high scalability and good detection performance, but complex structure and low speed. The most representative single-stage pedestrian detection algorithm includes YOLO series [42], SSD [8], RFB [43], M2Det [44], RetinaNet [45], etc., with fast detection speed, but relatively low detection performance. However, as technical progresses in YOLO series, single-stage detection algorithms have outperformed two-stage detection algorithms not only in detection speed but also in detection accuracy. Among these single-stage detection algorithms, the YOLOv5 detection algorithm is particularly suitable for pedestrian detection because of its fast detection speed, high detection accuracy, and good deployment on hardware device [46]. There are four common detection algorithms in the YOLOv5 series, i.e., YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. From YOLOv5s to YOLOv5x, the detection accuracy increases steadily, while the detection speed decreases rapidly and the network complexity increases significantly [47].

Eight typical object detection algorithms are selected for pedestrian detection algorithm selection and verification, namely, SSD, RFB, RetinaNet, M2Det, YOLOv3, YOLOv4, YOLOv5s and YOLOv5m. The experimental dataset consists of 17,587 images containing people selected from the public dataset VOC and 3119 pedestrian images with a resolution of 2448 $\times$ 2048 collected in the laboratory, totaling 20,706 images. Image samples of the dataset are shown in Figure 4. During the experiment, the same parameters are used to train the model, and the same model performance indices, that is, average precision (AP) and frame per second (FPS), are selected to evaluate the model. The experimental results are shown in Table 1. The performance indices of YOLOv5s and YOLOv5m are significantly better than the other six algorithms. For YOLOv5s, the AP is 89.35%, the FPS is 73, and the model parameter amount is 26.88 MB, while for YOLOv5m, the AP is 90.36%, the FPS is 60, and the model parameter amount is 80.23 MB. The AP of YOLOv5s is only 1.01% lower than that of YOLOv5m, but the FPS of YOLOv5s is 21.7% higher than that of YOLOv5m and the parameter amount of YOLOv5s is 66.5% lower than that of YOLOv5m. The FPS and parameter amount of YOLOv5s are significantly better than those of YOLOv5m. Therefore, on the premise of ensuring the detection accuracy, YOLOv5s with the fastest detection speed and the smallest model parameter amount is selected as the basic network for improving the pedestrian detection performance in this paper.

The attention mechanism consistent with human perception is beneficial for the pedestrian detection network to focus on pedestrian quickly. In most pedestrian detection scenes, pedestrian objects usually have characteristics of multi-scale variation and spatial position variation due to the movement of pedestrian parallel to and perpendicular to the shooting direction. Hence, both channel attention and spatial attention should be considered. Therefore, the mixed attention mechanism is selected to optimize the YOLOv5s network.

Table 1. Performance comparison of different pedestrian detection algorithms.

Algorithm	AP (%)	FPS	Parameter Amount (MB)
SSD	82.42	31	90.27
RFB	81.57	23	141.67
RetinaNet	74.63	11	138.86
M2Det	81.29	25	226.03
YOLOv3	83.89	14	234.98
YOLOv4	83.06	46	244.30
YOLOv5s	89.35	73	26.88
YOLOv5m	90.36	60	80.23

(a) (b)

Figure 4. Image examples of the dataset. (**a**) Image selected from VOC dataset. (**b**) Collected image.

Pedestrian feature extraction was carried out in the backbone network of YOLOv5s. Glenn Jocher et al. found that the last layer of the backbone network C3 is the best choice for replacement in the process of optimizing YOLOv5s with C3 Transformer (C3TR) [47], i.e., replacing the attention module for the C3 module in the last layer of the backbone network of YOLOv5s. The CBAM mixed attention module and CA mixed attention module are used to replace the C3 module in the last layer of the backbone network of YOLOv5s. Meanwhile, the SE and ECA channel attention modules are used to complete the comparative experiment.

The improved YOLOv5s network is denoted as CBAM-YOLOv5s, CA-YOLOv5s, SE-YOLOv5s and ECA-YOLOV5s, respectively. The model parameter amount and model compression ratio are shown in Table 2. When the input image size is 640 × 640, the model parameter amount of YOLOv5s is 26.88 MB. Compared with these data, the model parameter amount of CBAM-Yolov5s is 22.50 MB, which is compressed by 16.29%. The model parameter amount of CA-YOLOv5s is 22.47 MB, which is compressed by 16.41%. The model parameter amount of SE-YOLOv5s is 27.63 MB, which is increased by 2.79%. The model parameter amount of ECA-YOLOV5s is 22.37 MB, which is compressed by 16.78%. ECA-YOLOV5s is the best, and CA-YOLOv5s is the second. The performance indices AP, recall and FPS of CBAM-YOLOv5s, CA-YOLOv5s, SE-YOLOv5s and ECA-YOLOV5s are shown in Table 3. For YOLOv5s, AP is 89.35%, recall is 82.09% and FPS is 73. Compared with this, CA-YOLOv5s is better than YOLOv5s in the AP index, CBAM-YOLOv5s and CA-YOLOv5s are better than YOLOv5s in the recall index, CBAM-YOLOv5s, CA-YOLOv5s and ECA-YOLOv5s are better than YOLOv5s in the FPS index. Only CA-YOLOv5s is better than YOLOv5s in all three indices.

In conclusion, CA-YOLOv5s is selected as the pedestrian detection algorithm in this paper, and its network structure is shown in Figure 5, in which the C3 module in the last layer of the backbone network is replaced with the CA attention module. The detailed network structure of the CA is shown in Figure 6, in which the attention weights in height and width directions of the input feature map can be obtained respectively. The feature visualization comparison is shown in Figure 7. Compared with YOLOv5s, the features of CA-YOLOv5s are more focused on the pedestrian region.

Table 2. Parameter amount and model compression ratio of different improved pedestrian detection algorithms with different attention modules.

No.	Detection Algorithm	Model Size	Parameter Amount (MB)	Model Compression Ratio (%)
1	YOLOv5s	640 × 640	26.88	—
2	CBAM-YOLOv5s	640 × 640	22.50	16.29
3	CA-YOLOv5s	640 × 640	22.47	16.41
4	SE-YOLOv5s	640 × 640	27.63	−2.79
5	ECA-YOLOv5s	640 × 640	22.37	16.78

Table 3. Performance comparison of different improved pedestrian detection algorithms with different attention modules.

