ITC 4/52
Information Technology and Control
Vol. 52 / No. 4 / 2023
pp. 1045-1057
DOI 10.5755/j01.itc.52.4.33766

Research on Pedestrian Detection Based on Multimodal Information Fusion

Received 2023/04/03 | Accepted after revision 2023/05/25

HOW TO CITE: Yang, X., Li, Z., Liu, Y., Huang, R., Tan, K., Huang, L. (2023). Research on Pedestrian Detection Based on Multimodal Information Fusion. *Information Technology and Control*, 52(4), 1045-1057. https://doi.org/10.5755/j01.itc.52.4.33766

Research on Pedestrian Detection Based on Multimodal Information Fusion

Xiaoping Yang, Zhehong Li
School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi, 541004, China
Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, Guangxi, 541004, China

Yuan Liu
College of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, Guangxi, 541004, China

Ran Huang, Kai Tan, Lin Huang
School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi, 541004, China
Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, Guangxi, 541004, China

Corresponding author: liuyuan@glmc.edu.cn

The automatic driving system based on a single-mode sensor is susceptible to the external environment in pedestrian detection. This paper proposes a fusion of light and thermal infrared multimodal pedestrian detection methodology. Firstly, 1 × 1 convolution and dilated convolution square measure are introduced within the residual network, and also the ROIAlign methodology is employed to exchange the ROIPooling methodology to map the candidate box to the feature layer to optimize the Faster R-CNN. Secondly, the generalized intersection over union (GIoU) loss function is employed as the loss function of prediction box positioning regression. Finally, to explore the performance of multimodal image pedestrian detection methods in different fusion periods in the improved Faster R-CNN, four forms of multimodal neural network structures are designed to fuse visible and thermal infrared pictures. Experimental results show that the proposed algorithm performs better on the KAIST dataset than current mainstream detection algorithms. Compared to the conventional ACF + T + THOG pedestrian detector, the AP is 8.38 percentage points greater. The miss rate is 5.34 percentage points lower than the visible light pedestrian detector.

KEYWORDS: Multimodal Pedestrian Detection; Faster R-CNN; Generalized Intersection Over Union; Feature Fusion.

1. Introduction

With the progress of computer vision technology, pedestrian detection plays a crucial role within the field of target detection, which is widely utilized in automatic driving, intelligent monitoring, public security, robots, video image analysis, and other fields [22]. The early pedestrian detection algorithm commonly uses the single-mode pedestrian detection method. In this method, the pedestrian target has a variety of transformations of its posture and scale. It is susceptible to external environmental factors such as light, occlusion, weather, and sophisticated background [26]. Therefore, researchers specialize in multimodal pedestrian detection, multimodal pedestrian detection [3] is a pedestrian detection technology that combines visible light and thermal infrared images. Its purpose is to simultaneously aggregate the detailed texture features of visible light images and the thermal radiation features of thermal infrared images and improve the visible light pedestrian detector in dim light, rain and snow, haze, and other weather conditions. The poor imaging effect of the image makes up for the defects of the thermal infrared image pedestrian detectors [28] with fewer texture details, low signal-to-noise ratio, and significant influence by background information and obtain complementary information of the target of interest in order to obtain more robust and accurate pedestrian detection results. Research on multimodal pedestrian detection techniques will ensure pedestrians' safety, accelerate the comprehensive landing of driverless car products, and assist the development of smart cities.

At present, deep learning has made significant progress in target detection tasks. A few researchers have begun applying deep learning techniques to detect pedestrians, particularly when using a deep convolutional neural network (CNN) [6]. In the target detection algorithm based on CNN, the Single Shot MultiBox Detector (SSD) [16] and You Only Look Once (YOLO) [22] algorithms describe a one-stage target detection approach. The prominent characteristic of the YOLO algorithm is that the original image is directly divided into grids and location regression and classification. The algorithm is fast, but the positioning accuracy is low, especially in small target detection. SSD algorithm performs regression and prediction based on feature maps of different scales, effectively improving targets' detection ability at different scales. However, the approach needs to address the issue of highly unbalanced positive and negative samples, making model training more challenging [5]. The other algorithm is a two-stage detection method represented by the Faster Region Convolutional Neural Network (Faster R-CNN) algorithm [23]. The goal is to filter the candidate box using a heuristic approach (e.g., selective search) or a Region Proposal Network (RPN), then determine whether the candidate box is inside the target box. Corrected is the goal classification or location. Faster R-CNN, in contrast to SSD and YOLO algorithms, first trains the RPN network using the extracted basic features so that it can generally distinguish between foreground and background and then uses the detection network to produce precise detection boxes. Faster R-CNN cuts through the candidate box detection bottleneck and integrates the entire algorithmic process of region formation, feature extraction, network training, target classification, and location regression into a complete end-to-end learning framework.

Although the visible and thermal infrared image fusion-based multimodal pedestrian detection technology can be designed to combine the characteristics, benefits, and technical advantages of conventional visible and thermal infrared images simultaneously, whether the fused image can weaken the surrounding background or strengthen the characteristics of pedestrians, the choice of image feature fusion method and period has a significant impact on the detection effect of the pedestrian detector. The existing visible and thermal infrared pedestrian detection research does not have a standard fusion method, which is still in the exploratory stage.

In summary, this paper constructs a multimodal pedestrian detection network based on Faster-RCNN and improves it to explore multimodal pedestrian detectors. The main contributions are as follows:

1 The feature extraction network is optimized, and structures such as 1 × 1 convolution and dilated convolution are introduced to enhance the expression of the network feature layer. In addition, the ROIAlign method replaces the ROIPooling method to map the candidate box to the feature layer to

eliminate the quantitative loss and improve the detection ability of small target pedestrians.

2 By optimizing the loss function and introducing GIOU into the loss function [24], the prediction frame moves to the target frame to improve the positioning accuracy and the target detection accuracy.

