Multimedia Tools and Applications (2022) 81:39655–39674
https://doi.org/10.1007/s11042-022-13058-w

# Pedestrian detection in infrared image based on depth transfer learning

Zhiwen Wang [1] · Jing Feng [2,3] · Yifeng Zhang [2]

Received: 19 April 2021 / Revised: 6 September 2021 / Accepted: 4 April 2022 /
Published online: 30 April 2022


**Abstract**
Because of the difficulty in feature extraction of infrared pedestrian images, the traditional methods of object detection usually make use of the labor to obtain pedestrian features, which suffer from the low-accuracy problem. With the development and the progress of science and technology, deep learning has gradually stepped into the problem of object detection, and achieved good results. In this paper, aiming at the defects of deep convolutional neural network, such as the high cost on training time and slow convergence, a new algorithm of MoblieNet V2(1.4) + SSD infrared image pedestrian detection based on transfer learning is proposed, which adopts a transfer learning method and the Adam optimization algorithm to accelerate network convergence. For the experiments, we augmented the OUS thermal infrared pedestrian dataset and our solution enjoys a higher mAP of 94.8% on the test dataset. The experimental results show that our proposed method has the characteristics of fast convergence, high detection accuracy and short detection time.

**Keywords** Deep learning · Transfer learning · SSD · Pedestrian detection

## 1 Introduction

With the development and progress of science and technology, pedestrian detection has become a hot topic in the field of artificial intelligence.Pedestrian targets are non-rigid objects, which are easily affected by posture, angle of view, target occlusion, clothing material and thickness, etc. It is difficult to detect [21]. At present, visible image, Lidar, infrared image and

✉ Zhiwen Wang
wzw69@126.com

[1] School of Computer Science and Telecommunication Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China

[2] School of Electrical and Information Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China

[3] South China Institute of Software Engineering, GU, Guangzhou 510990 Guangdong, China

so on are the main methods of pedestrian detection at night.Visible image is strongly affected by lighting conditions, and the image quality at night is poor, so it can not detect pedestrians very well [9]. Lidar, which relies on ranging principles, has the advantages of high accuracy and less interference. It can be imaged night or day.However, Lidar is generally used to detect large-scale objects, such as ships, vehicles, etc. Pedestrian targets are relatively small in scale, and their point cloud data is small and sparse. Lidar is a small and weak target, so detection is very difficult.In infrared imaging technology, the brightness of an object is determined by temperature and thermal radiation, and pedestrian temperature is usually higher than background, so pedestrian targets are generally brighter than background. Therefore, the research on infrared pedestrian detection technology has great market application potential.The research results can be applied to assisted driving, unmanned vehicle, intelligent video monitoring and intelligent transportation.

At present, the commonly used pedestrian detection methods include traditional methods, feature-based extraction methods, multi-information fusion methods and the emerging in-depth learning methods. Traditional detection methods disadvantage low detection accuracy, poor generalization ability and low robustness, which make it difficult to meet the application requirements. Feature-based methods are provided with higher accuracy than traditional methods, but they rely on manual feature extraction, poor generalization, and long operation time, which cannot meet the practical application. In order to further improve the detection accuracy, some scholars have proposed a method of multi-information fusion. Although this method reduces the rate of leak detection, it is difficult to apply widely because of the high cost of equipment.

In recent years, with the rapid development of artificial intelligence, Convolutional Neural Network (CNN) [8] has made great achievements in solving many current object detection problems. Its high detection accuracy and short detection time occupy the dominant position and become the mainstream method of solving target detection problems. Network structures such as SPP-net [7], Fast R-CNN, Faster R-CNN [14], R-FCN [4], YOLO [13], SSD [18] appeared successively. Compared with traditional target detection methods, CNN does not need to set a specific feature through experts. It can learn target features automatically through a large number of samples and has strong generalization ability.

In order to reduce the false detection rate, Yong et al. [23] optimize the detection process by using the correlation between pedestrian target and background, and use a multi-task depth model to coordinate operation. However, this algorithm produces more candidate frames, takes a lot of computations and time, and does not use regression operation on the generated detection window, which results in inaccurate positioning of pedestrian targets. Girshick et al. proposed a target detection algorithm [2] called region-nominated convolution network (R-CNN). It uses traditional methods to extract about 2000 region nominations in the image, then uses CNN to extract the feature vectors and input them into SVM, and finally uses regression to correct the candidate box. R-CNN improves the detection accuracy, but this method has complex training steps and takes a long time. To solve this problem, Che Kai et al. [6] proposed an improved Fast R-CNN infrared pedestrian detection algorithm. An adaptive ROI area extraction algorithm is proposed to reduce the number of ROI areas generated by the network while ensuring the accuracy of pedestrian recognition in infrared images, which greatly reduces the amount of computation and speeds up the pedestrian detection in infrared images. Then, in order to make the positioning box more accurate, three different scales of a priori box are selected to calculate its confidence level, and the coordinates of the results are weighted. Experiments show that this feature fusion algorithm improves the reliability and

accuracy of pedestrian detection in infrared images by comparing with the traditional algorithm combining HOG features with SVM, the multi-feature fusion algorithm and the original Fast R-CNN algorithm. However, using non-deep learning to extract candidate regions takes a long time to detect.

Wang Dianwei et al. [1] proposed an improved algorithm for YOLOv3 for pedestrian detection of infrared video, which has high miss detection rate and low accuracy. There are three main improvements. First is to cluster the candidate boxes by K-means to select the optimal size and number. Second, fine-tune the VOC dataset-trained network with different resolution images to further improve the accuracy. Third, the network is trained with different sizes of images so that the network can detect multi-scale images. Through experimental comparison and analysis, this method can improve the detection accuracy without restricting the user input image size, and has a good detection effect for infrared pedestrians. However, for targets that overlap or are close to each other, the miss rate is high and the generalization ability is weak.

Migrative learning can make up for this shortage and make full use of it before applying it to new fields, which has good generalization performance. At the same time, because in practice, the threshold for deep learning is very high. Training a network requires a large number of samples and hardware resources. This makes in-depth learning not widely used. This paper first changes the basic network of SSD (single shot multibox detector) to a lighter MobileNet network, which greatly reduces the computational load and improves the training time. And In order to get better results on small samples and low-configuration hardware, this paper uses a combination of migration learning and in-depth learning to make full use of previously labeled data, while ensuring the accuracy of the model on new tasks.

In this paper, using infrared image information, based on the SSD network, combined with migration learning, the pedestrian detection of infrared image is studied, which mainly consists of five parts, the main contents are as follows:

1. Introduction. This paper mainly introduces the current night pedestrian detection methods and the research results of some scholars. Points out the problems in the current research and the solutions presented in this paper.
2. Model introduction. The deep learning model used in this paper is SSD network. Change its feature extraction network VGG-16 to a lighter MobileNet network. MobileNet V2 (1.4) network is used after considering both accuracy and parameter quantity.
3. Transfer learning. This paper mainly introduces a part of the theoretical knowledge of migration learning and the migration strategies used in this paper.
4. Experimental results and analysis. First, the hardware configuration used in this paper is introduced. Then, the OUS thermal infrared dataset and the preprocessing method for the dataset are described. It is also expanded by transformations such as shift and random clipping. Finally, the experimental results of this paper are displayed and compared.
5. Summary. The main work done in this paper is summarized.