No.	Detection Algorithm	AP (%)	Recall (%)	FPS
1	YOLOv5s	89.35	82.09	73
2	CBAM-YOLOv5s	89.19	82.99	74
3	CA-YOLOv5s	89.99	82.82	75
4	SE-YOLOv5s	88.34	81.34	72
5	ECA-YOLOv5s	88.94	81.32	78

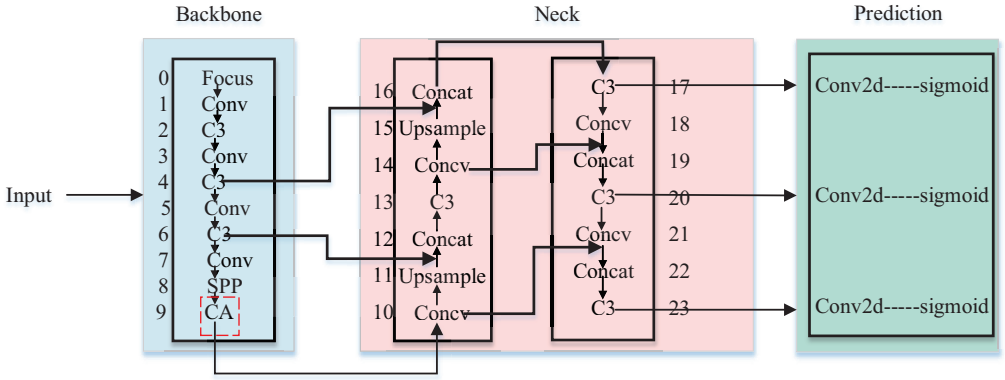


Figure 5. Network structure of CA-YOLOv5s.

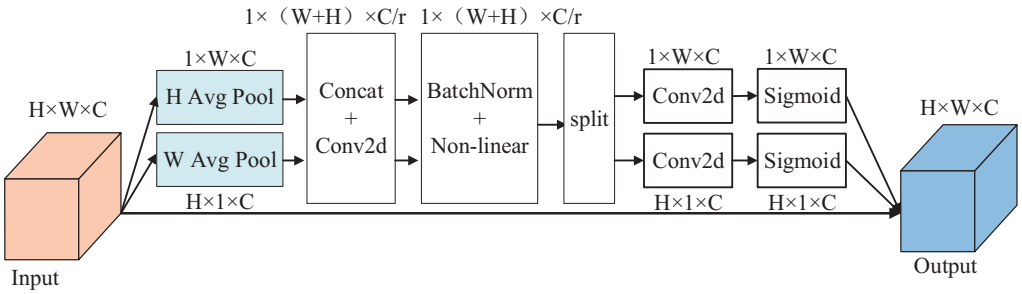


Figure 6. Detailed network structure of CA.

Figure 7. Feature visualization comparison of the last layer of the backbone network between YOLOv5s and CA-YOLOv5s. (**a**) Input image. (**b**) Feature visualization of C3 in YOLOv5s. (**c**) Feature visualization of CA in CA-YOLOv5s.

The output of the proposed CA-YOLOv5s pedestrian detection algorithm is shown in Figure 8. Figure 8a contains a real pedestrian and a PPWB, and Figure 8b contains a real pedestrian and a PPWNB. The output includes the bounding box of the detected pedestrian, the coordinate information of the bounding box, the label and the confidence. Table 4 illustrates the coordinates of the top left corner and the bottom right corner of the bounding box in Figure 8a, as well as the label and confidence of the detected pedestrian. As shown in Figure 8, the real pedestrian, the PPWB and the PPWNB are all detected as a pedestrian by the CA-YOLOv5s algorithm. Therefore, the real pedestrian and pseudo pedestrian should be further distinguished on this basis.

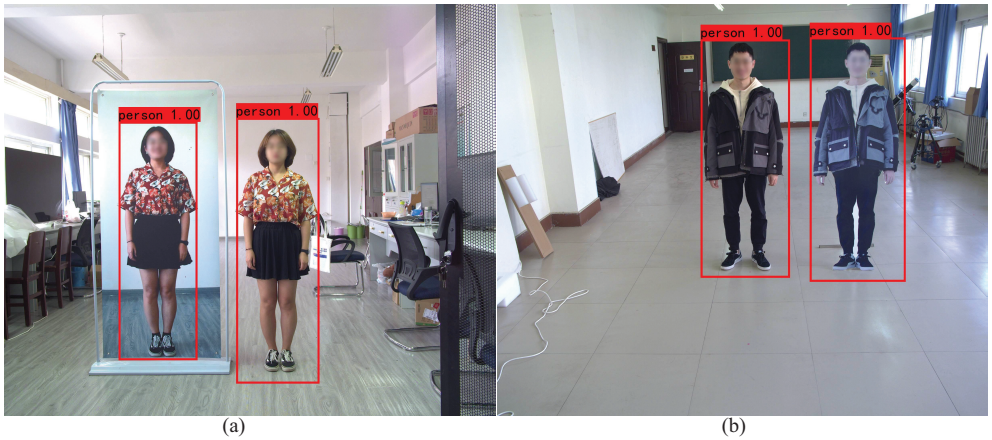


Figure 8. Output images of CA-YOLOv5s pedestrian detection algorithm. (**a**) Containing PPWB. (**b**) Containing PPWNB.

Table 4. Output data of Figure 7a.

No.	Label and Confidence	Top Left Corner, Bottom Right Corner
1	person 1.00	(570, 585), (962, 1770)
2	person 1.00	(1155, 566), (1569, 1886)

3.2. Binocular Stereo Matching and Ranging

In the binocular stereo matching module, the extracted left-view pedestrian region ROI_L and the right-view pedestrian region ROI_R are stereo matched by SURF matching [36], so as to obtain the multiple matched feature point pairs $(\mathbf{p}_{Li}, \mathbf{p}_{Ri}), i = 1, 2, \ldots, N$ and the corresponding 2D coordinates $\mathbf{P}_{Li}(x_{Li}, y_{Li})$ (in ROI_L) and $\mathbf{P}_{Ri}(x_{Ri}, y_{Ri})$ (in ROI_R). Figure 9 shows a pair of extracted pedestrian regions and their matching result. Then the calibration parameters f_L (left focal length), f_R (right focal length), $\mathbf{R}$ (rotation matrix) and $\mathbf{T}$ (translation matrix) of the binocular stereo camera are calculated by Zhengyou Zhang's calibration method [37], wherein $\mathbf{R} = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix}$ and $\mathbf{T} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$.

(a) (b)

Figure 9. A pair of extracted pedestrian regions and their matching results. (**a**) Pedestrian region pair. (**b**) Matching results.