3 Four neural network structures fusing visible and thermal infrared images are designed to explore the performance of multimodal image pedestrian detection methods in different fusion periods in the improved Faster R-CNN. Finally, it was tested on the KAIST dataset and found that the pixel fusion pedestrian detector has a better missed detection rate and accuracy than other pedestrian detectors.

2. Correlation Study

2.1. Pedestrian Detection Based on Visible Light Image

Many different strategies have been put out in recent years to improve the performance of visible light pedestrian detectors, primarily to address the issues of pedestrian occlusion, congestion, and scale difference. Angelov et al. [1] use a cascade classifier to increase the accuracy of deep neural networks. The disadvantage is that time consumption will increase as the image size increases ALFNet was proposed by Liu et al. [15], expanding the SSD target identification method with cascade and multistage ideas. The cascade structure is employed to increase IoU (intersection-over-union) continually. While improving the accuracy of pedestrian detection, it can also enjoy the speed of SSD. Xie et al. [26] extracted the image foreground by the Gaussian mixture model to avoid the interference of complex backgrounds. Huang et al. [9] proposed a Region NMS technique effectively eliminates superfluous frames without producing a significant amount of false positives by using the viewable region with reduced occlusion to tackle the pedestrian detection problem in crowded settings. Zhang et al. proposed the AR-CNN algorithm [27]. By enhancing the secondary classification operation of the target candidate box and the loss function of the Faster R-CNN algorithm, the detection effect of occluded pedestrians is increased. From the mask-guided attention mechanism perspective, Pang et al. proposed the MGAN algorithm (Mask-guided Attention Network) [21] to increase the detection effect of occluded pedestrians. Lin et al. [13] incorporated multiscale pedestrian attention mechanisms to enhance the ability of the network to recognize small-scale pedestrian targets and employed convolution layers with various resolutions and receptive fields for detection. Pang et al. [20] suggested a multiscale MCF network based on JCS-Net to improve the knowledge of tiny target pedestrians. With the concept of a multi-stage progressive localization mechanism, Liu et al. presented the ALFNet method [15], which increased the detection effect of multiscale pedestrians.

2.2. Pedestrian Detection Based on Multimodal Images

In order to effectively remove the pedestrian shadow region, Choi et al. [4] devised a combined bilateral filtering technique that included the edge information of the visible light picture with the white space of the thermal image. Hwang et al. proposed the ACF + T + THOG algorithm, which is additionally ofttimes applied as a Baseline algorithm, using a massive public multimodal dataset (KAIST) [10]. Two modal features were extracted using the ACF feature extraction methodology, multiple modal features were fused using the feature cascade approach, and targets were classified using the improved boosting decision tree (BDT). Wagner et al. [17] created two decision networks (early-fusion and late-fusion) based on CaffeeNet and applied DNN to multimodal pedestrian identification for the first time. Experimental results on the KAIST dataset show that late-fusion performance is significantly better than early-fusion.

Based on Faster R-CNN at various DNN stages, Liu et al. [14] created four bi-branch convolutional neural network fusion architectures. The Halfway Fusion construction is the best. Faster RCNN-C and Faster RCNN-T are two separate pedestrian detectors that were trained separately. It was discovered that the two detectors provided complementary information while identifying human occurrences, proving the rationality and necessity of multimodal image fusion detection for pedestrians. Konig et al. constructed Fusion RPN + BDT fusion network by fusing dual branch RPNs on middle convolution features [12]. To fill the gap in pixel-level picture fusion, Hou [8] developed

a technique based on the SSD framework. In place of directly predicting the border box, Alexander [18] suggested an anchorless small-scale multimodal pedestrian identification approach that learns pedestrian representation based on object center and scale. The outcomes demonstrate the efficacy of the method for small-scale pedestrian detection.

3. Proposed Method

3.1. Design of Pedestrian Detection Network Based on Faster R-CNN

In order to enable the detection algorithm to locate and detect pedestrian targets quickly and effectively, this paper uses ResNet (Deep Residual Network) [11] to replace the traditional VGG16 (Visual Geometry Group Network) [2] as the shared convolution layer to obtain more abundant semantic information. It introduces a 1×1 convolution design in the residual network structure, reducing the dimension of the feature map and the model parameters. The dilated convolution design is introduced to obtain more dense and expressive features in the same receptive field and enhance the capacity on the network to recognize small target pedestrians. In addition, the ROIAlign method replaces the ROIPooling method to map the candidate box to the feature layer to eliminate the quantitative loss. Finally, the design and optimization of the loss function in this paper are proposed so that the network can better regress the candidate boxes to increase pedestrian detection accuracy and reduce the missed detection rate.

3.1.1. Faster R-CNN Network Structure

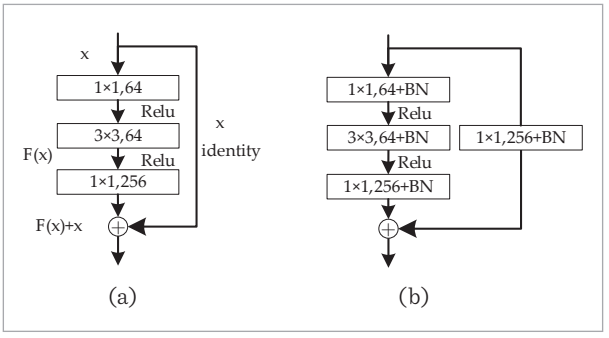
The feature extraction network (Backbone), Region Proposal Network (RPN), and detection network comprise the Faster R-CNN network. The feature extraction network completes the feature extraction of the input image to obtain the feature map, which can be replaced by basic feature extraction networks such as ResNet and VGG. The candidate box extraction network generates a high-quality target candidate box (region proposal) on the feature map; finally, the feature map and the generated candidate boxes are sent to the detection network and processed into fixed-size feature vectors by ROIPooling operation. Lastly, the target classification and boundary box regression are realized through full connection.