## 2 Overview of deep learning algorithms

### 2.1 Introduction to deep learning algorithm

In 2006, Hinton et al. [5], the father of Neurology, studied and proposed deep learning algorithm. In this algorithm, there are many neurons, which are independent of each other.

By connecting with the upper neurons, the input of the lower neurons is received, and more semantic information with abstract representation is obtained in the form of combination and transformation, so as to achieve the purpose of internal distributed characteristics of learning objectives. It belongs to a deeper neural network. Deep learning algorithm includes many key technologies, including CNN, adaptive learning algorithm and so on.

## 2.2 Convolutional neural network

### 2.2.1 Network structure

CNN is a kind of multi-layer neural network, which is mainly used to recognize two-dimensional features. CNN's network structure is mainly composed of five parts: input layer, convolution layer, pooling layer, full connection layer and output layer, which have strong invariance in the face of translation, scaling and other deformations [19] [17]. CNN network structure is shown in Fig.1.

### 2.2.2 Training process

In essence, the input and output of CNN is a kind of nonlinear mapping, which is mainly learned by training the known model [22]. In order to ensure the success of training, CNN will take initial value operation on network weight before model training, so as to carry out a series of training work. The specific training process is divided into forward propagation and back propagation stages.

CNN forward propagation refers to the process in which data samples are input to neural network and output results are obtained after each hidden layer. Set the input of current layer as $x^{l-1}$ and the output as $x^{l}$. the relationship between them is calculated by Eq. (1).

$$x^{[l]} = f\left(W^{[l]}x^{[l-1]} + b^{[l]}\right) \tag{1}$$

Where, $l$ represents the number of layers of the neural network; $W$ represents the weight; $b$ represents the bias; $f(\cdot)$ represents the activation function.

When CNN completes a forward propagation, it also needs to define a certain error to express the network state after the completion of propagation. CNN's back propagation stage is the process of optimizing the network. The error is back propagation from the back to the front. After receiving the error, the upper neurons update their own weights, and continue to

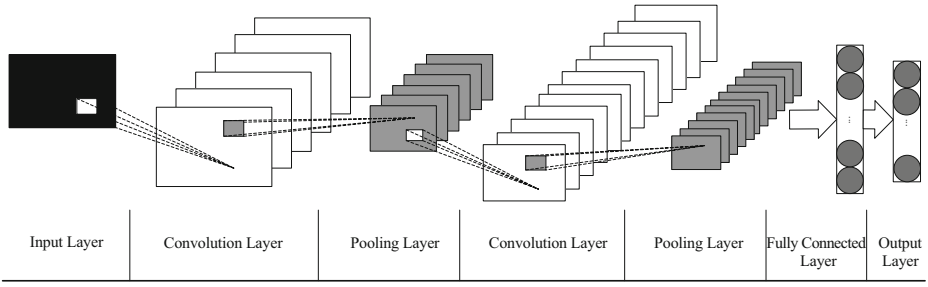



**Fig. 1** Typical CNN network structure

iterate the above process until convergence, and finally get the $w$ and $b$ which make the error the smallest. The error of each output neuron and the whole network is minimized, so that the actual output value is closer to the ideal value.

## 2.3 SSD model introduction

### 2.3.1 SSD network structure

The Single Shot MultiBox Detector (SSD) algorithm is composed of the basic network VGG-16, and then several volume layers. The network structure of SSD is shown in Fig. 2, which classifies the characteristic graphs of different sizes and can detect the targets at various scales. It can be seen from the figure that the characteristic map sizes of the model are 38 × 38, 19 × 19, 10 × 10, 5 × 5, 3 × 3, 1 × 1 [15]. After the network extracts the features, it is input into the subsequent network to calculate the target location and target type. For targets of different sizes, the previous solution is to perform multi-scale transformation on the original image and detect them at different scales, while SSD algorithm solves this problem by introducing the concept of default boxes. Default boxes is a candidate box with a fixed aspect ratio and size. Fixed areas are selected in different convolution layers. Different receptive fields of different convolution layers are used to detect targets of different sizes. Based on these areas, the coordinates, size and object types of the candidate box are regressed through network calculation. Finally, end-to-end training is realized through multi-scale loss function.

### 2.3.2 Mobilenet V2 + SSD network structure

The original SSD algorithm uses VGG-16 to extract features. The top-5 test accuracy of VGG-16 on ImageNet dataset is 90.1%, which is higher than that of most networks. However, it has more than 100 million parameters, so it has high computational complexity and slow calculation [24]. After that, some scholars put forward MoblieNet network structure, in which

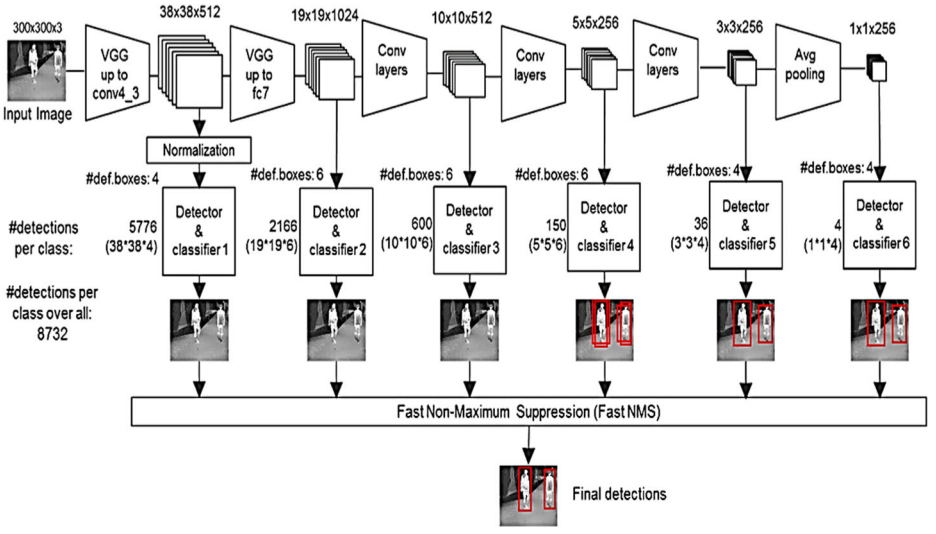



**Fig. 2** SSD network structure

MoblieNet V1 parameters are more than 4 million, which is much less than VGG-16. Moreover, MoblieNet V1's top-5 accuracy on ImageNet dataset is 89.85%, slightly lower than that of VGG-16, but the difference is not significant. Since MoblieNet V1 uses relu activation function, it will cause data collapse [20]. In order to solve this problem, they proposed MoblieNet V2 network. By changing the network structure and activation function, MoblieNet V2 will not lose too much information and retain more complete features. Its precision is higher than MoblieNet V1, and the parameter quantity is reduced by more than 700,000. Therefore, in this paper, MoblieNet V2 (1.4) network is used to replace VGG-16 for feature extraction, which not only improves the detection accuracy, but also greatly shortens the detection time. This paper compares the top-1 and top-5 test accuracy, number of parameters and CPU running time of VGG-16 and MoblieNet networks. The results are shown in Table 1.

