Hindawi
Advances in Multimedia
Volume 2022, Article ID 8940743, 9 pages
https://doi.org/10.1155/2022/8940743

Research Article

Pedestrian Detection Algorithm Combining Attention Mechanism and Nonmaximum Suppression Method

Duo Pan and Xuemei Zhou

College of Intelligent Manufacturing and Information Engineering, Sichuan Technology &Business College, DuJiangYan 611830, China

Correspondence should be addressed to Duo Pan; pan_duo72@sina.com

Received 18 January 2022; Revised 14 February 2022; Accepted 25 February 2022; Published 21 March 2022

Academic Editor: Qiangyi Li

With the involution of pedestrian detection technology, higher requirements are put forward for the detection accuracy under the conditions of insufficient light, target occlusion, and too small scale. Without information and multiscale pedestrian target, visible light single-mode pedestrian detection algorithm has poor performance. To solve the above problems, a pedestrian detection algorithm combining attention mechanism and nonmaximum suppression method is proposed in this study, aiming to improve the accuracy of pedestrian detection. In addition, residual network ResNet-50 and IoU (intersection over union) loss function are also adopted to improve pedestrian detection accuracy. Attention mechanism was used to optimize and highlight pedestrian area features, and meanwhile, the nonmaximum suppression method was applied to improve the robustness of the algorithm. Experimental results show that the detection accuracy of the proposed algorithm is significantly higher than that of the traditional convolutional neural network algorithm.

1. Introduction

Pedestrian detection, which is a hot and difficult problem in computer vision research and has been difficult to solve for a long time, is gradually applied to every aspect of daily life due to the rapid development of image processing technology [1]. However, with the complexity and diversity of human body posture, the problems of insufficient light and occlusion are more serious. Hence, it is of great difficulty to accurately detect pedestrians in various scenes [2]. Pedestrians are constantly moving in video frames or images, resulting in different positions and attitudes of the same pedestrian in different video frames or images. Additionally, the influence of objective factors such as shooting or burning will increase the difficulty of pedestrian detection as well [3]. The force majeure such as occlusion or visual blind area in the picture makes pedestrian detection become a very serious and time-sensitive problem in computer vision and has become a continuous topic and hot spot related to traffic and safety [4]. Through pedestrian detection, researchers can make use of real-time detection and high recognition rate to accurately obtain the pedestrian detection situation in the region. Then, on this basis, the relevant analysis and providing traffic protection can be explored, and also, mathematical statistics is applied to estimate and forecast [5]. Pedestrian detection methods based on various situations and directions emerge one after another on account of the advent of pedestrian detection. Among them, pedestrian detection methods based on machine learning gradually occupy the mainstream field and continue to be optimized. In this method, pedestrian feature classifier can be established by the matching feature extraction model, so as to achieve the purpose and result of pedestrian detection [6].

The main problems of pedestrian detection are as follows. (1) Local occlusion will greatly reduce the amount of information required for detection, leading to missed detection [7]. (2) It is difficult to extract features with strong discrimination from small-size pedestrians, leading to unsatisfactory detection results. Presently, most of the frontier pedestrian detection research work is based on deep

learning. Researchers proposed a multiscale pedestrian detection method so as to better detect different pedestrian instances with large size differences [8]. MS-CNN [9] proposed to obtain regional proposal frames on multiple feature maps of different scales and introduced context information into detection features to assist in improving detection effects. SA-Fast R-CNN [10] introduces multiple subnetworks to detect pedestrian instances of different scales and then selects the detection result of the subnetwork with the largest response as the final detection result.

Some methods improve pedestrian detection via introducing semantic segmentation tasks. F-DNN + SS [11] uses a single shot multibox detector (SSD) [12] as a basic detector to find candidate pedestrian instances. Then, semantic segmentation network is introduced to correct the detection results obtained by SSD. The final result is corrected by calculating the overlap rate of the mask obtained by the prediction box and the semantic segmentation module. The method proves that the effect of pedestrian detection can be improved by introducing the semantic segmentation module. SDS-RCNN [13] adds semantic segmentation tasks in the stage of candidate region generation and candidate region classification. During training, joint optimization is carried out to enhance the feature expression in the network and make the detection of occlusion pedestrians more robust. FRCN + A + DT [14] firstly added semantic segmentation branch in the region proposal stage of faster R-CNN [15] to improve the quality of the proposal frame. Then, in the detection stage, feature transformation is learned so that pedestrians and nonpedestrians can be better distinguished, so as to deal with occlusion.

Attention mechanism has also been adopted by researchers to deal with occlusion, thus improving pedestrian detection quality. The faster R-CNN + ATT [16] probe uses attention mechanism to weight different convolutional feature channels to deal with local occlusion of pedestrians. SSA-CNN [17] integrates semantic segmentation features as self-attention into the two stages of regional proposal and classification to improve the robustness of occlusion pedestrian detection.