3.1.2. Improved Residual Learning Module

The residual learning module is a ResNet network structure model proposed by Dr. He in 2015 as a solution to the issue of gradient information disappearance or gradient explosion in deep networks. The residual idea is to add the original input feature x to F (x) after convolution, pooling, and nonlinear activation functions. It is continuously learning new features without sacrificing performance to enhance network performance. Figure 1(a) shows the structure of the residual learning module.

Figure 1

Improved residual learning unit. (a) Residual Learning Unit of ResNet50 networks; (b) Residual Learning Unit with 1×1 Convolution



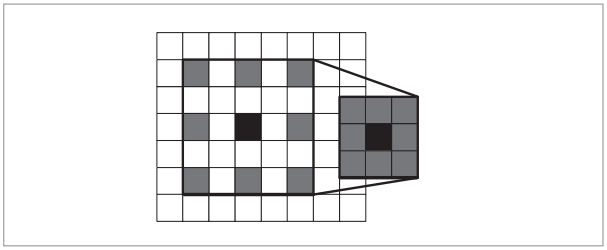
By introducing nonlinear activation functions (e.g., relu), 1 × 1 convolution strengthens the nonlinear characteristics of the network and enhances the ability of the network to express complex features. Controlling the number of convolution kernels reduces feature dimension, speeds up network operation, and improves training efficiency. With these benefits of the 1 × 1 convolution kernel in mind, this study introduces the convolution kernel and batch normalization (BN) approach to quicken the convergence of the ResNet50 network. Figure 1(b) shows the organizational structure of the improved residual learning module.

3.1.3. Introducing Dilated Convolution

Dilated convolutions inject holes into the standard convolution and introduce the dilation rate (dr) to control the size of the expansion of the convolution kernel so that the hole convolution operation in the feature layer has a larger field of vision than the general convolution operation without losing resolution [19], which is convenient for better aggregation of im-

age feature information. The convolution hole with a dilation rate of two is shown in Figure 2. In pedestrian detection tasks, detecting small target pedestrians from long distances is challenging. Dilated convolution can obtain denser and more expressive features in the same receptive field, which has more advantages for detecting small target pedestrians. With different expansion rates, the receptive field will be different. This paper captures multiscale information by adjusting the expansion rate parameters.

Figure 2
The Dilated Convolution of dr = 2



3.1.4. ROIAlign Replaces ROIPooling

The process of Faster R-CNN mapping the candidate box to the feature map is ROIPooling, which involves the transformation of the candidate frame size. The nearest neighbor interpolation is used, and the boundary is quantified as integer point coordinates, which loses a certain spatial accuracy, and causes the position deviation of the target to affect the subsequent regression and positioning of the target. The ROIAlign proposed in Mask RCNN cancels the quantization operation and obtains the pooled results with floating-point accuracy by bilinear interpolation to increase the detection accuracy of small targets.

3.2. Design of Multimodal Fusion Network

Based on the improved Faster R-CNN algorithm, this paper constructs Pixel Fusion Network, Early Fusion Network, Middle Fusion Network, and Late Fusion Network. The pixel fusion method is designed to retain the color information of the image better to improve the quality and clarity of the image. The design of the latter three fusions is to comprehensively utilize different feature expressions to better fuse image information, improve the accuracy and robustness of target detection, and further explore the influence of different fusion periods on the detection effect of pedestrian detectors.

Summarize the above four fusion network structure. Its characteristics are shown in Table 1.

Table 1
Features of four converged network architectures

Network structure	Fusion method	Integration phase
Pixel fusion network	Color space fusion	Input layer
Early fusion network	Feature map channel fusion	After Stage 0
Middle fusion network		After Stage 2
Late fusion network		After Stage 4

3.2.1. Pixel Level Fusion Network Structure

The pixel fusion network is based on the fusion of visible light and thermal infrared images at the pixel level as the input image without changing the Faster R-CNN network structure. Compared with the original network, only the image fusion process is added before the input layer. This paper applies the color space fusion method to achieve pixel fusion. The fusion method is shown in Figure 3. Firstly, the visible light image is converted from RGB color space to HSV color space, where H represents hue, S represents saturation, and V represents transparency. The raw thermal infrared image is then fused with the H channel of the visible light's HSV color space. The fused H channel replaces the original H channel in the HSV space. Finally, the image is transformed from HSV space to RGB color space by the color model transformation method to obtain the fused image. It can be seen from the figure that the visible light image with thermal

Figure 3
Fusion Process of Visible and Infrared Images

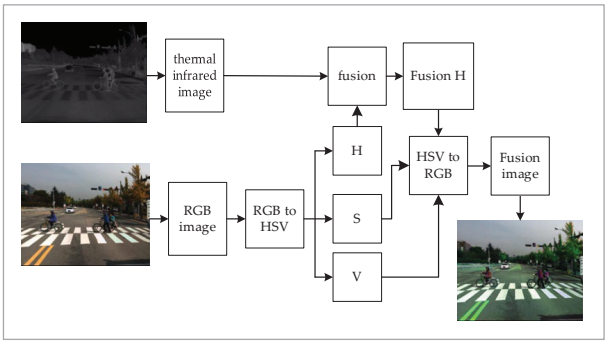
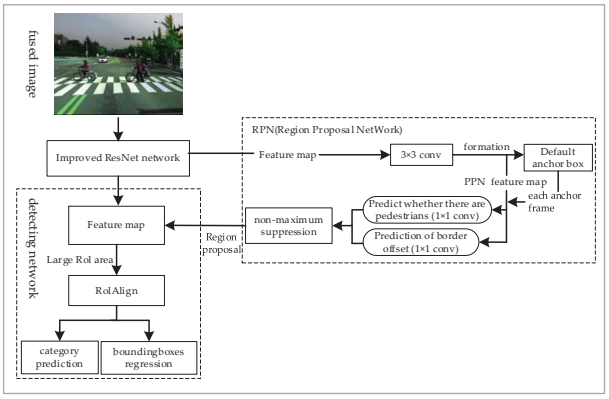


Figure 4
Input the fused image into Faster-RCNN



infrared image information becomes clearer in color, making pedestrians more identifiable.