In order to further improve the detection accuracy and reduce the training time, the original feature extraction network VGG-16 of SSD network is replaced by mobilenet V2 network. The improved network structure is shown in Fig.3. Extract conv11, conv13 and conv14_ 2, Conv15_ 2, Conv16_ 2 and conv17_ 2-layer feature map for prediction classification. In Fig.3, the red convolution box is the depth separable convolution, and the white convolution is the ordinary convolution [16].

### 2.3.3 SSD network loss function

SSD loss function is defined as the weighted sum of localization loss (LOC) and confidence loss (CONF), which can be calculated by Eq. (2).

$$L(x,c,l,g) = \frac{1}{N}\left(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)\right) \tag{2}$$

In Eq. (2), $N$ is the number of positive samples of the prior box, $c$ is the predicted value of category confidence, $l$ is the predicted value of the corresponding boundary box of the prior box, $g$ is the position parameter of the real boundary box, and $\alpha$ is the weight coefficient.

Position error is defined by Smooth L1 loss and Eq. (3) and Eq. (4).

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k} smooth_{L1}\left(l_i^m - \widehat{g}_j^{\,m}\right) \tag{3}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

**Table 1** comparison of feature extraction network schemes

| network model | Top-1(%) | Top-5(%) | MParams | CPU(ms) |
|---|---|---|---|---|
| VGG-16 | 71.5 | 90.1 | 138 | 714 |
| MoblieNet V1 | 70.81 | 89.85 | 4.2 | 123 |
| MoblieNet V2 | 71.8 | 91 | 3.47 | 80 |
| MoblieNet V2(1.4) | 75 | 92.5 | 6.06 | 149 |

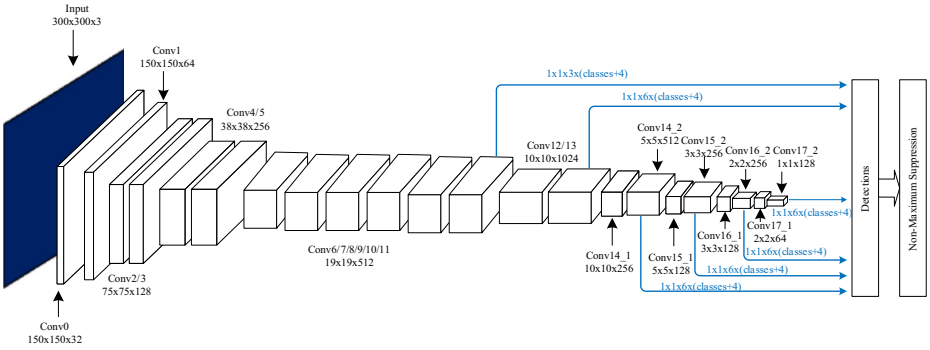



**Fig. 3** Mobilenet V2 + SSD network structure

Since $l$ is the encoded value, $g$ should be encoded first to get $\hat{g}$, and the calculation equation is Eq. (5) and Eq. (6).

$$\hat{g}_j^{cx} = \left(g_j^{cx} - d_i^{cx}\right)/d_i^{w}/variance[0] \qquad \hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^{h}/variance[1] \tag{5}$$

$$\hat{g}_j^{w} = \log\left(g_j^{w}/d_i^{w}\right)/variance[2] \qquad \hat{g}_j^{h} = \log\left(g_j^{h}/d_i^{h}\right)/variance[3] \tag{6}$$

In Eqs. (5) and (6), $d$ is the prior box position and *variance* is the super parameter, which is used to adjust the detection value to scale $\hat{g}$. For the confidence error, the softmax loss is calculated by Eq. (7).

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p}\log\left(\hat{c}_i^{p}\right) - \sum_{i\in Neg}\log\left(\hat{c}_i^{0}\right) \quad \text{where } \hat{c}_i^{p} = \frac{\exp\left(\hat{c}_i^{p}\right)}{\sum_p \exp\left(\hat{c}_i^{p}\right)} \tag{7}$$

In Eq. (7), $x_{ij}^{p} \in \{1,0\}$ is a parameter index, when $x_{ij}^{p} = 1$, it means that the $i$ prediction boundary box matches the j real boundary box, and the boundary box category is $p$. $c$ is the predictive value of class confidence. The higher the probability prediction of $p$, the smaller the loss, and the probability is generated by softmax.

## 3 Infrared pedestrian detection based on depth transfer learning

### 3.1 Transfer learning

In CCN, we repeatedly convolute the local region of a picture to reduce the area and increase the number of channels [25]. From the interpretation of the convolution layer by Matthew D. Zeiler and rob Ferguson, it can be seen that the shallow convolution layer summarizes the more geometric abstract content. As the convolution layer becomes deeper and deeper, the content of clustering becomes more and more specific. So if we keep the first n layers of convolutional neural network and cut off the unwanted ones, the outline of training in the first

few layers is still very helpful to us. Based on this, the same function can be realized in the new field through the data characteristics and model parameters learned in the previous field.

Transfer learning is a machine learning method, which is to transfer knowledge from one domain (source domain) to another domain (target domain), so that the target domain can achieve better learning effect [12]. The process of transfer learning is shown in Fig.4.Through the data features and model parameters learned in the previous field, the same function can be realized in the new field. Transfer learning is conducive to accelerate the convergence of the model, fine data characteristic classification, and improve the classification effect.

Generally speaking, the amount of data in the source domain is sufficient, while the amount of data in the target domain is very small. This scenario is very suitable for migration learning. For example, we need to classify a task. The data in the task is not enough (target domain), but there are many related training data (source domain). In this case, if appropriate transfer learning methods can be adopted, the classification and recognition results of tasks with insufficient samples can be greatly improved [19].

Migration learning has many advantages, such as small demand for data; Good generalization and good results can be obtained on similar and different data sets; The training model is stable and robust. It is easy to debug and can improve network performance. It is helpful to accelerate the convergence speed of the model, refine the data feature classification and improve the classification effect.