In addition, RPN + BF [18] combined with the advantages of deep learning and machine learning proposed that faster R-CNN region proposal network should be first employed to generate candidate boxes. Then, depth features are extracted from the convolutional layer of the backbone network. Finally, the random forest method is exploited to classify the obtained features. Instead of using only a single frame image as input, Zelenov et al. [19] use time-consistent information in the video to improve pedestrian detection results. This method iteratively searches the corresponding part of the covered pedestrian in the current frame in the adjacent frame to form temporal tube and then performs adaptive weighting on the features from the temporal tube. Then, it gathers to the current frame to enhance the feature representation of the blocked pedestrian in the current frame.

Although there have been some good research achievements in pedestrian detection, the problems of local occlusion and small target size still deserve further discussion. In this study, the innovations and contributions of it are listed below.

(1) A pedestrian detection algorithm combining deep residual network and attention mechanism is proposed. The algorithm includes three aspects: algorithm model modification, attention-attracting mechanism, and nonmaximum suppression (NMS) technology.

(2) The branches of two complete convolution layers are used to predict the pixel-level boundary boxes and confidence scores, respectively. Meanwhile, IoU loss function optimization network is introduced to improve the accuracy of pedestrian detection.

(3) In the deep residual network, the attention mechanism is used to improve the algorithm's ability to understand complex scenes, reduce the interference of useless information such as occlusion and density, and achieve accurate extraction of effective target information for pedestrians. Finally, the nonmaximum suppression technique is utilized to improve the performance of the algorithm.

The structure of this paper is listed as follows. The recommendation algorithm proposed in this study is described in Section 2. Section 3 focuses on the experiment and analysis. Section 4 is the conclusion.

2. The Recommendation Algorithm Proposed in this Paper

2.1. Related Knowledge

2.1.1. Residual Network. In traditional convolutional neural networks, multilayer features become more abundant with the superposition of network layers. Therefore, the deeper the network layer is, the better the image processing effect is. Nevertheless, the simple superposition network will cause the problem of gradient disappearance and hinder the convergence of the model. Initialization and regularization can ensure the convergence of tens of layers of networks. However, in the deeper network, the effect becomes worse when the accuracy reaches saturation.

In view of the above situation, ResNet [20] introduces residual learning to solve the problem that deep network is difficult to optimize, and its module structure is shown in Figure 1. Let $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + M_1(M_2(F^c_{\text{max}})))$ represent the optimal mapping and use stacked nonlinear layers to fit another mapping $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + W_1(M_2(F^c_{\text{max}})))$. Then, the optimal mapping can be expressed as $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + M_1(M_2(F^c_{\text{max}})))$. Residual mapping adds fast connections to feedforward networks and performs simple identity mapping. In this way, no additional parameters and computational complexity are added, and it is easier to optimize than the original mapping [21].

The normal directly connected convolutional neural network is quite distinguished from ResNet, which has a bypass feeder that connects the feeder directly to the

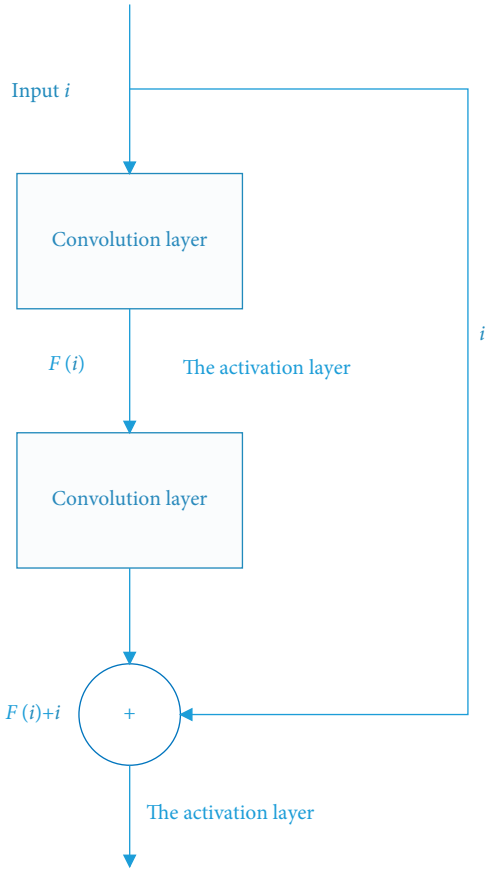


Figure 1: The residual network structure.

subsequent layer, allowing the subsequent layer to learn the residual directly. This structure is also known as direct connection or skip pass. In the traditional convolutional layer or full connection layer, there is more or less the problem of information loss in the process of information transmission. ResNet solves this problem to a certain extent. The integrity of information by passing input information directly to output is protected. The whole network only needs to learn input and output differences, simplifying the learning goal and difficulty.