In the binocular stereo ranging module, the 3D coordinates $\mathbf{P}_i(x_i, y_i, z_i)$ of the matched feature point pair $(\mathbf{P}_{Li}, \mathbf{P}_{Ri})$ are calculated using $\mathbf{P}_L(x_L, y_L)$, $\mathbf{P}_R(x_R, y_R)$, f_L, f_R, $\mathbf{R}$ and $\mathbf{T}$ according to Equation (17) [37]. All these spatial feature points $\mathbf{P}_i(x_i, y_i, z_i)$ form a feature point set $S = \{\mathbf{P}_j(x_j, y_j, z_j), j = 1, 2, \ldots, M\}$. The space distance d_i between each spatial feature point $\mathbf{P}_i$ in S and the optical center $\mathbf{O}_L$ of the left-view camera, namely the origin of the world coordinate system, is calculated according to Equation (18). The mean value $\bar{d}$ and standard deviation σ of all d_i are derived according to Equations (19) and (20). The absolute difference $|\Delta d_i|$ between each d_i and the mean value $\bar{d}$ is computed according to Equation (21).

$$\begin{cases} x = zx_L/f_L \\ y = zy_L/f_L \\ z = \dfrac{f_L(f_R t_x - x_R t_z)}{x_R(r_7 x_L + r_8 y_L + f_L r_9) - f_R(r_1 x_L + r_2 y_L + f_L r_3)} \\ \quad = \dfrac{f_L(f_R t_y - y_R t_z)}{y_R(r_7 x_L + r_8 y_L + f_L r_9) - f_R(r_4 x_L + r_5 y_L + f_L r_6)} \end{cases} \tag{17}$$

$$d_i = \sqrt{x_i^2 + y_i^2 + z_i^2}, \quad i = 1, 2, \ldots, N \tag{18}$$

$$\bar{d} = \sum_{i=1}^{N} d_i / N, \quad i = 1, 2, \ldots, N \tag{19}$$

$$\sigma = \sqrt{\sum_{i=1}^{N} \left(d_i - \bar{d}\right)^2 / N}, \quad i = 1, 2, \ldots, N \tag{20}$$

$$|\Delta d_i| = \left|d_i - \bar{d}\right| \tag{21}$$

Since there may exist some mismatched feature points in S, the direct use of these feature points in the plane fitting process will lead to a large deviation in the fitting plane, and will affect the final real and pseudo pedestrian judgment. Therefore, the mismatched feature points in S should be eliminated first. If $|\Delta d_i| > \sigma$, it is considered that $\mathbf{P}_i$ is not within the constraint range of the space distance standard deviation σ in S and is an outlier, which should be removed. If $|\Delta d_i| \leq \sigma$, it is considered that $\mathbf{P}_i$ is within the constraint range of the space distance standard deviation σ in S and is a matched point, which should be reserved. Finally, a matched feature point set $S_{\text{match}} = \{\mathbf{P}_j(x_j, y_j, z_j), j = 1, 2, \ldots, M\}$ is obtained, wherein $M \leq N$. Compared with S, the precision of the fitting plane and the accuracy of SVM prediction can be improved by eliminating the mismatched feature points and reserving only the correctly matched feature points. So far, the pedestrian region extraction, binocular stereo matching and binocular stereo ranging have been realized, and the 3D information required for the real and pseudo pedestrian judgment is acquired.

3.3. SVM Prediction

In the human visual system, real and pseudo pedestrians are distinguished according to the difference of the 3D information. In the proposed method, this process can be achieved by predicting the 3D information by SVM. The mean values of all M feature points $\mathbf{P}_j$ in S_{match} in the x, y, z directions are firstly calculated, and a new point $\overline{\mathbf{P}}(\bar{x}_{\text{match}}, \bar{y}_{\text{match}}, \bar{z}_{\text{match}})$ can be obtained, as expressed in Equation (22). The space distance $\bar{d}_{\text{match}}$ of $\overline{\mathbf{P}}$ is derived to represent the space distance between the pedestrian and the camera, as expressed in Equation (23).

$$\overline{\mathbf{P}}(\bar{x}_{\text{match}}, \bar{y}_{\text{match}}, \bar{z}_{\text{match}}) = \left(\sum_{j=1}^{M} x_j / M, \sum_{j=1}^{M} y_j / M, \sum_{j=1}^{M} z_j / M\right) \tag{22}$$

$$\bar{d}_{\text{match}} = \sqrt{\bar{x}_{\text{match}}^2 + \bar{y}_{\text{match}}^2 + \bar{z}_{\text{match}}^2} \tag{23}$$

As shown in Figure 10, the feature points in S_{match} are distributed in a spatial range with a certain thickness for the real pedestrian, while the feature points in S_{match} are almost on the same plane for the pseudo pedestrian. Therefore, the real and pseudo pedestrian can be distinguished according to the standard deviation $\sigma_{d_{fit}}$ of the distance d_{fit_j} from all the feature points $\mathbf{P}_j$ in S_{match} to their fitting plane α_{Fit}.

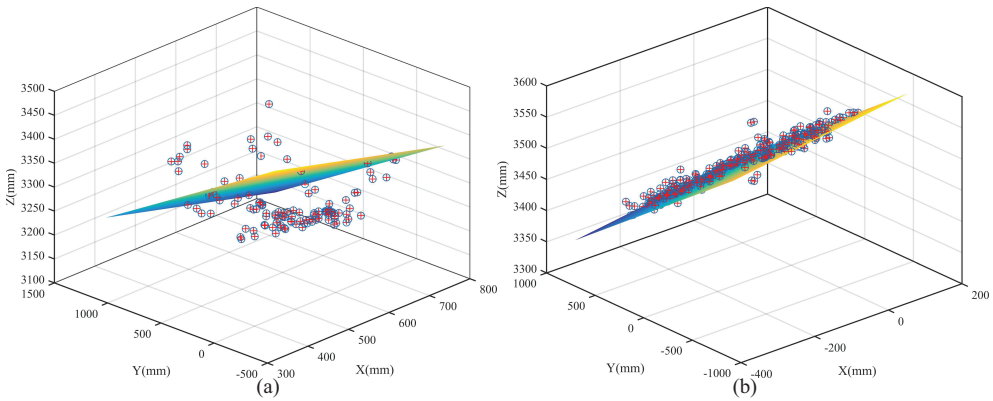


Figure 10. Feature point distribution of real and pseudo pedestrians. (**a**) Real pedestrian. (**b**) Pseudo pedestrian.

For plane fitting, the random sample consensus (RANSAC) plane-fitting algorithm can fit most points to be fitted, and eliminate invalid points according to a preset threshold, which will effectively reduce the interference from matching errors [48–51]. The threshold TH should be pre-determined before the plane fitting with RANSAC. The TH can be set according to the human body error tolerance ε, which is half of the human body thickness. Not only is the human body thickness related to the chest thickness, but it is also related to the clothes to wear. In the national standard GB/T 10000 [52], a total of 47 basic human size data from six regions of the country are provided. Among them, the bare chest thickness is $W \in [0.155\text{ m}, 0.268\text{ m}]$, then $W/2 \in [0.077\text{ m}, 0.134\text{ m}]$. Considering another thickness increment, 0.03 m, of the clothes, the human body error tolerance is $\varepsilon \in [0.077\text{ m}, 0.164\text{ m}]$.