The fused image is input into the Faster R-CNN network for pedestrian detection, as shown in Figure 4. Considering the classification and regression of small target pedestrians, ROIAlign replaces ROIPooling in the original Faster R-CNN network

3.2.2. Early Fusion Network Structure

The early fusion network structure is shown in Figure 5. At the same time, the visible light image and the thermal infrared image are input into the feature extraction network (e.g., ResNet50) for convolution to extract features. After the feature extraction of the early ResNet50 network, the feature layers corresponding to the two are obtained, respectively. The obtained feature layers are fused, that is, the superposition operation of the channels. The five blue boxes in the figure represent the convolutional layer, the five stages corresponding to the Backbone network in ResNet50 (Stage 0 to Stage 4). The fusion methods used in the early, middle, and late stages are all feature map channel fusion, identified by the green box in the fusion network structure diagram. The difference is that the fusion period is different. Early and middle fusion require size transformation of the fused feature maps, denoted by the yellow boxes, to meet the size requirements of subsequent network layers. Other parts of the algorithm remain unchanged.

3.2.3. Middle Fusion Network Structure

The middle fusion network structure is shown in Figure 6. The visible light and thermal infrared images

Figure 5
Different Early Fusion Network Structure

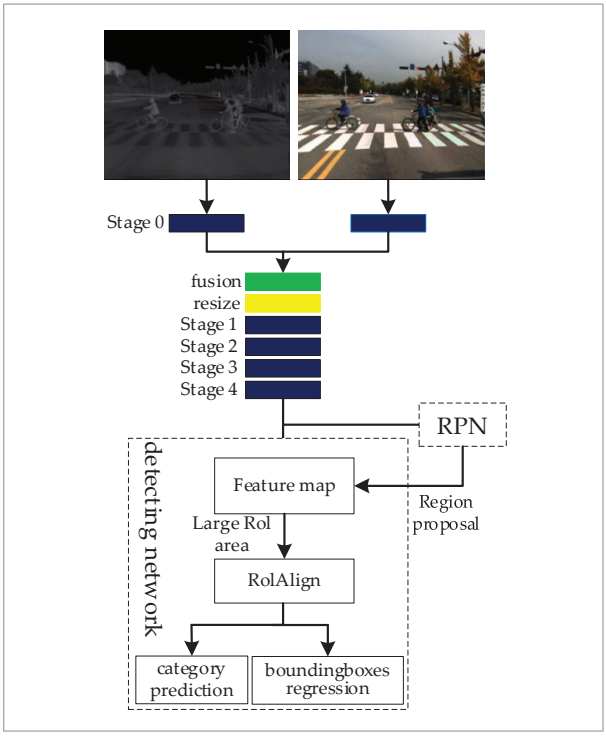
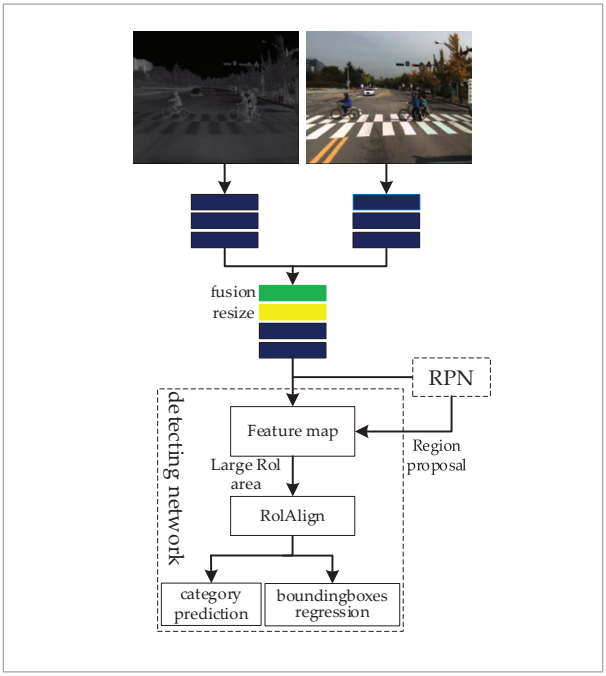


Figure 6
Middle Fusion Network Structure

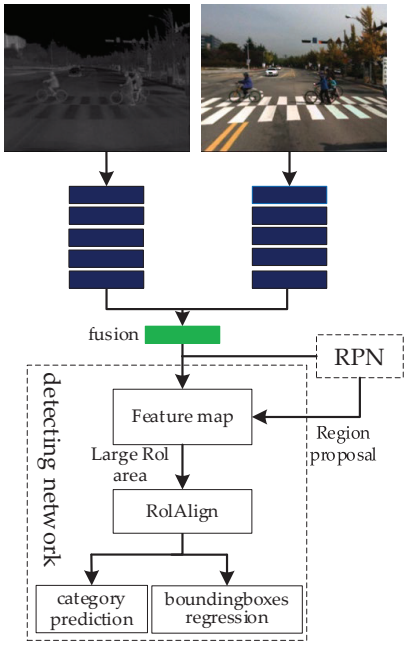


are, respectively, input into the ResNet50 for feature extraction. The features are fused in the middle of the feature extraction network (i.e., after Stage 2), and the fused feature map proceeds to subsequent stages for feature extraction.

3.2.4. The Late Fusion Network Structure

The later fusion network is similar to the middle stage. After the feature extraction of visible and thermal infrared images (i.e., Stage 4), the feature channel is fused to obtain the fused feature map and input it into the detection network. The network structure is shown in Figure 7.

Figure 7
Late Fusion Network Structure



3.3. Design of Loss Function

The Loss function of the Faster-RCNN RPN network has two components: the Loss of Target Detection and the Loss of Regression Prediction, as shown in Formula 1.

$$L(\{p_i\},\{t_i\})=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i L_{reg}\left(t_i,t_i^*\right) \quad (1)$$

Among them, i represents the i detection box, is the probability of targeting anchor prediction, p_i^* is the label of anchor corresponding to Ground Truth

$$\text{GTlabel}: p_i^*=\begin{cases}0 & Negative\ samples\left(background\right)\\ 1 & Positive\ samples\left(pedestrians\right)\end{cases}.$$

$t_i=\{t_x,t_y,t_w,t_h\}$, represents the four parameterized coordinates corresponding to the detection box, namely the normalized offset and scaling scale between the anchor and the detection box, and t_i^* represents the normalized offset and scaling scale between the anchor and the truth value.