2.1.2. IoU Loss Function. For each pixel (x, y) in the image, the bounding box of the true value can be defined as a 4-dimensional vector. i_t, i_b, i_l, i_r represent the distance between the current pixel position (x, y) and the upper and lower boundaries of the true value. Comments (x, y) are omitted for simplicity of calculation. Therefore, the predicted boundary box is defined as $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + M_1(M_2(F^c_{\text{max}})))$.

IoU is the intersection ratio between the prediction box and the real box, and the loss function of IoU is

$$\text{IoU loss} = -\ln\frac{\text{Intersection}(\text{Prediction}, \text{Ground truth})}{\text{Union}(\text{Prediction}, \text{Ground truth})}, \quad (1)$$

where Prediction indicates the predicted value, Ground truth is the true value, Intersection means intersection, and Union means intersection.

2.1.3. Attention Mechanism. At present, the most commonly used attention mechanisms in image processing include channel attention mechanism and spatial attention mechanism.

(1) Channel Attention Mechanism. Channel attention mechanism pays more attention to the channel information of image input and extracts the accuracy of feature classification through feature extraction of channel information. The channel attention module firstly carries out maximum pooling and average pooling, respectively, for the spatial dimension compression of the input feature graph $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + M_1(M_2(F^c_{\text{max}})))$, where C represents the number of channels in the input feature graph and H and W represent the length and width of the feature graph. Then, the channel attention map is calculated by sharing multilayer perceptron (MLP). Finally, activation function sigmoid is used for output, and channel attention feature graph $W_c(F) = \sigma(M_1(M_2(F^c_{\text{avg}})) + M_1(M_2(F^c_{\text{max}})))$ is obtained [22]. Its network structure is shown in Figure 2. And its calculation equation is where F is the input feature, F^c_{avg} and F^c_{max} represent average pooling and maximum pooling, respectively, and M_1 and M_2 represent the weights of two layers in a multilayer perceptron.

$$W_c(F) = \sigma\left(M_1\left(M_2\left(F^c_{\text{avg}}\right)\right) + M_1\left(M_2\left(F^c_{\text{max}}\right)\right)\right), \quad (2)$$

(2) Spatial Attention Mechanism. Spatial attention mechanism mainly focuses on the location information of the target in the image and selectively aggregates the features of each space through the weighted sum of spatial features. Input feature graph $F_s = 1/C\sum_{x\in C}F(x) + \max_{x\in C}F(x)$, and perform maximum pooling and average pooling successively for the input features, as shown in equation (3). Then, it is processed by $7 * 7$ convolution kernel and sigmoid activation function, as shown in equation (4). Feature $F_s = 1/C\sum_{x\in C}F(x) + \max_{x\in C}F(x)$ of spatial attention weight is obtained, and its network structure is shown in Figure 3:

$$F_s = \frac{1}{C}\sum_{x\in C}F(x) + \max_{x\in C}F(x), \quad (3)$$

$$W_S = \sigma\left(f^{7\times 7}(F_s)\right). \quad (4)$$

2.1.4. Nonmaximum Suppression Method. For detection tasks, the nonmaximum suppression (NMS) algorithm is a postprocessing algorithm for redundancy removal of detection results. Greedy clustering is performed based on a fixed distance threshold. That is, the detection results with high scores are greedily selected and the adjacent results exceeding the preset threshold are deleted to achieve a trade-off between recall rate and accuracy. IoU loss function was used to extract all pedestrian detection frames within the threshold. The above detection frames are sorted according to their scores, and the one with the highest scores is selected. Then, how much the other boxes overlap with the current box is calculated. If the degree of overlap is greater

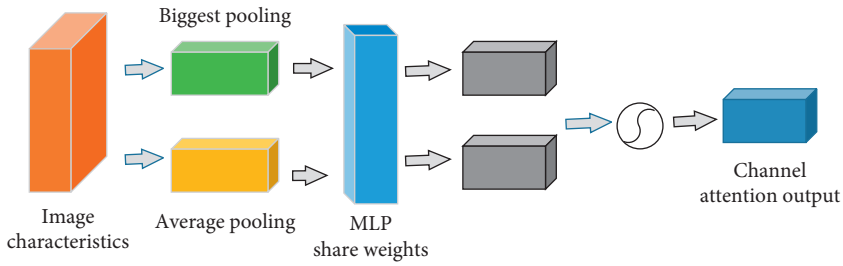


Figure 2: The channel attention structure.

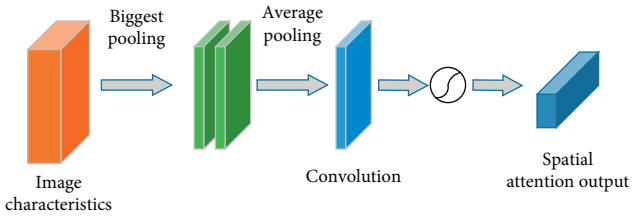


Figure 3: The spatial attention structure.

than a certain threshold, it will be removed, since there may be several high-scoring boxes in the same pedestrian, but only one is needed. The equation for NMS is as follows:

$$s_x = \begin{cases} s_x, & \mathrm{IoU}(N, i_x) < T, \\ 0, & \mathrm{IoU}(N, i_x) \geq T, \end{cases} \tag{5}$$

where N represents the box to be processed which is the boundary box with the highest score. The candidate window that overlaps with N is i_x. The score of the window whose overlap degree IoU is less than the threshold T is reserved; the score of the window whose overlap degree IoU is greater than the threshold T is set to 0.