### 3.2 Migration strategy

The designed algorithm is used to learn the features of the target in the data set and train the weight model which can detect the target. In this paper, migration learning is used to detect the infrared pedestrian image, that is to initialize the network parameters based on the network weight that MoblieNet V2(1.4) + SSD has trained and iterated 100,000 times on PASCAL VOC data set. The training iterative weight model has the ability to extract features, and this method can improve the network training effect and accelerate convergence. The main steps are:

1. Use PASCAL VOC data set to train MoblieNet V2(1.4) + SSD network and save the weight;
2. After transforming the ous infrared pedestrian data set into three channel image, we expand the data, divide the training set and test set into 8:2, and transform them into tfrecord format file;
3. The training set is used to fine tune the pre-trained MoblieNet V2(1.4) + SSD network. Firstly, the pre-trained weight of PASCAL VOC data set is converted to the weight on the hot infrared pedestrian data set of ous. The output layer layer, conv4, is then directly

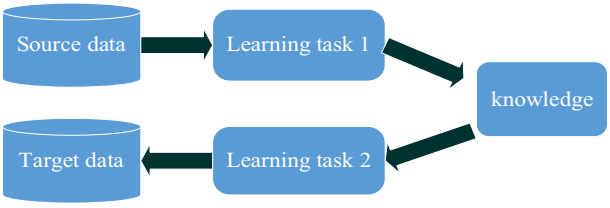



**Fig. 4** Transfer learning flow chart

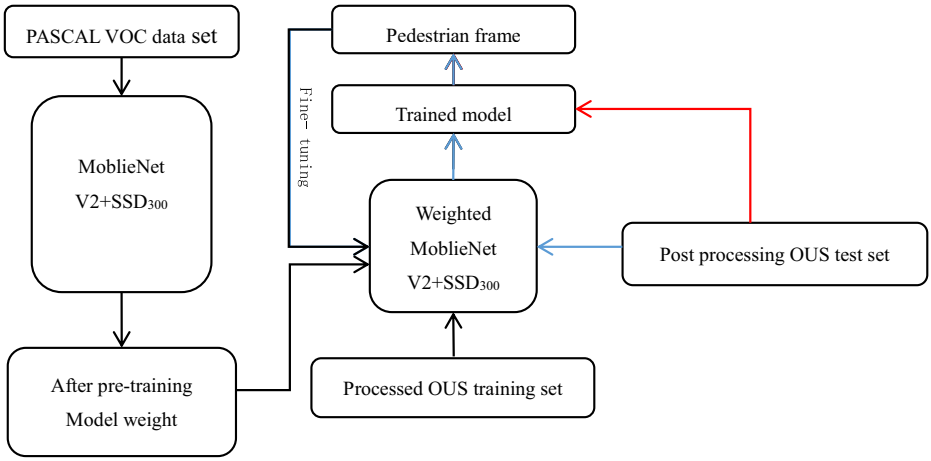



**Fig. 5** MoblieNet V2(1.4) + SSD migration learning flow chart

involved Conv11, Conv13, Conv14_2, Conv15_2, Conv16_2 and Conv17_2. The network weight of layer 2 is relearned. Then, the multi-scale feature map is fused by convolution layer to generate the bounding box containing the probability of interested objects, and the NMS is used to generate the detection results. Finally, the error of the model in the training set is used to iterate the training model to get a reasonable model for data fitting;

4. Input the test set into the trained model, and adjust the super parameters according to the results;
5. Test set is used to evaluate the model and verify the robustness of the model.

The overall experimental flow is shown in Fig. 5. The black arrow represents the training process, blue represents the verification process, and red represents the test process.

## 4 Experimental results and analysis

### 4.1 Experiment configuration and super parameter setting

The hardware and software configuration of this experiment is shown in Table 2.

In the process of network training, impulse constant is 0.9, initial learning rate is 0.0004, multi-step strategy learning, weight attenuation coefficient is 0.0005, and beach size is 16. At the same time, the learning rate and regularization coefficient are adjusted by observing the change of loss function and the difference of training and testing accuracy.

**Table 2** Experimental configuration

| name | related configuration |
|---|---|
| operating system | Windows 10 Enterprise 2015 |
| CPU | Inter Core i7 |
| Deep learning framework | Tensorflow-1.13.0rc2 |
| data processing | Python 3.5, Open CV |

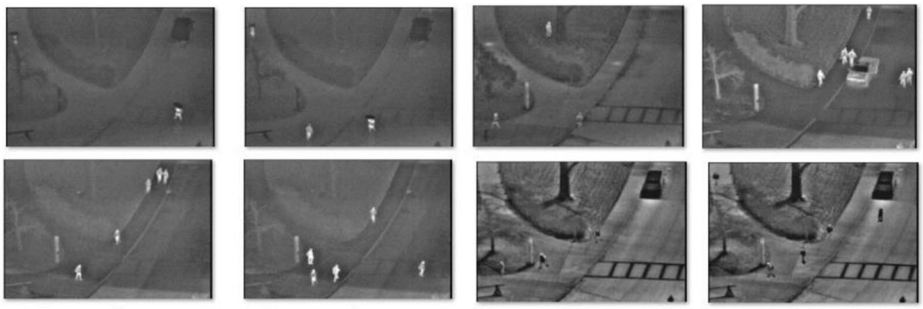


**Fig. 6** Example of OSU thermal infrared pedestrian data set

### 4.2 Data set

In this paper, the OSU hot infrared pedestrian data set is a part of otcbvs benchmark database, which contains 10 hot infrared image sequences under fixed monitoring perspective, totaling 284 images. The image capture environment is diverse, including sunny, cloudy and rainy days [11]. Figure 6 is partial example of this dataset.

In Fig. 6, the weather conditions of the first and second images in the first row are light rain, the third image is haze weather, and the fourth image is sunny. The first and second images in the second row were taken in the morning, and the weather condition was less cloud. The third and fourth images were taken in the afternoon, and the weather condition was less cloud.

#### 4.2.1 Data preprocessing

The infrared pedestrian image is preprocessed for its low signal-to-noise ratio and low contrast. In terms of denoising, several main denoising methods are compared. The experimental results are shown in Table 3. Wavelet adaptive threshold denoising performs better in both structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) of images with the same noise than each filter method. It not only filters out noise, but also retains good edge characteristics. In this article, it is the first step for image processing. In this paper, the image is decomposed by two layers using a coif2 discrete wavelet basis.

By comparing the histograms of several algorithms, an adaptive histogram equalization method with limited contrast is used to improve the image contrast. The implement-tation steps of the preprocessing are shown in Fig. 7.