The classification tasks use cross-entropy loss and the calculation of $L_{cls}\left(p_i,p_i^*\right)$ as shown in Formula 2.

$$L_{cls}\left(p_i,p_i^*\right)=-\log\left[p_ip_i^*+\left(1-p_i\right)\left(1-p_i^*\right)\right]. \quad (2)$$

When p_i^* is 0, the regression loss is 0, and when p_i^* is 1, the return loss needs to be considered. The calculation formula of regression loss, $L_{reg}\left(t_i,t_i^*\right)$ in Formula 1 is shown in Formula 3.

$$L_{reg}\left(t_i,t_i^*\right)=R\left(t_i-t_i^*\right), \quad (3)$$

where R is:

$$smooth_{L1}\left(x\right)=\begin{cases}0.5x^2 & if\ |x|<1\\ |x|-0.5 & otherwise\end{cases} \quad (4)$$

x represents the value of $t_i-t_i^*$.

Aiming at the problems of complex environmental changes, rain and snow weather, and small target pedestrian detection in pedestrian detection tasks, GIoU is used as the loss function of Faster to increase the accuracy of further target positioning R-CNN prediction box positioning regression.

GIoU introduces penalty terms based on IoU. According to the GIoU principle, $GIoU\le IoU$ and $0\le IoU\le 1$ so $0\le IoU\le 1$. In addition, when the C-frame is introduced, and A and B are not intersected, $GIoU=-1+\frac{A\cup B}{C}$, it is vital to maximizing GIoU to reduce the loss value, which then calls for the C-frame to be minimum or maximum so that A and B are always close. To effectively improve the target positioning accuracy. The improved Faster R-CNN total loss function formula is as follows:

$$L\left(\{p_i\},\{t_i\}\right)=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i p_i^*\left(1-GIoU_{prdict(t_i)}^{groundtruth\left(t_i^*\right)}\right) \quad (5)$$

4. Experimental Results and Analysis

4.1. Datasets and Preprocessing

The experimental portion of this study uses the public KAIST multimodal pedestrian dataset. The entire original KAIST pedestrian dataset consists of 45 168 pairs of images. Each pair of images contains two modal information of RGB color image and thermal infrared image, and 103 128 dense annotations [10]. The original data set is extracted every ten frames to produce the experiment's reduced data set. Finally, 2000 pairs of multimodal images make up the test set, compared to 2516 pairs in the training set.

4.2. Parameter Setting of the Experimental Environment Set

This experiment uses PyTorch open source framework to implement training based on GeForce GTX 1660Ti GPU. During training, conv1 and conv2 parameters are fixed, and the parameters of the remaining convolution layer need to be adjusted according to the backpropagation algorithm. The experimental model consisted of 150 epochs, where the initial 50 epoch learning rate was set to $1e^{-4}$, each epoch decayed by 0.95. The learning rate of the subsequent 100 epochs began from $1e^{-5}$, and each epoch decayed by 0.95.

4.3. Evaluation Criterion

Currently, the commonly used evaluation indexes of pedestrian detection include average precision (AP), mean average precision (mAP), frame per second (FPS), Precision, and Recall. Miss Rate is the loss rate, representing the proportion of undetected pedestrians within the total variety of pedestrians. FPPI represents the number of pedestrians, which will be correctly retrieved in each graph. Following are the calculation formulas of Precision, Recall, Miss Rate, AP, mAP, Miss Rate, and FPPI:

$$Precision=\frac{TP}{TP+FP} \quad (6)$$

$$Rrecall=\frac{TP}{TP+FN} \quad (7)$$

$$AP=\frac{p_1+p_2+\cdots+p_n}{n} \quad (8)$$

$$mAP=\frac{\sum_{i=0}^{k}AP_i}{k}=\frac{AP_1+AP_2+\cdots+AP_k}{k} \quad (9)$$

$$Miss\ Rate=1-Recall \quad (10)$$

$$FPPI=\frac{FP}{the\ number\ of\ image}. \quad (11)$$

The performance evaluation indices are the AP, Miss Rate-FPPI curve, and log-average miss rate. The area of the curve formed by exactitude as the vertical axis and recall as the horizontal axis is known as AP. The longitudinal axis of the Miss Rate-FPPI curve is the miss rate, and the horizontal axis is FPPI. The lower the Miss Rate-FPPI curve, the better the effect. The log-average miss rate is similar to mAP. The vertical axis of the curve is the log of miss rate, and the horizontal axis is FPPI. The smaller the index value, the higher the detector performance.

4.4. Ablation Experiments

Based on the visible light and thermal infrared datasets selected, this paper used ResNet50 as the backbone network. Ablation experiments compared traditional Faster R-CNN and improved Faster R-CNN algorithms. Table 2 shows that mAP increased by 2.53 percentage points after improving the backbone feature extraction network by introducing 1×1 convolution and dilated convolution. Selecting ROIAlign's pooling operation increased mAP by 3.5 percentage points. Improving the loss function increased mAP by 3.49 percentage points. Combining all improvements increased mAP by 4.55 percentage points.

Table 2
Ablation analysis experimental

1×1 convolution	dilated convolution	ROIAlign method	GIOU loss	V-AP	T-AP	mAP
				76.87	75.25	76.06
√				78.58	75.83	77.20
	√			79.82	75.16	77.49
√	√			81.68	75.50	78.59
		√		80.44	78.69	79.56
			√	80.88	78.23	79.55
√	√	√	√	82.03	79.13	80.58

Note: V-AP represents the AP value corresponding to the visible pedestrian detector, and T-AP represents the AP value corresponding to the thermal infrared pedestrian detector.