2.2. Pedestrian Detection Algorithm Based on Deep Residual Network. In view of the existing problems in pedestrian detection, such as dense pedestrians, insufficient light, and wearing masks, this study uses ResNet-5 as the backbone network and attracts attention mechanism and NMS optimized detection method.

2.2.1. Improvement of Deep Residual Network. The network structure of the backbone network RESNET-50 is designed to obtain deeper image parameters, of which the network structure is shown in Figure 4. The input image is passed into the residual block after the first convolutional pooling. In each subsequent stage, Conv + Batch operation is required, namely, Conv Block, which then passes through multiple identity blocks with the same input and output dimensions. After the convolution from stage 2 to stage 5, the flatten layer is passed to compress the data into a one-dimensional array through a $7 * 7$ average pooling layer (AVG Pool), and then, it is connected with the full connection layer.

This study removed the full connection layer in RESNET-50. At the same time, two full-volume base

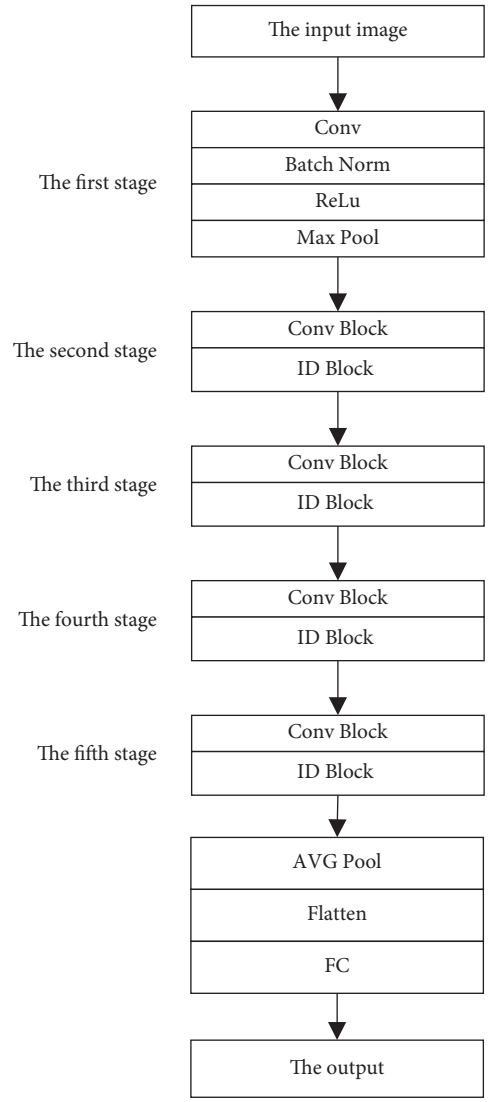


Figure 4: ResNet-50 network architecture.

branches were added to predict pixel-level boundary boxes and confidence scores, respectively, as shown in Figure 5.

As can be seen from Figure 5, a convolution layer is added at the end of resNET-50 phase 4. The step size is 1 and the kernel size is $512 * 3 * 3 * 1$. Then, linear interpolation is carried out to adjust the feature map to the original image size. Finally, a channel feature map with the same input image size is obtained after the element map is aligned with

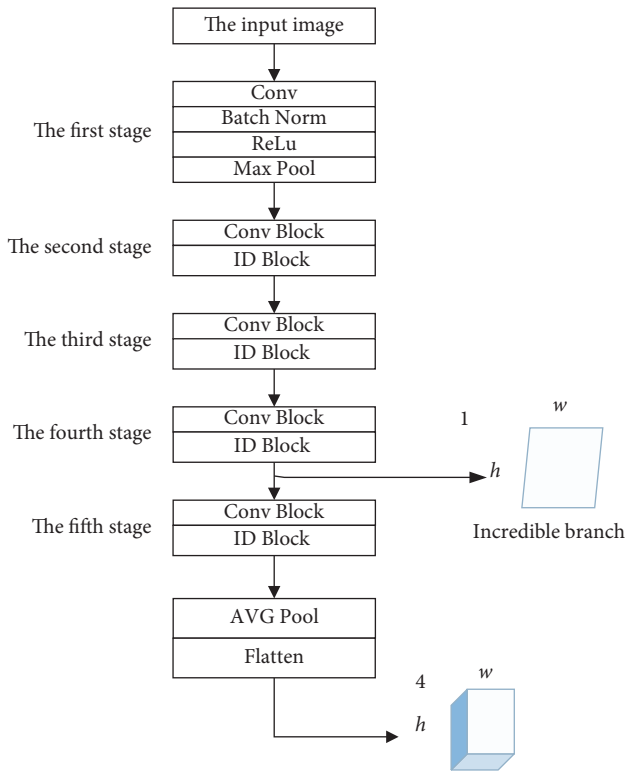


Figure 5: Confidence and prediction branches in deep residual networks.

the input image. An S-shaped cross entropy loss is used to generate a confidence heat map.