In order to verify the generality of this method, five images of different weather in ous thermal infrared pedestrian database are selected for experiment. The simulation results are shown in Fig. 8. The first column in the figure is the original image, the second column is the

**Table 3** PSNR and SSIM values of each filtering method

| evaluating indicator | 5×5 template median filtering | 3×3 template median filtering | 5×5 template median filtering | Gaussian filtering | Bilateral filtering | adaptive median filtering | Wavelet adaptive threshold |
|---|---|---|---|---|---|---|---|
| SNR | 19.5501 | 21.7274 | 20.9566 | 21.1829 | 20.8755 | 21.7390 | 25.1936 |
| PSNR | 29.1098 | 31.2872 | 30.5163 | 30.7427 | 30.4352 | 31.2988 | 34.7533 |
| SSIM | 0.5848 | 0.6400 | 0.5863 | 0.6485 | 0.6441 | 0.6387 | 0.8240 |

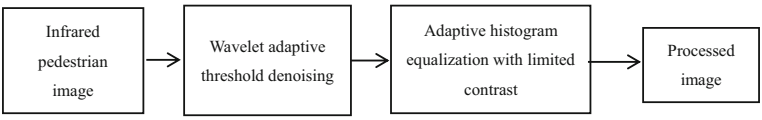



**Fig. 7** infrared image preprocessing

result of wavelet adaptive threshold denoising, and the third column is the result of the second column image after adaptive histogram equalization with limited contrast. It can be seen that the original image after the above process becomes smooth and clear, enlarges the details, and improves the contrast and brightness.

### 4.2.2 Data set expansion

As there are only 284 images in OUS data set, in order to achieve better results, this chapter uses data enhancement method to expand the data set, mainly for turning up and down 10% of the images, mirror turning 50% of the images, left and right translation 20%, shear transformation, image scaling 80% to 120%, random rotation, after the expansion of the data set to 2823 images. Figure 9 is partial image of the expanded dataset.

Because the infrared image is a single channel image, and the MoblieNet V2(1.4) + SSD network requires the input image to be a three-channel image, this paper uses the method of assigning the original channel value to RGB three channels to transform the infrared image into three channels. Then, the 2823 pedestrian images are labeled with LabelImg, and the correspond-ing .xml format label of the infrared pedestrian image is constructed. Finally, the data set label is converted to .csv format, and the image name, image width and height, target category and coordinates contained in the image label of .xml format are written into the table

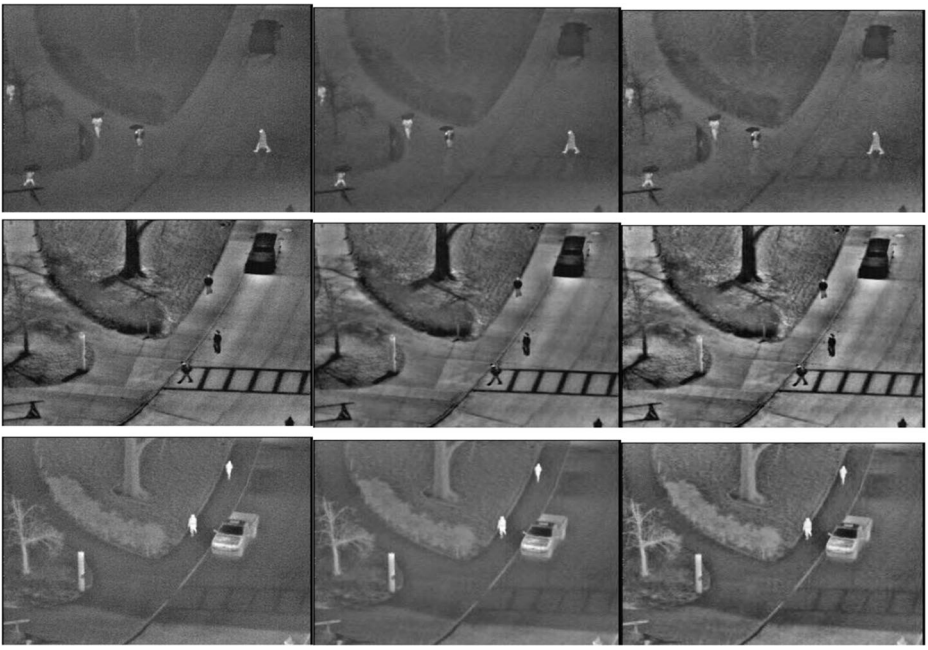


**Fig. 8** Simulation results

**Fig. 9** is the partial image of the expanded dataset

file. At the same time, the data set is divided into training set and test set in the proportion of 8:2. The specific values of the data set are shown in Table 4.

In SSD network, the label information contained in .xml format and .cvs format files can not be processed directly. It needs to be further converted into tfrecord format files that can be read by the network .tfrecord format file is a binary file that is more convenient to copy and move, and does not need a separate label file. Convert the image and label in .xml format to .tfrecord format file for reading.

### 4.3 Evaluating indicator

When IOU > 0.5 in the experimental results, it indicates that the prediction is successful. Generally, precision (P), recall (R) and mean average precision (map) are used to evaluate the model. Table 5 shows the confusion matrix.

The accuracy rate, also known as the precision rate, indicates how many real cases are predicted out of the positive cases, and is calculated by Eq. (14).

$$\mathrm{Pre}cision = \frac{TP}{TP + FP} \tag{14}$$

Recall rate indicates how many of all positive samples are predicted correctly, which is calculated by Eq. (15).

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{15}$$

Average precision (AP) is a value between 0 and 1, which can measure the quality of the model. Use formula (16) and formula (17) to calculate.

$$AP = \sum_{r=0}^{1} (r_{n+1} - r_n) p_{interp}(r_{n+1}) \tag{16}$$

**Table 4** Specific data set values

| data set | Number of images | Number of pedestrians |
|---|---|---|
| Training set | 2258 | 5568 |
| Test set | 565 | 1336 |
| total | 2823 | 6904 |

**Table 5** confusion matrix

| The truth | Forecast results | |
|---|---|---|
| | Positive example | Negtive example |
| Positive example | TP(True Positive) | FN(False Negtive) |
| Negtive example | FP(False Positive) | TN(True Negtive) |

$$p_{interp}(r_{n+1}) = \max_{\widetilde{r} \geq r_{n+1}} p\left(\widetilde{r}\right) \tag{17}$$

mAP means that the mean accuracy rate is average for all classes of AP.

### 4.4 Experimental result

To verify the accuracy of the weight model to detect the target, we mainly look at the average accuracy mean (mAP) calculated after detecting all the data in the test set. In this paper, Adam descent method is used to train all pictures in batches, which improves the speed of updating all parameters when the gradient descends, and obtains the optimal value of parameter solution as much as possible. In this paper, the number of pictures per input model training is 16, the number of iterations is 9000, and the accuracy of the final test set is 94.8%. According to the map calculated by every 10,000 verification, the curve of accuracy increasing with the number of steps is drawn as shown in Fig. 10.

Figure 10 shows the convergence curve of the loss value in the training process, the abscissa is the number of iterations, and the maximum number of iterations is 9000. Finally, the total loss decreased to about 1.94, and the result of network training was ideal.

Localization_loss in Fig. 11 is the loss of boundary box regression. Classification_loss is the loss that classifies the detected objects into various categories. The total_loss is total loss.

It can be seen from Figs. 10 and 11 that the curve of both accuracy and loss is fluctuant, not smooth. This is because the Adam descent method is adopted in this paper. Each training is only a batch of images in all data sets. The loss and accuracy of this batch of images are calculated. It can be seen from Figs. 9 and 10 that although the curve is fluctuating, the overall trend of accuracy is The potential is rising, the loss is falling, and the model is converging gradually.