RANSAC plane fitting is performed on S_{match} according to TH, and a spatial plane α_{Fit} is obtained, as shown in Equation (24). The distance d_{fit_j} from all M feature points $\mathbf{P}_j$ in S_{match} to the fitting plane α_{Fit} is computed, as shown in Equation (25). The standard deviation $\sigma_{d_{fit}}$ of d_{fit_j} is derived, as shown in Equation (26).

$$Ax + By + Cz + D = 0 \tag{24}$$

$$d_{fit\_j} = \frac{\left|Ax_j + By_j + Cz_j + D\right|}{\sqrt{A^2 + B^2 + C^2}} \tag{25}$$

$$\sigma_{d_{fit}} = \sqrt{\frac{\sum_{j=1}^{M}\left(d_{fit_j} - \sum_{j=1}^{M} d_{fit}/M\right)^2}{M}} \tag{26}$$

Figure 11 is a distribution diagram of the randomly selected real and pseudo pedestrian experimental data in the $\bar{d}_{\text{match}}$ and $\sigma_{d_{fit}}$ coordinates. The horizontal axis $\bar{d}_{\text{match}}$ is the space distance between the pedestrian and the camera, and the vertical axis $\sigma_{d_{fit}}$ is the standard deviation of the distance from all feature points in the human region to the fitting plane. The blue circle represents the real pedestrian, and the red asterisk represents the pseudo pedestrian. As can be seen from Figure 11, within the spatial range of 2–12 m, the experimental data conform to the first-order linear separability law. Thus, the binary classification method can be selected for the real and pseudo pedestrian classification.

Common binary classifiers include Bayesian classifier [53], decision tree classifier [54], back propagation (BP) classifier [55] and SVM classifier [56]. As shown in Figure 11, the two input variables of the classifier are positively correlated. The input variables are required to be independent to each other for the Bayesian classifier, so it is not applicable. There are still a few points in S_{match} with relatively large matching error, which will lead to overfitting, so the decision tree classifier is not applicable either. Meanwhile, the binary classification problem may have multiple feasible solutions, and the BP classifier can only work out one feasible solution but not the optimal solution. The SVM classifier is the statistically optimal solution among many feasible solutions, and has higher generalization performance than the BP network. Therefore, the SVM classifier is chosen to classify the data in this paper. The training and predicting process of SVM classifier is expressed in Algorithm 1.

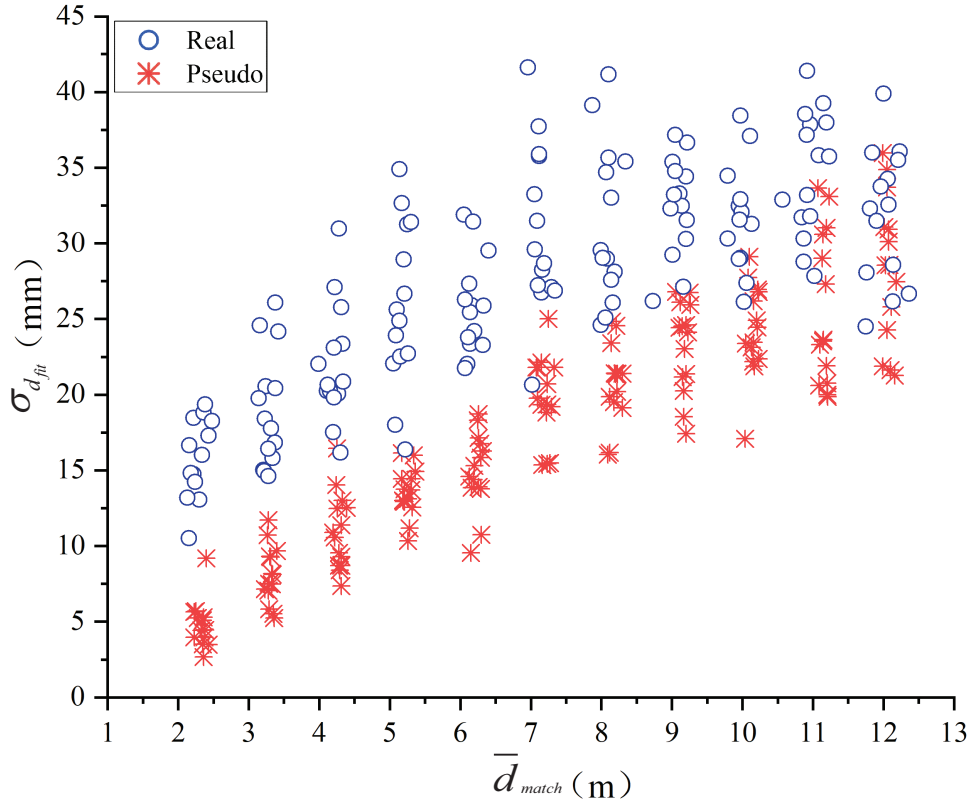


Figure 11. Distribution diagram of the real and pseudo pedestrian in $\bar{d}_{\text{match}}$ and $\sigma_{d_{fit}}$ coordinates.

Algorithm 1 The SVM training and predicting process.

Input: Label, $\overline{d}_{re}$, σ_{re} of the targets in the training set bounding box; $\overline{d}_{re}$, σ_{re} of the targets in the new bounding box;

Output: Label of the targets in the new bounding box;

1: Put Label, $\overline{d}_{re}$, σ_{re} of the target in the training set bounding box into the SVM for training;
2: The true and pseudo classification model is obtained by SVM training;
3: Send the $\overline{d}_{re}$ and σ_{re} of the target in the new bounding box to the true and pseudo classification model for prediction;
4: **return** The label of the target in the new bounding box.

For the training process of the SVM classifier, the input is a first-order linear separable training set $TS = \{(\mathbf{x}_i, y_i), i = 1, 2, \cdots N\}$, wherein, $\mathbf{x}_i\left(\bar{d}_{\text{match\_}i}, \sigma_{d_{fit-}i}\right)$ is the feature vector, also known as an instance; and $y_i \in \{-1, 1\}$ is the class label of $\mathbf{x}_i$. If $\mathbf{x}_i$ corresponds to the real pedestrian, $y_i = 1$; and if $\mathbf{x}_i$ corresponds to the pseudo pedestrian, $y_i = -1$. The output is the maximal margin separation hyperplane (MMSH) and the real and pseudo pedestrian classification model.

The optimization process for linear separable SVM can be expressed by Equation (27) [57]:

$$\begin{aligned} &\min_{\omega,b} \tfrac{1}{2}\|\boldsymbol{\omega}\|^2 \\ &s.t. \quad y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{27}$$

Wherein $\boldsymbol{\omega}$ and b are the normal vector and intercept of the separation hyperplane, and the optimal solutions $\boldsymbol{\omega}^*$ and b^* are the normal vector and intercept of the MMSH, which is represented by Equation (28).

$$\boldsymbol{\omega}^* \cdot \mathbf{x} + b^* = 0 \tag{28}$$

The real and pseudo pedestrian classification model can be represented by Equation (29) and can be used in the predicting process of the SVM classifiers.

$$f(\mathbf{x}) = \mathrm{sgn}(\boldsymbol{\omega}^* \cdot \mathbf{x} + b^*) = \begin{cases} -1, & \text{pseudo pedestrian} \\ 1, & \text{real pedestrian} \end{cases} \tag{29}$$

Next, the threshold TH for plane fitting is increased from 0.07 m to 0.17 m, with a step of 0.01 m. The performance indices of the SVM classification results with different TH are compared, and the optimal threshold TH_{opt} is selected. In the TH optimization experiment, 64 volunteers acted as real pedestrians, and two flat panels with photos of person and two human-shaped signboards were used as pseudo pedestrians. In total, 1000 images with single pedestrian were captured, from which 783 images were randomly selected, including 394 real pedestrians and 389 pseudo pedestrians. Then, 626 images were randomly selected from the 783 images as the training set, and the remaining 157 images were used as the verification set. Table 5 shows the performance comparison of the SVM classification results for different TH.