In order to predict the bounding box heat map, the convolution layer was added at the end of resNET-50 stage 5, whose core size was $512 * 3 * 3 * 4$. As in stage 4, the feature map is adjusted to the original image size and aligned with the input image. In addition, the ReLU [23] layer is inserted to ensure that the boundary box prediction is nonnegative, and the prediction boundary is optimized with the IoU loss function. The weighted average of the two branch losses is the final loss.

The confidence branch at the end of resNET-50 phase 4 is connected. Since the bounding box for IoU loss calculation is a whole, the bounding box branch is inserted at the end of stage 5. Therefore, a larger visual field of perception is needed, and the boundary box of the object can be predicted intuitively from the confidence heat map. In this way, bounding box branching is regarded as a top-down strategy that abstracts bounding boxes from confidence heat maps.

2.2.2. Introduction of Attention Mechanism. Attention mechanism in convolution block of network structure is introduced in this study. Given an intermediate feature graph, enter $I \in \mathbb{R}^{C\times H\times W}$. The trunk consists of two groups of residual units. The branch is composed of a group of residual units, channel attention module, and spatial attention module. The intermediate feature map first generates a one-dimensional channel attention diagram $M_C \in \mathbb{R}^{C\times 1\times 1}$ through channel attention module. Then, two-dimensional spatial attention force $M_s \in \mathbb{R}^{l\times H\times W}$ is generated through spatial attention module. In the figure, $\otimes$ is the multiplication of the corresponding matrix elements. When the channel attention modules are multiplied, the one-dimensional channel attention force is first expanded to $M_C \in \mathbb{R}^{C\times H\times W}$ and then multiplied. When multiplying the spatial attention modules, the two-dimensional spatial attention force is also expanded to $M_s \in \mathbb{R}^{C\times H\times W}$ along the channel dimension before multiplying.

The above process can be regarded as the combination of channel and spatial attention learning. The maximization of mutual information between levels is realized, which leads the model to learn more significant pedestrian-related information in iterative training.

2.2.3. Overall Network Structure. Figure 6 shows the overall network structure of pedestrian detection and applies the attention mechanism to the entire residual network. The useful information of images is made to flow effectively in the network, the useful information of key parts of pedestrians is captured, the detection ability of covered pedestrians is improved, and the pedestrians with confidence and heat maps of boundary boxes are accurately located. The ellipse was used to fit the pedestrian on the threshold confidence heat map. Since the pedestrian ellipse was too rough to locate the object, the center pixels of these pedestrian ellipses were further selected and corresponding boundary boxes were extracted from these selected pixels.

2.2.4. Nonmaximum Suppression Algorithm. The nonmaximum suppression algorithm (NMS) calculates the area of each detection frame and sorts it according to the score. Then, calculate the checking-union ratio between the remaining checking-union ratio and the checking-union ratio with the current maximum score and delete the checking-union ratio greater than the set threshold. Repeat the above process until the candidate detection frame is empty and finally get the best pedestrian detection frame.

3. Experimental Results and Analysis

This study studies the algorithm based on PyTorch deep learning framework. The deep learning platform is PyTorch0.4, the compilation environment is PyTorch4.0, and the operating system is Ubuntu18.04. The hardware platform is Dell T7810 workstation, with Intel E5-2620 (2.1 GHz) CPU, 16 GB memory, and Nvidia Quadro P4000 GPU. In this experimental environment, the detection speed of this study reaches 20 frame/s, which has a certain real-time performance.

In order to analyze and compare the algorithm in this study with other algorithms, PASCAL VOC2007, INRIA, and KAIST datasets are selected as experimental datasets. The VOC2007 dataset containing pedestrians was used as the training set. INRIA and positive sample images from INRIA dataset were used as test sets. The positive samples in the test set were used as validation sets.