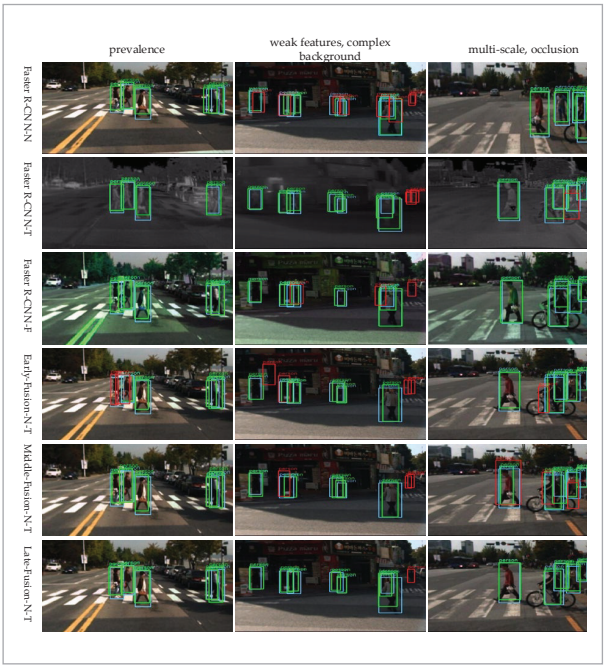
4.5. Design of Pedestrian Detection Network Based on Faster R-CNN

In this paper, six groups of experiments are designed to explore and study the pedestrian detection problem of visible and thermal infrared images, namely, the visible light image pedestrian detector (Faster R-CNN-N). Thermal infrared image pedestrian detector (Faster R-CNN-T), pixel fusion pedestrian detector (Faster R-CNN-F), early fusion pedestrian detector (Early-Fusion-N-T), middle fusion pedestrian detector (Middle-Fusion-N-T), late fusion pedestrian detector (Late-Fusion-N-T).

After the training, the model parameters are obtained, and the test set is tested. Figure 6 displays the pedestrian detection results of six pedestrian detectors. The detector accurately identifies the forecast as a pedestrian in the green box, which indicates this. The detector marks the prediction as inaccurate despite the red box indicating it is pedestrian. The blue box shows Ground Truth.

Figure 8 displays several pedestrian detection outcomes from the pixel fusion pedestrian detector, the visible pedestrian detector, and the thermal infrared pedestrian detector. There was a false detection in the visible image pedestrian detector and the thermal infrared pedestrian detector in the first line of images. False detection was removed from the pedestrian detector of the pixel fusion image. In the second line of images, there were two false detections in the thermal

Figure 8
Detection results of different pedestrian detectors (the green box represents TP, the red box represents FP, and the blue box represents ground truth)



infrared pedestrian detector but none in the visible pedestrian detector or the pedestrian detector of the pixel fusion image. The experimental results demonstrate that the pixel fusion visible light image and the thermal infrared image have the advantages of com-

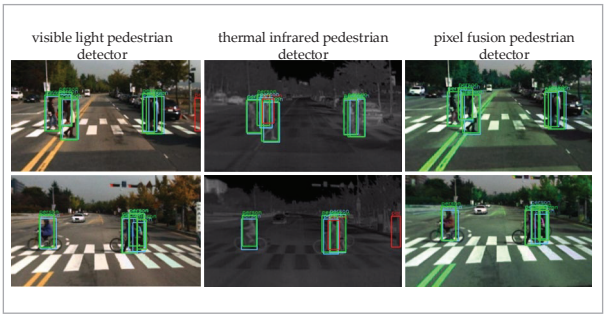
plementary information, which makes the pixel fusion pedestrian detector have excellent performance. To a certain extent, the pixel fusion pedestrian detector can eliminate the false pedestrians predicted by the visible light pedestrian detector and the thermal infrared pedestrian detector.

Figure 9 displays several pedestrian detection outcomes from the pixel fusion pedestrian detector, the visible pedestrian detector, and the thermal infrared pedestrian detector. There was a false detection in the visible image pedestrian detector and the thermal infrared pedestrian detector in the first line of images. False detection was removed from the pedestrian detector of the pixel fusion image. In the second line of images, there were two false detections in the thermal infrared pedestrian detector but none in the visible pedestrian detector or the pedestrian detector of the pixel fusion image. The experimental results demonstrate that the pixel fusion visible light image and the thermal infrared image have the advantages of complementary information, which makes the pixel fusion pedestrian detector have excellent performance. To a certain extent, the pixel fusion pedestrian detector can eliminate the false pedestrians predicted by the visible light pedestrian detector and the thermal infrared pedestrian detector.

The experimental results of the ACF + T + THOG pedestrian detector, FasterRCNN pedestrian detector, and the six pedestrian detectors designed in this paper on the KAIST dataset are shown in Table 3. The AP of the traditional FasterRCNN, the visible light,

Figure 9

Comparison of pedestrian detection results of pixel fusion pedestrian detector, visible pedestrian detector, and thermal infrared pedestrian detector



thermal infrared, and pixel fusion algorithm based on the improved FasterRCNN is superior to the ACF + T + THOG pedestrian detection algorithm based on traditional machine learning, which verifies that convolutional neural network has unique advantages in the field of target detection. Among the six pedestrian detectors, the pixel fusion pedestrian detector is superior to others in AP and miss rate. It is superior to the pedestrian detector of the fusion network in detection speed and model size, achieving the best pedestrian detection effect. The AP of the pixel fusion pedestrian detector is 8.38 percentage points higher than that of the ACF + T + THOG pedestrian detector, and the miss rate is 5.34 percentage points lower than that of the visible light pedestrian detector Faster R-CNN-N, 10.52 percentage points lower than the Faster R-CNN-C pedestrian detector in reference

Table 3

Comparison of experimental results

Experiment	AP(%)	miss rate (%)	Log average miss rate(%)	Fps(f/s)	Model size (M)	MIoU/%
ACF+T+THOG [10]	74.86	64.46	/	32.00	/	/
Faster RCNN-C [14]	/	61.19	/	/	/	/
Faster R-CNN-N	82.03	56.01	48	5.68	108	52.4
Faster R-CNN-T	79.13	56.12	51	6.07	108	52.4
Faster R-CNN-F	83.24	50.67	47	5.53	108	52.4
Early-Fusion-N-T	66.11	70.94	63	4.29	108	52.4
Middle-Fusion-N-T	67.15	72.27	63	4.47	148	71.8
Late-Fusion-N-T	76.39	57.35	52	3.04	159	77.1

[14], and 13.79 percentage points lower than the ACF + T + THOG pedestrian detector.