As can be seen from Table 5, when TH = 0.15 m, the SVM classification model for real and pseudo pedestrian can achieve the best performance in both accuracy and recall, and can achieve the second-best performance in precision, which is only 0.02% lower than the best one. Therefore, the optimal threshold TH_{opt} is 0.15 m. For TH_{opt}, the optimal solutions of the MMSH by SVM training are $\boldsymbol{\omega}^* = \begin{bmatrix} -0.69369225 & 0.26863033 \end{bmatrix}$ and $b^* = -1.41798519$, which can be further substituted into Equations (28) and (29) to obtain the labels of the real and pseudo pedestrians in the bounding box.

Table 5. Performance comparison of SVM classification for different TH.

TH	Accuracy	Precision	Recall
0.07	78.34%	85.29%	70.73%
0.08	83.44%	91.89%	77.27%
0.09	85.99%	86.05%	88.10%
0.1	82.80%	87.50%	80.46%
0.11	84.71%	87.50%	80.77%
0.12	86.62%	87.18%	86.08%
0.13	87.26%	89.74%	85.37%
0.14	84.08%	82.28%	85.53%
0.15	91.72%	90.91%	92.11%
0.16	88.54%	90.00%	87.80%
0.17	85.35%	83.75%	87.01%

4. Experiments

In the practical real and pseudo pedestrian detection test, two industrial cameras and a laptop are used. The Hikvision MV-CA050-11UC industrial camera has a resolution of 2448 × 2048, with a Wallis WL1608-5MP fixed-focus lens of 8 mm. The laptop is equipped with an Intel Core i7-10750H CPU, 16 GB RAM, and a Nvidia RTX2060 6G graphics card. The cell size of the calibration board is 30 mm × 30 mm. Two groups of experiments are conducted with different arrangement mode of pedestrians, i.e., equidistant arrangement mode and random arrangement mode. In the testing experiment, 71 volunteers acted as real pedestrians, and two flat panels with person photos and two human-shaped signboards were used as pseudo pedestrians. A total of 455 testing images with no occlusion were captured in the two groups of experiments, among which 212 are real pedestrians and 243 are pseudo pedestrians. In the first group of experiment with pedestrians in equidistant arrangement mode, a total of five shooting scenes were designed, that is, the pedestrian number was increased from one to five successively, and every one image was collected every one meter. In the second group of experiments with pedestrians in a random arrangement mode, a total of three shooting scenes were designed, that is, the pedestrian

number was increased from three to five successively, and the pedestrians stood randomly. The main purpose of the proposed real and pseudo pedestrian detection method with CA-YOLOv5s based on stereo image fusion is to solve the problem of pseudo pedestrian detection, therefore, the testing experiments are mainly designed to verify the effect of pseudo pedestrian detection. For this reason, in the experiment, at least one pseudo pedestrian exists in each image where there is more than one pedestrian in it. As shown in Table 6, the real and pseudo pedestrian number setting is designed for five different total numbers, ranging from 1 to 5.

Table 6. Real and pseudo pedestrian number setting.

Pedestrian Number	Real Pedestrian Number	Pseudo Pedestrian Number
1	1 0	0 1
2	1 0	1 2
3	1 2 0	2 1 3
4	1 2 3 0	3 2 1 4
5	1 2 3 4	4 3 2 1

4.1. Experiments in Equidistant Arrangement Mode

Figure 12 shows the point plots of the real and pseudo pedestrian detection results by the proposed method for 1–5 pedestrians arranged equidistantly, wherein the real pedestrian is represented by the label RP, and the pseudo pedestrian is represented by the label PP. A dot line represents the data points of a same pedestrian at different distances, different dot lines for different pedestrians. The MMSH is represented by a red line. If the data point is above the MMSH, it means that the pedestrian detected is a real one; if not, a pseudo one. If a RP data point is below the MMSH or a PP data point is above the MMSH, error detection occurs. For the same group of pedestrians, every image is collected every one meter at a distance from 2 m to 12 m, and 10 images can be collected for each group of pedestrians. However, when collecting images of five equidistantly arranged pedestrians, the target may not be captured due to the close distance, but the detection result will not be affected. For example, in Figure 12e, only nine images of RP10 are collected. As shown in Figure 12, the proposed method can correctly detect most data points of the real or pseudo pedestrians, but also with a small amount of error detections. The pedestrian becomes smaller with the increase in the distance, and the features become not obvious, and hence, the number of mismatched feature points increases, and the standard deviation from the feature point set to the fitting plane becomes inaccurate, resulting in the wrong classification of pedestrians. The number of error detection instances increases with the distance.

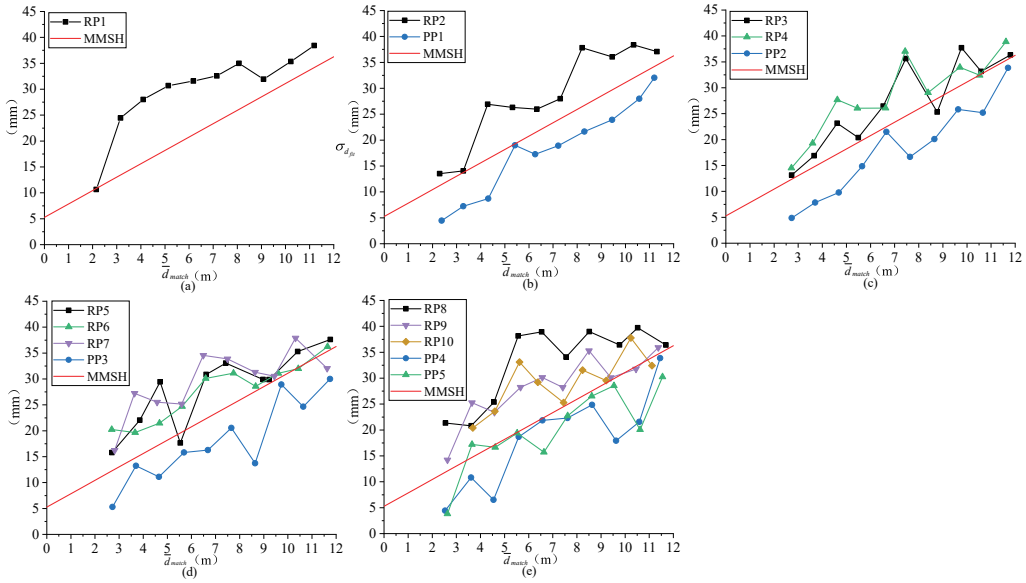


Figure 12. Point plots of the real and pseudo pedestrian detection results for different number of pedestrians in equidistant arrangement mode. (**a**) One pedestrian. (**b**) Two pedestrians. (**c**) Three pedestrians. (**d**) Four pedestrians. (**e**) Five pedestrians.