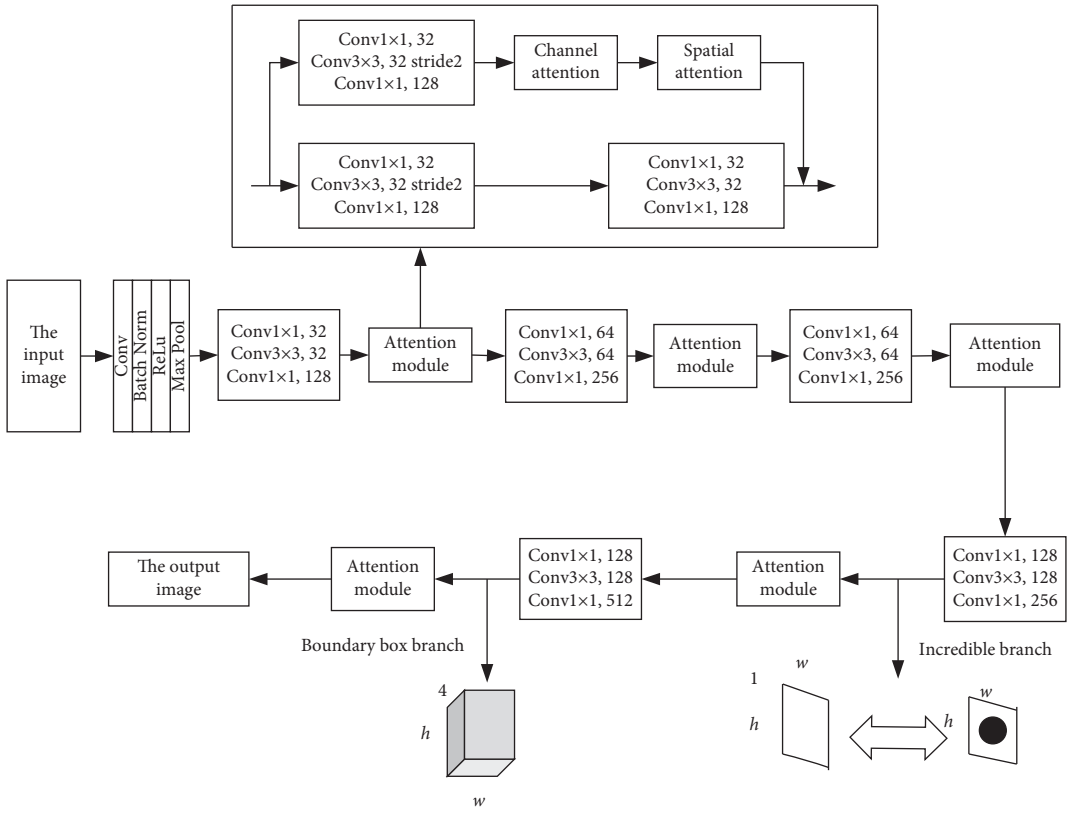


Figure 6: Network structure of deep residual network.

3.1. Experimental Parameter Setting. The initial learning rate was set as 0.001. With the increase of training rounds, the learning rate was reduced to 0.0001 to approach the optimal solution of the model. As the default anchor size of the algorithm in this study is obtained, it is not applicable to pedestrian detection task. Pedestrian targets are mostly long and narrow individuals. In this study, anchor points with dimensions of [48,157][34,104][84,50], [27,80][26,63] [25,40], and [18,54][16,44][13,24] were obtained by clustering algorithm as large, medium, and small pedestrian target detection frames.

3.2. Comparison of Experimental Results of Different Fusion Strategies. The experimental results of the accuracy comparison of different fusion strategies are shown in Table 1. Spinello and Siegwart [24] represent the direct fusion method commonly used by other multimodal pedestrian detection algorithms. In [25], an unimproved deep residual network detection method is represented. This study presents a pedestrian detection algorithm combining attention mechanism and nonmaximum suppression method.

As can be seen from Table 1, the proposed method combining deep residual network and attention mechanism in this study has great performance improvement, and attention mechanism is helpful for multimodal pedestrian detection task.