The comparison of miss rate-fppi curves of six pedestrian detectors is shown in Figure 10. In the diagram, take the miss rate when the fppi equals 10^{-1}, where the pixel fusion pedestrian detector obtains the lowest miss rate. The best performance of the feature map channel fusion network detector is the late fusion network structure, and the miss rate is 13.59 percentage points and 14.92 percentage points lower than the early fusion pedestrian detector and the middle fusion pedestrian detector, respectively.

Figure 10

Comparison of miss-rate-fppi curves of different pedestrian detectors

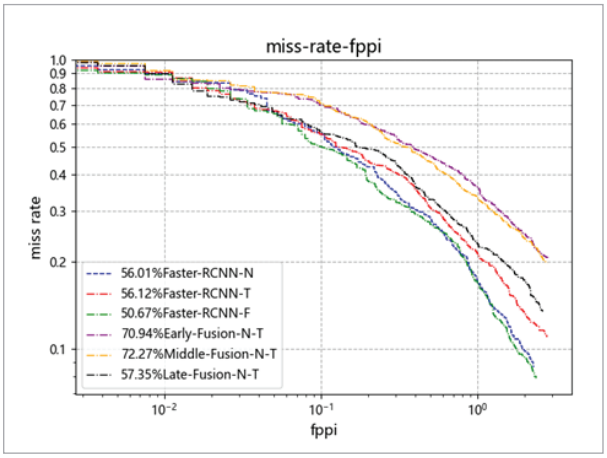
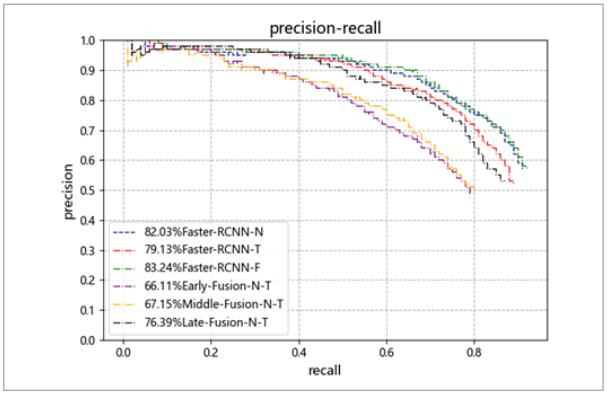


Figure 11 compares the precision-recall curves of six pedestrian detectors. The best AP is still obtained using the pixel fusion pedestrian detector. The AP of the late fusion pedestrian detector in the fusion network is 10.28 percentage points and 9.24 percentage points higher than that of the early and middle stages, respectively.

Of the six experimental groups, the pixel fusion pedestrian detector has the lowest missed detection rate and the most excellent accuracy for detecting pedestrians. Compared to a single visible picture or thermal infrared image, the pixel fusion image of visible and thermal infrared images provides complementary information. The latter three pedestrian detectors based on feature map channel fusion are inferior to others in detection accuracy and missed detection rate. The possible reasons are as follows. Firstly, the fusion network intensifies the complexity of pedestrian features but does not extract more effective pedestrian features. Secondly, the fusion network strengthens the characteristics of the surrounding background while strengthening pedestrian features, making it challenging to identify pedestrian targets. Because the fusion network is a dual-branch network, the model parameters are more than the single-branch network, which makes the FPS of the fusion network pedestrian detector lower.

Figure 11

Comparison of precision-recall curves of different pedestrian detectors



5. Conclusions

Four neural network structures for fusing visible and thermal infrared pictures are developed based on the improved Faster R-CNN, and the experimental study is done using the public KAIST multimodal pedestrian data set. It is discovered that the pixel fusion pedestrian detector outperforms existing pedestrian detectors in terms of missed detection rate and detection accuracy. The later fusion pedestrian detector achieves the best pedestrian detection performance among the latter three pedestrian detectors based on feature map channel fusion. It demonstrates that the pedestrian detector combined with thermal infrared and visible images has certain benefits over the pedestrian detector of a single image, which has some reference significance for the ensuing study on the multimodal pedestrian detector.

Author Contributions

Conceptualization, Y. X. P and L. Z. H.; methodology, Y. X. P and L. Z. H.; software, L. Z. H and H. R.; valida-