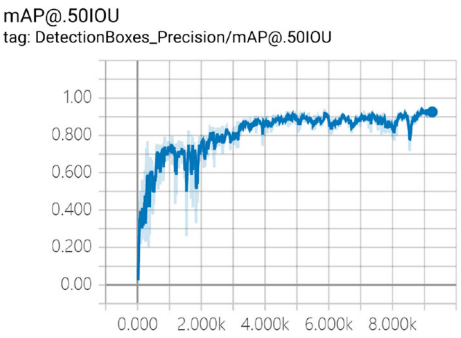
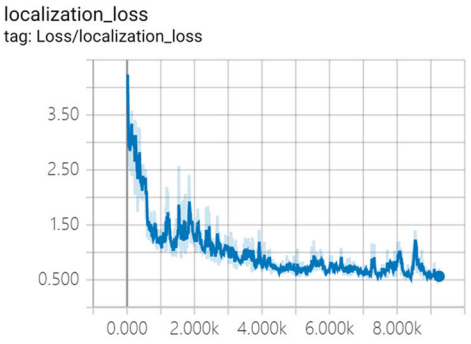
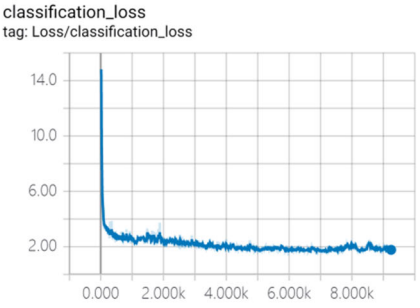


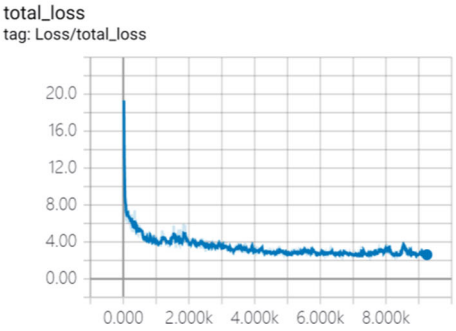

**Fig. 10** Map function curve




(a) localization_loss function curve




(b) classification_loss function curve

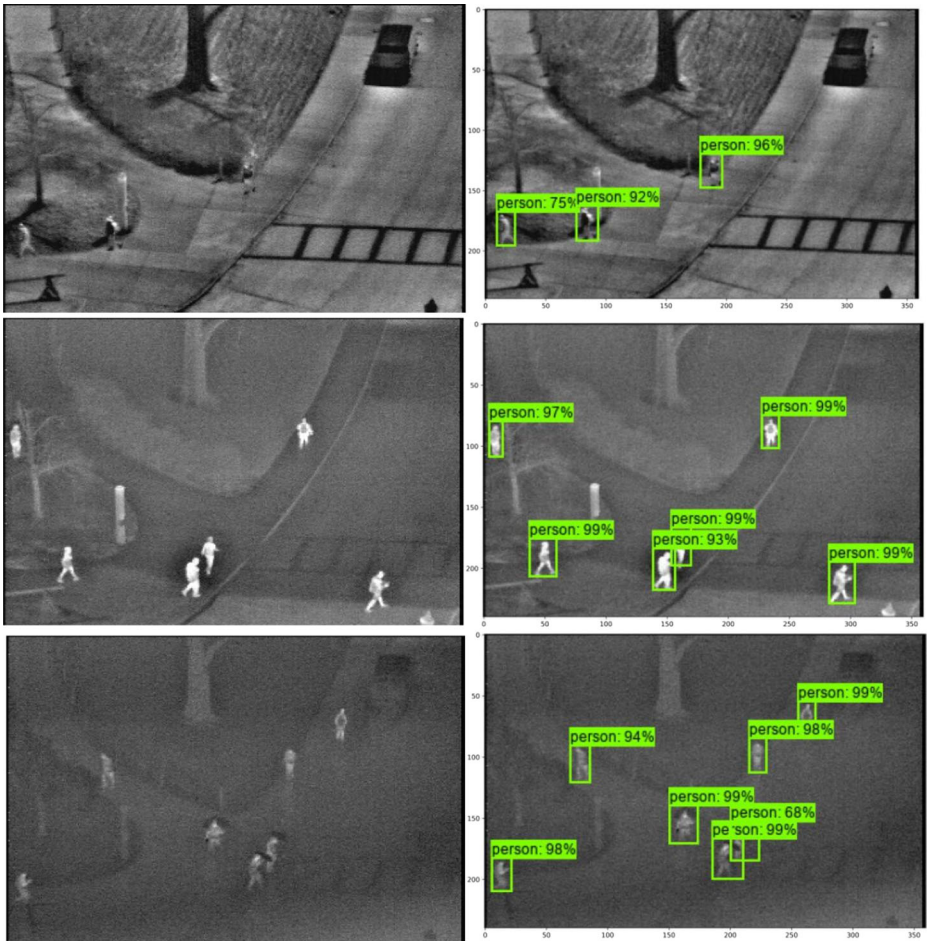



(c) Total loss curve

**Fig. 11** Loss function curve (**a**) localization_loss function curve (**b**) classification_loss function curve (**c**) Total loss curve

From Fig. 10, it can be seen that with the increase of training times, the map is constantly improving, and the curve change of accuracy rate between 0 and 2000 iterations is very obvious, which means that the model is in the learning stage; after 3200 iterations, the curve of accuracy rate is basically stable, and there is no significant change, because the model is gradually converging The update of the number is fitting to the optimal solution. When the number of training reaches 5000, the accuracy tends to saturation. When the number of training reaches 9000, the average accuracy of MoblieNet V2(1.4) + SSD algorithm is about 94.8%.

The data set of this paper consists of the original OUS thermal infrared pedestrian data set and the expanded image based on it. Figures 12 and 13 are partial image detection results. In the test, the algorithm network is built first, and then the weight model of 9000 training times is directly called to calculate the position offset, target category and predefined box position of the target in the picture. According to the predefined box position and position offset, the final position of the target (target center, target length and target width) is obtained. The yellow green color is set to draw the target frame, and the target is given as a certain type Probability. Figure 12 shows the processing results of the original ous thermal infrared data set image in the data set of this chapter. When the prediction probability is more than 50%, the target is considered to be a pedestrian. In Fig. 12(a), the probability of the first image classified as a pedestrian for three targets is 75%, 92% and 96% from left to right. All the three images can be predicted successfully.