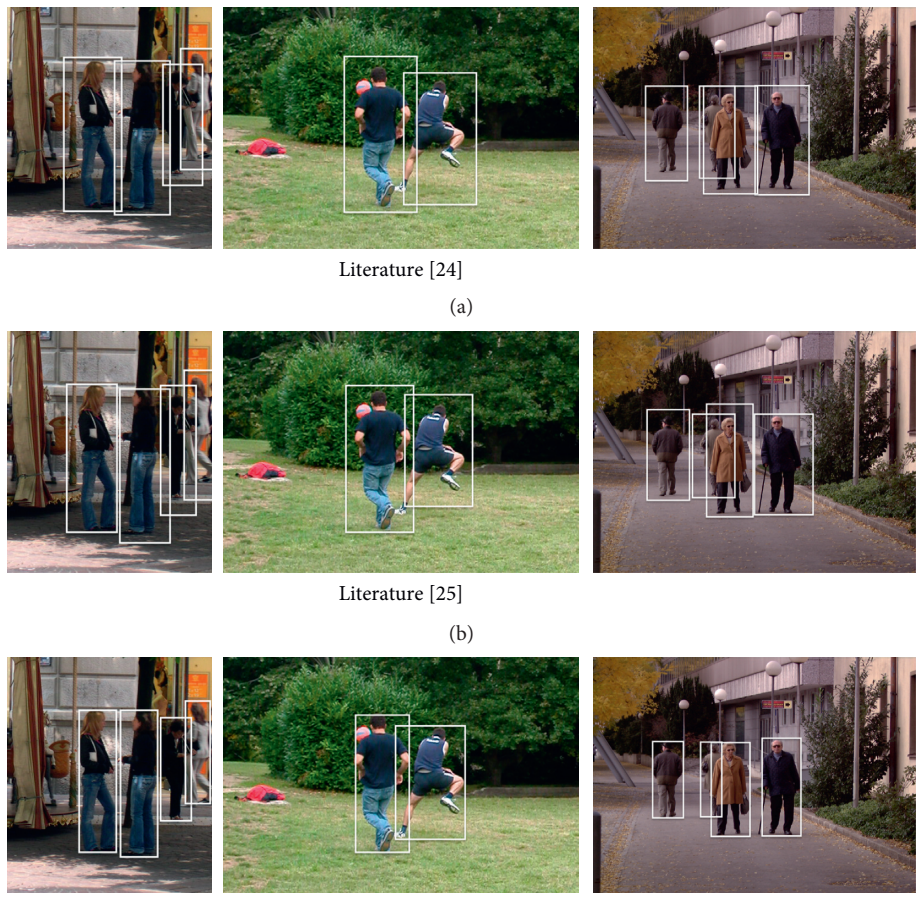
Part of the detection effect in the daytime is shown in Figure 7. The top row is the direct fusion detection results commonly used by other multimodal pedestrian detection algorithms, and the bottom row is the detection results of this algorithm. The rectangle in the figure is the detection result box, and the ellipse is the missed pedestrian target. The targets in Figure 7, which are difficult to detect due to their small size and mutual occlusion, are accurately detected. Experiments show that the proposed algorithm can improve the performance of pedestrian detector by integrating deep residual network and attention mechanism.

Table 1: The experimental results compared with different fusion modules.

Methods	All day	Day	Night
Literature [24]	82.67	83.29	82.13
Literature [25]	90.37	90.11	90.48
Proposed	92.58	93.26	91.49

The comparison between the proposed algorithm and other pedestrian detection algorithms based on the fusion of visible and infrared light is shown in Table 2, among which, the results of other comparison algorithms come from [26–28]. As can be seen from Table 2, the accuracy of the proposed algorithm is improved and compared with the comparison algorithm, and the speed of the proposed algorithm has a great advantage over algorithms with similar accuracy.

3.3. Comparison of Experimental Results with Single-Mode Pedestrian Detection Algorithm. Nguyen et al. [29] only use visible light for pedestrian detection algorithm. In the face of



Literature [24]

(a)

Literature [25]

(b)

Proposed

(c)

Figure 7: Comparison results during the day. (a) Literature [24]. (b) Literature [25]. (c) Proposed.

Table 2: Comparison of detection results of different algorithms.

Methods	AP (%)			Speed (frame/s)
	All day	The day	The night	
Literature [26]	72.42	75.75	63.79	31
Literature [27]	89.26	90.04	87.88	2.6
Literature [28]	90.14	90.83	87.91	1.9
Proposed	93.61	94.41	92.56	1.8

insufficient illumination, some scholars preprocess low-illumination images by means of exposure enhancement. McDonald et al. [30] use an algorithm for pedestrian detection after image exposure enhancement. The pedestrian detection algorithm combined with attention mechanism and nonmaximum suppression method in this study is compared with the above two algorithms, and the results are shown in Table 3.

As can be seen from Table 3, the visible light single-mode pedestrian detection method has advantages in speed. However, it is not as accurate as the algorithm in this study, especially at night, and its performance is poor. After the preprocessing of image exposure enhancement, the accuracy of the algorithm is improved, whereas, the accuracy is lower than that of the proposed algorithm.

Table 3: Performance comparison results of different algorithms.

Methods	AP (%)			Speed (frame/s)
	All day	The day	The night	
Literature [29]	73.62	80.21	59.88	35.60
Literature [30]	82.46	84.61	78.62	35.40
Proposed	92.63	93.4	91.57	20.00

4. Conclusion

With the further application of pedestrian detection technology, the accuracy and speed of pedestrian detection are required. In the real scene, the background load of pedestrian detection is changeable, and furthermore, various problems, such as insufficient illumination, target occlusion, and too small scale, often occur. Therefore, a pedestrian detection algorithm combining attention mechanism and nonmaximum suppression method is proposed in this study. Firstly, a deep residual network is introduced and an attention mechanism is added to the network, enhancing the expression ability of feature maps on the channel. In addition, the ability of context connection and feature description can be strengthened in feature map space via suppressing useless feature information. IoU loss function is

employed to optimize the performance of pedestrian detection, and the nonmaximum suppression method is added in the detection process to make the location more accurate. Experimental results illustrate that the proposed algorithm has higher accuracy and better detection effect in different environments such as insufficient illumination, target occlusion, and too small scale. In the future work, we will systematically analyze the performance of the algorithm and study it in the direction of industrial application.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Sichuan Technology & Business College.