tion, Y. X. P., L. Z. H., H. R. and T. K.; formal analysis, L. Z. H; investigation, Y. X. P, L.Z.H. and H. R.; resources, Y. X. P., L. Y. and H. L.; data curation, Y. X. P.; writing—original draft, Y. X. P., L. Z. H. and H. R.; writing—review and editing, Y. X. P. and L. Z. H.; visualization, Y. X. P. and L. Z. H; supervision, L.Y. and H.L.; project administration, L. Y.; funding acquisition, Y. X. P., L. Y. and H. L. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the Doctoral Research Start-up Fund of Guilin Medical College under Grant No.31304019011, in part by the National Natural Science Foundation of China under Grant No. 62166012, in part by the National High-tech R & D Program of China under Grant No. 2013AA12210504, in part by the Key Scientific and Technological Projects in Guangxi Zhuang Autonomous Region under Grant No. AC1638012 and No.AD18281068, and in part by the Science and Technology Plan of Qingxiu District Science and Technology Bureau of Nanning City, Guangxi Zhuang Autonomous Region under Grant No. RZ19100041.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., Ferguson, D. Real-Time Pedestrian Detection with Deep Network Cascades. British Machine Vision Conference, 2015, 32, 1-12. https://doi.org/10.5244/C.29.32
2. Avianto, D., Harjoko, A., Afiahayati. CNN-Based Classification for Highly Similar Vehicle Model Using Multi-Task Learning. Journal of Imaging, 2022, 8(11), 293. https://doi.org/10.3390/jimaging8110293
3. Chen, Y., Zhu, Y. Multimodal Pedestrian Detection Network Based on Modal Adaptive Weight Learning Mechanism. Optical Precision Engineering, 2020, 28(12), 2700-2709. https://doi.org/10.37188/OPE.20202812.2700
4. Choi, E. J., Park, D. J. Human Detection Using Image Fusion of Thermal and Visible Image with New Joint Bilateral Filter. 5th International Conference on Computer Sciences and Convergence Information Technology, 2010, 882-885. https://doi.org/10.1109/ICCIT.2010.5711182
5. Choi, H. T., Lee, H. J., Kang, H., Yu, S., Park, H. H. SSD-EMB: An Improved SSD Using Enhanced Feature Map Block for Object Detection. Sensors, 2021, 21(8), 2842. https://doi.org/10.3390/s21082842
6. Ding, L., Wang, Y., Laganière, R., Huang, D., Zhang, H. A Robust and Fast Multispectral Pedestrian Detection Deep Network. Knowledge-Based Systems, 2021, 227(6), 73-85. https://doi.org/10.1016/j.knosys.2021.106990
7. Feng, Y. P., Guan, Y. Y., Yang, X. R., Liu N., Wang Z. H. Real-time Pedestrian Detection Algorithm Based on Attention Mechanism. Electronic Measurement Technology, 2021, 44(17), 123-130.
8. Hou, Y. L, Song, Y., Hao, X., Yan, S., Qian, M. Multispectral Pedestrian Detection Based on Deep Convolutional Neural Networks. IEEE International Conference on Signal Processing, Communications and Computing, 2017, 1-4. https://doi.org/10.1109/ICSPCC.2017.8242507
9. Huang, X., Ge, Z., Jie, Z., Yoshie, O. NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing. IEEE Conference on Computer Vision and Pattern Recognition, 2020, 10750-10759. https://doi.org/10.1109/CVPR42600.2020.01076
10. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I. S. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. IEEE Conference on Computer Vision and Pattern Recognition, 2015, 1037-1045. https://doi.org/10.1109/CVPR.2015.7298706
11. He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
12. König, D., Adam, M., Jarvers, C., Layher, G., Teutsch, M. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, 243-250. https://doi.org/10.1109/CVPRW.2017.36
13. Lin, C. Z., Lu, J. W., Wang, G., Zhou, J. Graininess-Aware Deep Feature Learning for Robust Pedestrian Detection. IEEE Transactions on Image Processing, 2020, 29, 3820-3834. https://doi.org/10.1109/TIP.2020.2966371
14. Liu, J. J., Zhang, S. T., Wang, S., Metaxas, D. N. Multispectral Deep Neural Networks for Pedestrian Detec-

tion. British Machine Vision Conference, 2016, 73, 1-12. https://doi.org/10.5244/C.30.73

15. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting. European Conference on Computer Vision, 2018, 11218, 618-634. https://doi.org/10.1007/978-3-030-01264-9_38
16. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, 2016, 9905, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2
17. Wagner, J., Fischer, V., Herman, M., Behnke, S. Multispectral Pedestrian Detection Using Deep Fusion Convolutional Neural Networks. Bruges: ESANN, 2016, 587, 509-514.
18. Wolpert, A., Teutsch, M., Sarfraz, M. S., Stiefelhagen, R. Anchor-free Small-scale Multispectral Pedestrian Detection. British Machine Vision Conference, September ,2020, 1-14.
19. Wu, C. L., Yang, G., Lin. Y. S. Semantic Segmentation of Glioma Images Based on Dense Multi-scale Dilated Convolution. Computer Application and Software, 2023, 40(01), 234-240.
20. Pang, Y., Cao, J., Wang, J., Han, J. JCS-Net: Joint Classification and Super-Resolution Network for Small-Scale Pedestrian Detection in Surveillance Images. IEEE Transactions on Information Forensics and Security, 2019, 14(12), 3322-3331. https://doi.org/10.1109/TIFS.2019.2916592
21. Pang, Y. W., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., Shao, L. Mask-Guided Attention Network for Occluded Pedestrian Detection. IEEE International Conference on Computer Vision, 2019, 4966-4974. https://doi.org/10.1109/ICCV.2019.00507
22. Redmon, J., Divvala S., Girshick R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition ,2016, 779-788. https://doi.org/10.1109/CVPR.2016.91
23. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
24. Rezatofighi, H., Tsoi, N., Gwak, J. Y., Sadeghian A., Reid I., Savarese S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. IEEE Conference on Computer Vision and Pattern Recognition, 2019, 658-666. https://doi.org/10.1109/CVPR.2019.00075
25. Shaikh, Z. A., Van Hamme, D., Veelaert, P., Philips, W. Probabilistic Fusion for Pedestrian Detection from Thermal and Colour Images. Sensors, 2022, 22(22), 8637. https://doi.org/10.3390/s22228637
26. Xie, Y. M., Wang, H. L. Improved Algorithm for Long-distance and Small-size Pedestrian Detection in Complex Background. Computer Engineering and Design, 2021, 42(05), 1323-1330. https://doi.org/10.16208/j.issn1000-7024.2021.05.018
27. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. IEEE International Conference on Computer Vision, 2019, 5126-5136. https://doi.org/10.1109/ICCV.2019.00523
28. Zhao, S., Chen, S. Y., Wang, Q. Y. Research on Infrared Image Pedestrian Detection Algorithm in Complex Night Scenes. Infrared Technology, 2021, 43(06), 575-582.