Table 7 presents the partial detailed data of Figure 13, wherein the label '1' represents the real pedestrian and the label '−1' represents the pseudo pedestrian. As shown in Table 7, the actual label and the predicted label are the same for most data. However, for the image with four pedestrians, the actual label for the third pedestrian is −1, while the predicted label is 1, and an error detection occurs. After judging the classification of the pedestrians in the bounding box, the predicted label is combined with the coordinate information of the bounding box for output. Figure 13 shows the output images of the corresponding pedestrians in Table 7. The real pedestrian is displayed in red bounding box marked as RP, while the pseudo pedestrian is displayed in a blue bounding box marked as PP. In Figure 13d, the third (from left to right) target is PP, but detected as RP, and an error detection occurs. This small number of error detection instances is caused by matching errors.

Table 7. Partial detailed data of Figure 13.

Pedestrian Number	No.	$\bar{d}_{\mathbf{match}}$ (m)	$\sigma_{d_{fit}}$ (mm)	Actual Label	Predicted Label
1	1	6.26	28.37	1	1
2	1	6.76	16.71	−1	−1
	2	6.57	31.53	1	1
3	1	6.53	26.45	1	1
	2	6.66	21.49	−1	−1
	3	6.62	26.08	1	1
4	1	3.72	11.19	−1	−1
	2	3.68	12.21	−1	−1
	3	3.71	19.68	−1	1
	4	3.67	19.81	1	1
5	1	3.75	20.98	1	1
	2	3.62	12.71	−1	−1
	3	3.70	14.54	−1	−1
	4	3.77	7.55	−1	−1
	5	3.85	26.15	1	1

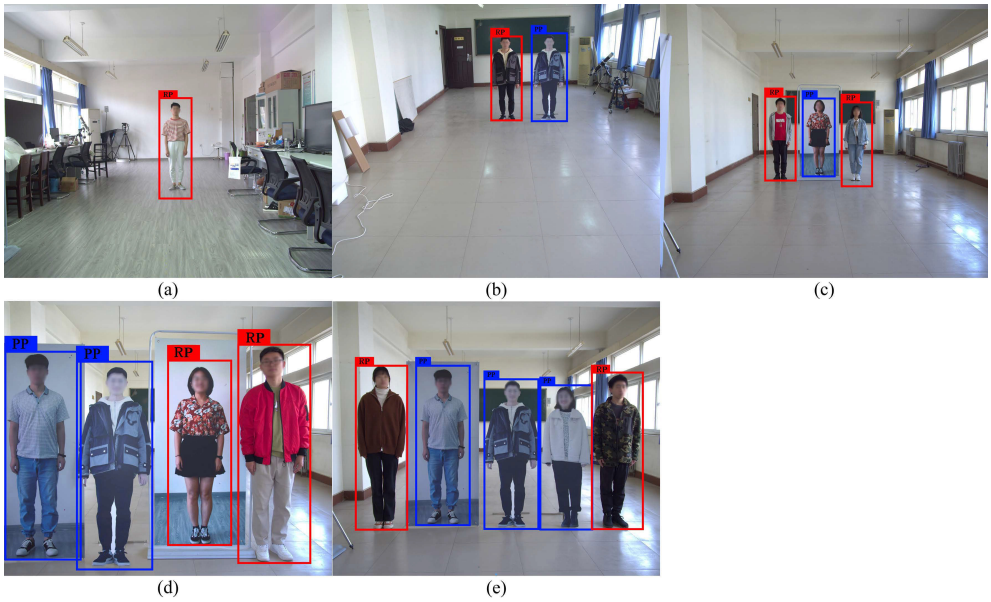


Figure 13. Exemplary images of the pedestrian detection outputs of the data in Table 7. (**a**) One pedestrian. (**b**) Two pedestrians. (**c**) Three pedestrians. (**d**) Four pedestrians. (**e**) Five pedestrians.

4.2. Experiments in Random Arrangement Mode

Figure 14 shows the point plots of the real and pseudo pedestrian detection results by the proposed method for 3–5 pedestrians arranged randomly. As shown in Figure 14, the proposed method can correctly detect most data points of the real or pseudo pedestrians, but also with a small amount of error detections. Table 8 presents the detailed data of Figure 14. As shown in Table 8, the actual label and the predicted label are the same for most data. However, for the image with five pedestrians, the actual label for the second pedestrian is 1, while the predicted label is −1, and an error detection occurs. Figure 15 shows the output images of the corresponding pedestrians in Table 8. In Figure 15c, the second (from left to right) target is RP, but detected as PP, and an error detection occurs. This small number of error detection instances is caused by the randomness of the feature points.

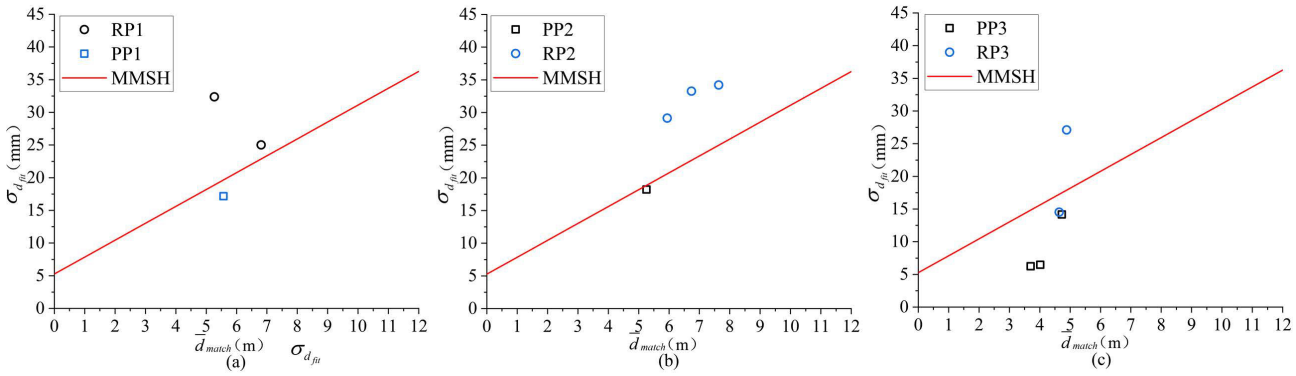
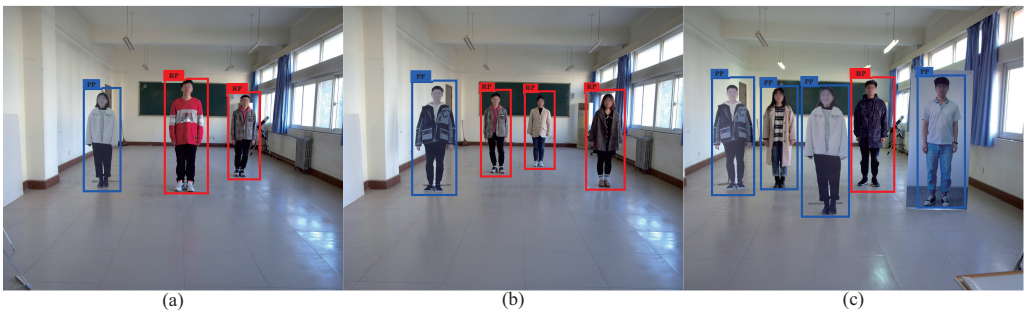


Figure 14. Point plots of the real and pseudo pedestrian detection results for different number of pedestrians in random arrangement mode. (**a**) Three pedestrians. (**b**) Four pedestrians. (**c**) Five pedestrians.