(a) Transformed image (b) detection result

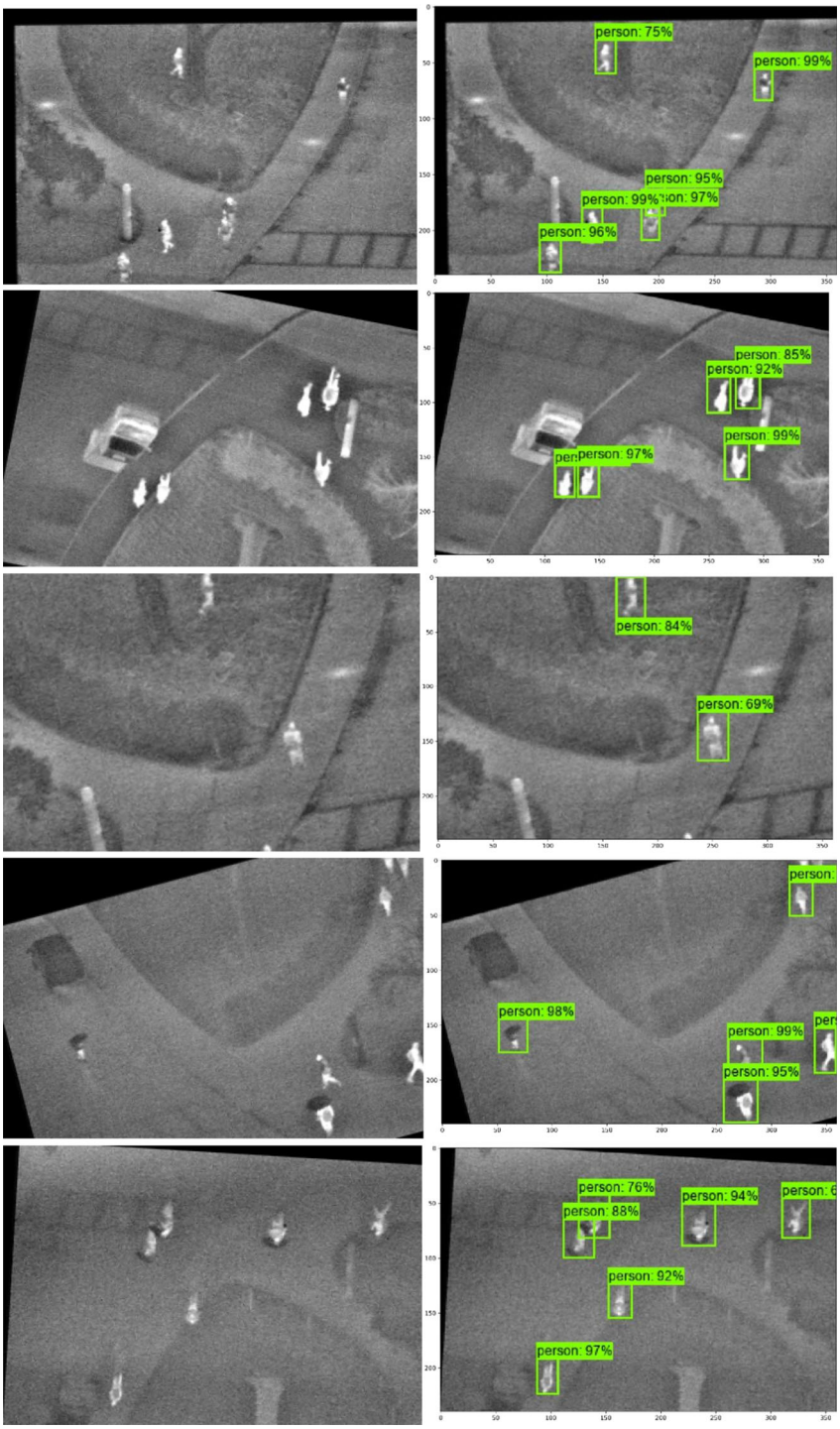
**Fig. 12** Pedestrian detection results in infrared image (**a**) Transformed image (**b**) detection result

Figure 13 shows the detection results of the image after the original OUS thermal infrared pedestrian data set image is expanded in the data set. The enlarged image is obtained by free combination of a series of transformations such as rotation, scaling, and flipping of the original image, as shown in Fig. 13(a). Figure 13(b) is the detection result. It can be seen that the 5 transformed images can be detected successfully.

## 4.5 Comparison of experimental results

### 4.5.1 Comparison of common network parameters and test time

Table 6 shows the base network VGG-16 in SSD, and compares the amount of operation and parameters of lightweight network MoblieNet V2(1.4). FLOPs are floating-point operators, which measure the complexity of the model.




(a) Transformed image (b) detection result

**Fig. 13** Pedestrian detection results of transformed infrared image (**a**) Transformed image (**b**) detection result

**Table 6** Comparison of parameter quantity and calculation quantity of basic network

| Basic network | FLOPs | MParams |
|---|---|---|
| VGG-16 [10] | $15.47*10^9$ | 138 |
| MoblieNet V2(1.4) | $0.72*10^9$ | 6.06 |

Table 7 shows the comparison of commonly used pedestrian detection network parameters and CPU test time. It can be seen that the network used in this article has fewer parameters and low computational complexity.

### 4.5.2 Comparison of data set before and after optimization

The network model can achieve better results and improve the detection accuracy by expanding the hot infrared pedestrian data set of ous. Table 8 shows the comparison of network performance after the expansion and optimization of data set.

### 4.5.3 Comparison of detection results of improved algorithm

In this paper, the average accuracy rate, recall rate and detection time are used as the evaluation criteria of different algorithms in the task of pedestrian target detection in infrared image, taking into account the two requirements of accuracy and recall, measuring the overall performance of the algorithm, and evaluating the detection performance of different algorithms for pedestrian target more comprehensively. Table 9 shows the comparison of mAP, R, and single picture detection time of the data set in this paper under different network models.

It can be seen from Table 9 that the accuracy of the method in this paper is significantly higher than that of other methods, and the detection time is greatly reduced. The detection effect of infrared image is better than fast RCNN, YOLO and SSD. Its mAP can reach 94.8%.

**Table 7** Comparison of pedestrian detection network parameters

| network model | MParams | CPU(ms) |
|---|---|---|
| SSD | 262 | 98.7 |
| YOLO v3 [3] | 235 | 82.4 |
| MoblieNet V2(1.4)+SSD | 124 | 71.2 |

**Table 8** Recognition results before and after data set optimization

| Data set processing method | mAP(%) |
|---|---|
| Before data expansion | 64.17 |
| After data expansion | 90.23 |
| After data optimization | 94.80 |

**Table 9** Identification results of different models in the data set of this paper

| Model | mAP(%) | R(%) | Test time (ms) |
|---|---|---|---|
| Faster-Rcnn | 79.1 | 74.35 | 2374 |
| YOLO | 75.25 | 69.79 | 1653 |
| SSD | 77.86 | 72.33 | 3428 |
| Methods of this paper | 94.8 | 85.94 | 530 |

**Table 10** Comparison of test results of different data sets

| data sets | mAP(%) |
|---|---|
| OUS | 94.8 |
| CVC-09 | 90.6 |

#### 4.5.4 Comparison of test results of different data sets

Table 10 shows the training results of different data sets under the network in this paper. It can be seen that the map of cvc-09 is significantly lower than that of the data sets used in this paper. This is because the image background in cvc-09 dataset is complex and there are many occlusion, which reduces the accuracy.