References

[1] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: a survey," 2019, https://arxiv.org/abs/1905.05055.

[2] X. Ke, X. Lin, and L. Qin, "Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios," *Machine Vision and Applications*, vol. 32, no. 2, pp. 1–23, 2021.

[3] F. He, *A High-Fidelity VR Simulation Study: Do External Warnings Really Improve Pedestrian Safe Crossing Behavior?*, University of Waterloo, Canada, 2021.

[4] A. Zarkeshev, *Information Management Models and Methods for Innovative Transportation Systems and services*, Budapest University of Technology and Economics, Hungary, 2020.

[5] D. Geronimo, A. M. Lopez, and A. D. Sappa, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2009.

[6] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Q. Liu, "A comparative study of state-of-the-art deep learning algorithms for vehicle detection," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 82–95, 2019.

[7] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D Based People Detection and Tracking for mobile Robots and Head-Worn cameras," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5636–5643, Hong Kong, China, June 2014.

[8] P. Yang, G. Zhang, and L. Wang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1125–1137, 2020.

[9] Z. Deng, B. Wang, Y. Xu, T. Xu, C. Liu, and Z. Zhu, "Multi-scale convolutional neural network with time-cognition for multi-step short-term load forecasting," *IEEE Access*, vol. 7, pp. 88058–88071, 2019.

[10] X. Zhao, W. Li, and Y. Zhang, "A Faster RCNN-Based Pedestrian Detection system," in *Proceedings of the 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Montreal, QC, Canada, September 2016.

[11] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A Deep Neural Network Fusion Approach to Fast and Robust Pedestrian detection," in *Proceedings of the 2017 IEEE winter Conference on Applications of Computer Vision (WACV)*, pp. 953–961, Santa Rosa, CA, March 2017.

[12] A. Kumar and S. Srivastava, "Object detection system based on convolution neural networks using single shot multi-box detector," *Procedia Computer Science*, vol. 171, pp. 2610–2617, 2020.

[13] L. Neumann, M. Karg, and S. Zhang, "Nightowls: A Pedestrians at Night dataset," *Asian Conference on Computer Vision*, pp. 691–705, Springer, Cham, 2018.

[14] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9557–9566, Seoul, Korea (South), November 2019.

[15] K. He, G. Gkioxari, and P. Dollár, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Cambridge, MA, USA, June 2017.

[16] F. Sandelin, *Semantic and Instance Segmentation of Room Features in Floor Plans Using Mask R-CNN*, Uppsala universitet, Sweden, 2019.

[17] Y. Zhao, H. Xu, T. Yang, S. Wang, and D. Sun, "A hybrid recognition model of microseismic signals for underground mining based on CNN and LSTM networks," *Geomatics, Natural Hazards and Risk*, vol. 12, no. 1, pp. 2803–2834, 2021.

[18] C. Li, D. Song, and R. Tong, "Multispectral Pedestrian Detection via Simultaneous Detection and segmentation," 2018, https://arxiv.org/abs/1808.04818.

[19] V. P. Zelenov, S. S. Bukalov, and A. N. Subbotin, "A new type of the dinitrogen pentoxide-acid interaction," *Mendeleev Communications*, vol. 27, no. 4, pp. 355-356, 2017.

[20] Z. Allen-Zhu and Y. Li, "What can ResNet learn efficiently, going beyond kernels," 2019, https://arxiv.org/abs/1905.10337.

[21] S. Tao, Y. Li, Y. Huang, and X. Lan, "Face detection algorithm based on deep residual network," *Journal of Physics: Conference Series*, vol. 1802, no. 3, Article ID 032142, 2021.

[22] B. Liang, Q. Liu, and J. Xu, "Target-specific sentiment analysis based on multi-attention convolutional neural network," *Computer Research and Development*, vol. 54, pp. 172–1735, 2017.

[23] A. F. Agarap, "Deep Learning Using Rectified Linear Units (relu)," 2018, https://arxiv.org/abs/1803.08375.

[24] L. Spinello and R. Siegwart, "Human detection using multimodal and multidimensional features," in *Proceedings of the 2008 IEEE International Conference on Robotics and Automation*, pp. 3264–3269, Pasadena, CA, USA, May 2008.

[25] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 187, Article ID 104964, 2020.

[26] S. Hwang, J. Park, and N. Kim, "Multispectral Pedestrian Detection: Benchmark Dataset and baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1037–1045, Boston, MA, USA, June 2015.

[27] J. Liu, S. Zhang, and S. Wang, "Multispectral Deep Neural Networks for Pedestrian detection," 2016, https://arxiv.org/abs/1611.02644.

[28] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

[29] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[30] R. J. McDonald, J. S. McDonald, D. F. Kallmes et al., "Gadolinium deposition in human brain tissues after contrast-enhanced MR imaging in adult patients without intracranial abnormalities," *Radiology*, vol. 285, no. 2, pp. 546–554, 2017.