Table 8. Detailed data of Figure 14.

Pedestrian Number	No.	$\bar{d}_{\text{match}}$ (m)	$\sigma_{d_{fit}}$ (mm)	Actual Label	Predicted Label
3	1	5.57	17.17	−1	−1
	2	5.27	32.38	1	1
	3	6.81	25.01	1	1
4	1	5.26	18.21	−1	−1
	2	6.74	33.25	1	1
	3	7.64	34.21	1	1
	4	5.94	29.13	1	1
5	1	4.73	14.16	−1	−1
	2	4.64	14.51	1	−1
	3	3.70	6.24	−1	−1
	4	4.89	27.09	1	1
	5	4.02	6.47	−1	−1



(a) (b) (c)

Figure 15. Exemplary images of the pedestrian detection output of the data in Table 8. (**a**) Three pedestrians. (**b**) Four pedestrians. (**c**) Five pedestrians.

Table 9 shows the performance indices of the pedestrian detection on the 455 testing images captured in the two groups of experiments, with TH_{opt} as 0.15 m. TP (True Positive) corresponds to the real label '1' and the predicted label '1'. FN (False Negative) corresponds to the real label '1' and the predicted label '−1'. TN (True Negative) corresponds to the real label '−1' and the predicted label '−1'. FP (False Positive) corresponds to the real label '−1' and the predicted label '1'. The accuracy is 93.85%, the precision is 93.81%, and the recall is 92.93%, achieving good performance for the real and pseudo pedestrian detection.

Table 9. Detection performance on the 455 testing images.

TH_{opt}	TP	FN	TN	FP	Accuracy	Precision	Recall
0.15	197	15	230	13	93.85%	93.81%	92.93%

4.3. Contrast Experiments

The performance of the proposed method is tested and compared with seven other pedestrian detection algorithms on a same test set. Table 10 shows the performance comparison of real and pseudo pedestrian detection among different algorithms. Considering that the number of the real pedestrian is much greater than that of the pseudo pedestrians in practice, 249 images were randomly selected from the 455 testing images captured in the two groups of experiments as the testing set in the comparison experiment, of which 212 were real pedestrians and 37 were pseudo pedestrians.

Table 10. Performance comparison of eight different pedestrian detection algorithms on the 455 testing images.

Pedestrian Detection Algorithms	Accuracy	Precision	Recall
SSD	86.35%	87.71%	97.64%
RFB	85.14%	86.31%	98.11%
RetinaNet	85.54%	87.93%	96.23%
M2Det	85.94%	89.69%	94.34%
YOLOv4	85.14%	85.14%	100.00%
YOLOv5s	85.14%	85.14%	100.00%
CA-YOLOv5s	85.14%	85.14%	100.00%
ours	93.17%	98.99%	92.92%

As shown in Table 10, the accuracy of the seven pedestrian detection algorithms, SSD, RFB, RetinaNet, M2Det, YOLOv4, YOLOv5s and CA-YOLOv5s ranges from 85.14% to 86.35%, the precision from 85.14% to 89.69%, and the recall from 94.34% to 100%. The recalls of YOLOv4, YOLOv5s and CA-YOLOv5s are all 100%, which indicates that these three algorithms can detect all the real pedestrians in the dataset. The precisions and accuracies are all 85.14%, which means that all the pseudo pedestrians are detected as real pedestrians, i.e., the real and pseudo pedestrians cannot be distinguished. For the proposed method, the accuracy is 93.17%, the precision is 98.99%, and the recall is 92.92%. The accuracy and precision of the real and pseudo pedestrian detection are significantly superior to the other algorithms. Therefore, the real and pseudo pedestrian detection method proposed in this paper with CA-YOLOv5s based on stereo image fusion can effectively detect pseudo pedestrians, and greatly improve the accuracy and precision of the pedestrian detection network for real and pseudo pedestrian detection.

5. Conclusions

To solve the problem of pseudo pedestrian detection, a bionic model for the real and pseudo pedestrian detection based on human stereo vision is constructed in this paper, and a detection method with CA-YOLOv5s based on stereo image fusion for the real and pseudo pedestrian detection is proposed. In the proposed method, the YOLOv5s pedestrian detection algorithm is improved by combining with the CA attention mechanism, which not only increases the detection accuracy, but also compresses the network model size. Then, stereo matching and ranging are performed on the detected pedestrian regions based on stereo image fusion so as to obtain the 3D information of the pedestrian. Next, the trained SVM classifier is used to predict the 3D information features of the real and pseudo pedestrians extracted by the plane fitting, which can effectively distinguish between the real and pseudo pedestrians. Experimental results show that the proposed method can correctly predict the real and pseudo pedestrians and effectively solve the problem that the existing pedestrian detection algorithms cannot distinguish between real and pseudo pedestrians well.

Author Contributions: Conceptualization, X.S. and L.Y.; methodology, L.Y. and G.L.; formal analysis, X.S. and L.Y.; data construction, G.L. and L.Z.; writing—original draft preparation, G.L.; writing—review and editing, L.Y., X.S., G.L. and L.Z.; supervision, L.Y., X.S., C.H. and Z.X.; funding acquisition, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the ZhongYuan Science and Technology Innovation Leading Talent Program under Grant 214200510013, in part by the National Natural Science Foundation of China under grant 62171318, in part by the Key Research Project of Colleges and Universities in Henan Province under Grant 21A510016 and Grant 21A520052, in part by the Scientific Research Grants and Start-up Projects for Overseas Student under Grant HRSS2021-36, and in part by the Major Project Achievement Cultivation Plan of Zhongyuan University of Technology under Grant K2020ZDPY02.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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