## 5 Conclusion

At present, pedestrian detection technology has become a research hotspot, which is widely used in vehicle assistant driving, driverless vehicle, intelligent video monitoring, intelligent transportation and other aspects. At night, due to the poor illumination and other conditions, the imaging effect of visible light camera is poor, which affects the effect of pedestrian detection. Infrared imaging technology can work around the clock by capturing the heat emitted by the object, which can effectively solve this problem. Traditional methods need complex feature design and feature extraction. In this paper, a detection method based on transfer learning is proposed by using deep learning method, and MoblieNet V2(1.4) + SSD is used in Pascal The network weight on VOC data set initializes the new MoblieNet V2(1.4) + SSD network. The weight of the layer directly related to the output layer will be learned again and again through the infrared image data set, and a reasonable model of data fitting is obtained, which is more suitable for infrared image pedestrian detection. Input the test set into the trained model, adjust the super parameters according to the results, and improve the accuracy of the network. Experiments show that this method improves the network accuracy, reduces the network training time and iteration times, and speeds up the network convergence. Through the migration learning, the complexity and running time of the network are reduced. The experimental results show that the network structure proposed in this paper improves the performance of VGG16 network and can solve the problem of pedestrian detection in infrared image Appendix Table 11.

# Appendix

**Table 11** Description of some symbols

| Symbol | Definition | Symbol | Definition |
|---|---|---|---|
| CNN | Convolutional Neural Network | $g$ | True bounding box location |
| $W$ | weight | $\widehat{g}$ | Prediction box location |
| $b$ | bias | d | A priori box location |
| SSD | Single Shot MultiBox Detector | P | Precision |
| LOC | localization loss | R | Recall |
| CONF | confidence loss | AP | Average Precision, |
| $N$ | number of positive samples of the prior box | mAP | mean Average Precision, |
| $c$ | predicted value of category confidence | PSNR | Peak Signal to Noise Ratio |
| $\alpha$ | weight coefficient | SSIM | Structural Similarity |

**Acknowledgments** This research was funded by the National Natural Science Foundation of China, grant number 6192007, 61462008, 61751213, 61866004; the Key projects of Guangxi Natural Science Foundation, grant number 2018GXNSFDA294001,2018GXNSFDA281009; the Natural Science Foundation of Guangxi, grant number 2018GXNSFAA294050, 2017GXNSFAA198365; 2015 Innovation Team Project of Guangxi University of Science and Technology, grant number gxkjdx201504; Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security, grant number MIMS19-04; Natural Science School-level Project of Software Engineering Institute of Guangzhou, grant number ky202108; Guangxi Postgraduate Education Innovation Project, grant number GKYC202106, GKYC202104, YCSW2021320; College Students' innovation and Entrepreneurship Project 202110594133, 202110594134.

**Authors contribution** Conceptualization, J.F and Z.w.W.; methodology, J.F; software, Y.h.W.; validation, J.F,Z.w.W.; formal analysis, J.F; investigation, Y.f.Z; data curation, Y.f.Z; writing—original draft preparation, J.F; writing—review and editing, Z.w.W; visualization, J.F; supervision, Y.f.Z; project administration, Z.w.W; funding acquisition, Z.w.W. All authors have read and agreed to the published version of the manuscript."

## Declarations

**Conflict of interest** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

# References

1. Dianwei W, Yanhui H, Daxiang L et al (2018) Improved yolov3 infrared video image pedestrian detection algorithm. J Xi'an Univ Posts Telecommun 23(4):48–67
2. Girshick R, Donahue J, Darrell T et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition 580–588
3. Jianting S, Guiqiang Z (2020) Improved yolov3 infrared image pedestrian detection algorithm. J Heilongjiang Univ Sci Technol 30(4):442–447
4. Jifeng D, Yi L, Kaiming H et al (2016) R-FCN: object detection via region-based fully Convolu- tional networks. Conference on Neural Information Processing Systems
5. Junyu Z, Yanming Z (2017) Overview of convolution neural network in image classification and target detection. Comput Eng Appl 53(13):34–41
6. Kai C, Zhengtao X, Yufen GC et al (2018) Research on infrared image pedestrian detection based on improved fast r-cnn. Infrared Technol 40(6):578–584

7. Kaiming H, Xiangyu Z, Shaoqing R et al (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. 2014 European Conference on Computer Vision 1904–1916
8. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
9. Liang C, Xiaoming D, Mingquan Z (2016) Convolutional neural network in image understanding. Acta Automat Sin 42(9):1300–1312
10. Liu X, Li FM, Liu SJ (2020) An infrared image pedestrian detection algorithm based on improved SSD algorithm. Electr Opt Control 27(1):42–46 59
11. Ming X, Xiaosheng Y, Dongyue C et al (2018) Pedestrian detection in complex thermal infrared monitoring scene. Chin J Image Graph 23(12):1829–1837
12. Qi L (2018) Fruit image recognition system based on deep learning. Agric Eng 8(10):31–34
13. Redmon J, Divvala S, Girshick R et al (2016) You only look once: Unified, real-time object detection. 2016 IEEE Conf Comput Vis Pattern Recogn 779–788
14. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection withRegion proposal networks. Adv Neural Inf Proces Syst 28:1137–1149
15. Simon M, Rodner E (2015) Neural activation constellations: Unsupervised part model discovery with convolutional network. 2015 IEEE International Conference on Computer Vision and Pattern Recognition 1143–1151
16. Song W, Shumin F (2019) Research and improvement of SSD (single shot multibox detector) target detection algorithm. Ind Control Comput 32(4):103–105
17. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition 1–9
18. Wang P, Wang Z, Lv D et al (2021) Low illumination color image enhancement based on Gabor filtering and Retinex theory. Multimed Tools Appl 80(12):17705–17719
19. Xudong X, Liqian M (2018) Control chart recognition based on transfer learning and convolutional neural network. Comput Appl 38(S2):290–295
20. Xudong L, Mao Y, Tao L (2017) A review of target detection based on convolutional neural network. Comput Appl Res 34(10):2881–2891
21. Yancheng W, Hongchang C, Shaomei L, Gao G (2018) Pedestrian recognition neural network model based on heterogeneity of pedestrian attributes. Comput Eng 44(10):196–203
22. Yandong L, Zongbo H, Hang L (2016) A review of convolutional neural networks. Comput Appl 36(9): 2508–2515
23. Yong LT, Ping L, Xiao GW et al (2015) Deep learning strong parts for pedestrian detection. *2015 IEEE International Conference on Computer Vision* 1904–1912
24. Zhihua Z (2016) Machine learning. Tsinghua Univ Press 121–139
25. Zhihua Z (2016) Machine learning. Tsinghua University Press