References

1. Pattanayak, S.; Ningthoujam, C.; Pradhan, N. A survey on pedestrian detection system using computer vision and deep learning. In *Advanced Computational Paradigms and Hybrid Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 419–429.
2. Zhang, S.; Xie, Y.; Wan, J.; Xia, H.; Li, S.Z.; Guo, G. WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild. *IEEE Trans. Multimed.* **2020**, *22*, 380–393. [CrossRef]
3. Dollár, P.; Appel, R.; Belongie, S.; Perona, P. Fast Feature Pyramids for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1532–1545. [CrossRef] [PubMed]
4. Cao, J.; Pang, Y.; Li, X. Learning Multilayer Channel Features for Pedestrian Detection. *IEEE Trans. Image Process.* **2017**, *26*, 3210–3220. [CrossRef]
5. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
6. Tesema, F.B.; Wu, H.; Chen, M.; Lin, J.; Zhu, W.; Huang, K. Hybrid channel based pedestrian detection. *Neurocomputing* **2020**, *389*, 1–8. [CrossRef]
7. Faster, R. Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *9199*, 2969239–2969250.
8. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
9. Gatto, B.B.; Souza, L.S.; dos Santos, E.M.; Fukui, K.; S Júnior, W.S.; dos Santos, K.V. A semi-supervised convolutional neural network based on subspace representation for image classification. *EURASIP J. Image Video Process.* **2020**, *2020*, 22. [CrossRef]
10. Cheng, H.; Zheng, N.; Qin, J. Pedestrian detection using sparse Gabor filter and support vector machine. In Proceedings of the IEEE Proceedings. Intelligent Vehicles Symposium, Las Vegas, NV, USA, 6–8 June 2005; pp. 583–587. [CrossRef]
11. Dollár, P.; Tu, Z.; Perona, P.; Belongie, S. Integral channel features. In *Proceedings of the British Machine Vision Conference*; BMVC Press: London, UK, 2009.
12. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. What can help pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3127–3136.
13. Yu, X.; Si, Y.; Li, L. Pedestrian detection based on improved Faster RCNN algorithm. In Proceedings of the 2019 IEEE/CIC International Conference on Communications in China (ICCC), Changchun, China, 11–13 August 2019; pp. 346–351.
14. Cao, J.; Pang, Y.; Xie, J.; Khan, F.S.; Shao, L. From handcrafted to deep features for pedestrian detection: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4913–4934. [CrossRef] [PubMed]
15. Zhang, S.; Yang, X.; Liu, Y.; Xu, C. Asymmetric multi-stage CNNs for small-scale pedestrian detection. *Neurocomputing* **2020**, *409*, 12–26. [CrossRef]
16. Xu, H.; Guo, M.; Nedjah, N.; Zhang, J.; Li, P. Vehicle and pedestrian detection algorithm based on lightweight YOLOv3-promote and semi-precision acceleration. *IEEE Trans. Intell. Transp. Syst.* **2022**. [CrossRef]
17. Lin, C.; Lu, J.; Zhou, J. Multi-Grained Deep Feature Learning for Robust Pedestrian Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3608–3621. [CrossRef]
18. Li, G.; Yang, Y.; Qu, X. Deep Learning Approaches on Pedestrian Detection in Hazy Weather. *IEEE Trans. Ind. Electron.* **2020**, *67*, 8889–8899. [CrossRef]
19. You, M.; Zhang, Y.; Shen, C.; Zhang, X. An Extended Filtered Channel Framework for Pedestrian Detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 1640–1651. [CrossRef]
20. Peng, B.; Chen, Z.B.; Fu, E.; Yi, Z.C. The algorithm of nighttime pedestrian detection in intelligent surveillance for renewable energy power stations. *Energy Explor. Exploit.* **2020**, *38*, 2019–2036. [CrossRef]
21. Noh, J.; Lee, S.; Kim, B.; Kim, G. Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
22. Lombacher, J.; Hahn, M.; Dickmann, J.; Wöhler, C. Potential of radar for static object classification using deep learning methods. In Proceedings of the 2016 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), San Diego, CA, USA, 19–20 May 2016; pp. 1–4. [CrossRef]
23. Shakeri, A.; Moshiri, B.; Garakani, H.G. Pedestrian Detection Using Image Fusion and Stereo Vision in Autonomous Vehicles. In Proceedings of the 2018 9th International Symposium on Telecommunications (IST), Tehran, Iran, 17–19 December 2018; pp. 592–596. [CrossRef]
24. Wei, W.; Cheng, L.; Xia, Y.; Zhang, P.; Gu, J.; Liu, X. Occluded Pedestrian Detection Based on Depth Vision Significance in Biomimetic Binocular. *IEEE Sens. J.* **2019**, *19*, 11469–11474. [CrossRef]

25. Zhao, Y.; Zhao, M.; Shi, F.; Jia, C.; Chen, S. Light-field imaging for distinguishing fake pedestrians using convolutional neural networks. *Int. J. Adv. Robot. Syst.* **2021**, *18*, 1729881420987400. [CrossRef]
26. Diner, D.B.; Fender, D.H. Stereoscopic Properties of the Human Visual System. In *Human Engineering in Stereoscopic Viewing Devices*; Springer: Boston, MA, USA, 1993; pp. 3–34.
27. Prasad, S.; Galetta, S.L. Anatomy and physiology of the afferent visual system. *Handb. Clin. Neurol.* **2011**, *102*, 3–19. [PubMed]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
31. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
32. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
33. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
34. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
35. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]
36. Bay, H.; Tuytelaars, T.; Gool, L.V. Surf: Speeded up robust features. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
37. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
38. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
39. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
40. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
41. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
42. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
43. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
44. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2det: A single-shot object detector based on multi-level feature pyramid network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9259–9266.
45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
46. Jiao, S.; Miao, T.; Guo, H. Image Target Detection Method Using the Yolov5 Algorithm. In *3D Imaging Technologies—Multidimensional Signal Processing and Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 323–329.
47. Available online: https://github.com/ultralytics/yolov5 (accessed on 5 August 2022).
48. Gallo, O.; Manduchi, R.; Rafii, A. CC-RANSAC: Fitting planes in the presence of multiple surfaces in range data. *Pattern Recognit. Lett.* **2011**, *32*, 403–410. [CrossRef]
49. Fan, M.; Jung, S.W.; Ko, S.J. Highly Accurate Scale Estimation from Multiple Keyframes Using RANSAC Plane Fitting with a Novel Scoring Method. *IEEE Trans. Veh. Technol.* **2020**, *69*, 15335–15345. [CrossRef]
50. Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; Guo, X. Locality preserving matching. *Int. J. Comput. Vis.* **2019**, *127*, 512–531. [CrossRef]
51. Fan, A.; Ma, J.; Jiang, X.; Ling, H. Efficient deterministic search with robust loss functions for geometric model fitting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
52. State Bureau of Technical Supervision. Chinese Adult Body Size: GB/T 10000-1988. 1988. Available online: https://www.chinesestandard.net/PDF.aspx/GBT10000-1988 (accessed on 5 August 2022).
53. Berrar, D. Bayes' theorem and naive Bayes classifier. *Encycl. Bioinform. Comput. Biol.* **2019**, *1*, 403–412. [CrossRef]
54. Priyanka; Kumar, D. Decision tree classifier: A detailed survey. *Int. J. Inf. Decis. Sci.* **2020**, *12*, 246–269.
55. Wang, J.Z.; Wang, J.J.; Zhang, Z.G.; Guo, S.P. Forecasting stock indices with back propagation neural network. *Expert Syst. Appl.* **2011**, *38*, 14346–14355. [CrossRef]

56. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
57. Wang, L. *Support Vector Machines: Theory and Applications*; Springer Science & Business Media: New York, NY, USA, 2005; Volume 177.