



# Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios

Xiao Ke<sup>1,2</sup> · Xinru Lin<sup>1</sup> · Liyun Qin<sup>1</sup>

Received: 15 March 2019 / Revised: 3 November 2020 / Accepted: 22 January 2021 / Published online: 24 February 2021  
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

## Abstract

Pedestrian detection and re-identification technology is a research hotspot in the field of computer vision. This technology currently has issues such as insufficient pedestrian expression ability, occlusion, diverse pedestrian attitude, and difficulty of small-scale pedestrian detection. In this paper, we proposed an end-to-end pedestrian detection and re-identification model in real scenes, which can effectively solve these problems. In our model, the original images are processed with a non-overlapped image blocking data augmentation method, and then input them into the YOLOv3 detector to obtain the object position information. LCNN-based pedestrian re-identification model is used to extract the features of the object. Furthermore, the eigenvectors of the object and the detected pedestrians are calculated, and the similarity between them are used to determine whether they can be marked as target pedestrians. Our method is lightweight and end-to-end, which can be applied to the real scenes.

**Keywords** Convolutional neural network · Deep learning · Data augmentation · Pedestrian detection · Pedestrian re-identification

## 1 Introduction

With the rapid improvement of computer technology and people's living standards, the world has undergone rapid changes, and artificial intelligence has become more and more integrated into our lives, in the form of technologies such as driverless vehicles, intelligent assisted driving, pedestrian analysis, intelligent robots, intelligent video surveillance, intelligent transportation, security systems, and glasses specially designed for blind people. Pedestrian detection and re-identification technology provide extremely important technical support for these applications, and it is also a core technology. The wide application and importance of this technology make it an important part in computer vision research. Scholars have also carried out a lot of

related research in this field. Although many well-known research institutes and key universities are carrying out related research, many issues in pedestrian detection and re-identification, such as low illumination, low contrast, occlusion, and insufficient feature representation, still have to be overcome. Therefore, further research in pedestrian detection and re-identification techniques is needed.

Several methods have been proposed for pedestrian detection. In [1], image features were extracted using the wavelet method, which is the first where a machine learning algorithm has been applied to pedestrian detection. In [2, 3], the authors proposed to use the HOG + SVM algorithm; In [4], the ACF algorithm was proposed; and Nam et al. proposed the Locally Decorated Channel Features (LDCF) method, which improves the performance of the ACF and achieves the best results on the shallow model. Deep learning techniques have also been introduced into the field of pedestrian detection. Krizhevsky et al. [5] using the convolutional neural network (CNN) method to win the ImageNet competition easily, which have attracted wide attention. In [6], the authors proposed the Region-based CNN (R-CNN) method and the authors in [7, 8] designed and improved the Faster R-CNN method based on R-CNN, which has greatly improved performance and speed. In [9], the detection

✉ Xiao Ke  
kex@fzu.edu.cn

<sup>1</sup> Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China

<sup>2</sup> Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou 350003, China

algorithm of single shot multibox detector (SSD) was proposed. The authors in [10] proposed the YOLO algorithm, and then designed and improved the YOLOv2 in [11], whose accuracy rate on the visual object classes (VOC) data set reached 78.6%, and whose FPS reached 40. In 2018, the YOLOv3 [12] was proposed, whose accuracy rate on the VOC data set is nearly 4% higher than the YOLOv2 and, unlike the YOLO and YOLOv2, is sensitive to small objects.

The pedestrian re-identification process is divided into two main parts. First, the pedestrian characteristics are extracted from the picture with good expression and strong robustness, and second, the model with the highest identification and matching ability is used to improve the image. In [38], the proposed algorithm is used to detect the face, and then facial distance measurements are obtained by the geometric-based facial distance measurement technique. Finally, the measured facial distances are evaluated with the physiognomy science to reveal the characteristic properties of person. In [39], the boundary of the face is determined by the feature surface method, and then the corresponding face distance measurement value is determined by the Euclidean distance measurement technology. References [38, 39] both extract facial features, and then evaluating by facial feature distance. The features we extract are different from face recognition, and we extract the body features. For example, in [13], a triangle model is constructed to represent the shape features of human body, and the color features of each part are extracted for pedestrian recognition. In [14], the authors proposed the local maximal occurrence (LOMO) method, which uses the maximum pooling method to process the blocks in the same horizontal strip and uses homomorphic filtering to minimize the influence of illumination changes on the image. The first deep learning algorithm to solve the pedestrian recognition problem are reported in [15, 16]. In [17], multiple large data sets were used as training and Softmax was used as a classification function in the classification network. The classification model has good performance on large data sets like PRW [18] and MARS [19],

and does not need to pick sample pairs. Tong Xiao et al. also used the framework of the Faster R-CNN algorithm for pedestrian detection and pedestrian re-identification [20, 21]. Liang Zheng et al. studied the combination of different conventional and deep learning pedestrian detection and pedestrian recognition methods, and proposed that, if the detection frame is good, the identification is directly helpful [18]. The authors in [22] proposed a distance metric learning algorithm for large margin nearest neighbor (LMNN), which improved the measurement effect. In [23], the distance metric algorithm keep it simple and straight (KISS) was proposed, which does not require iterative learning, so it is suitable for large-scale data. In [24], the authors developed the local fisher discrimination analysis (LDAF) and applied it to pedestrian recognition technology. In [40], more attention is paid to semantic information, a data-driven pedestrian re-identification model based on hierarchical semantic representation is proposed, which enhances the semantic representation of features with hierarchical mid-level ‘attributes’. In [41], a kernel-PCA-based spatial descriptor is generated and evaluated the descriptor using known distance metric learning methods.

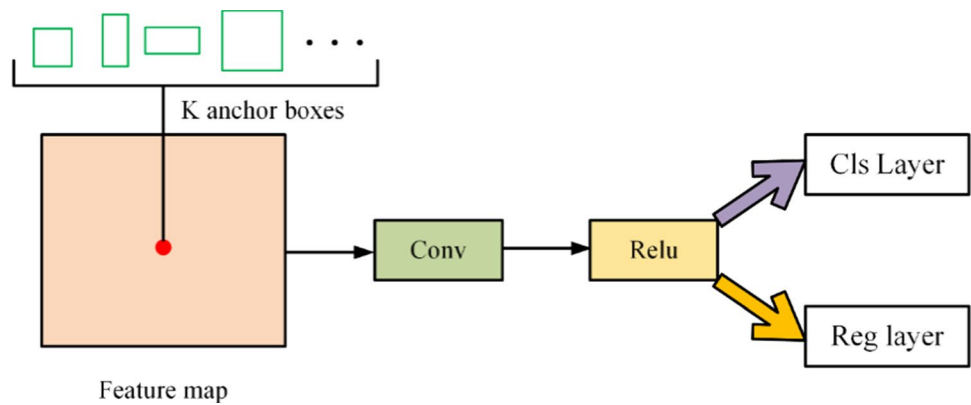
The main contributions of this paper can be summarized as follows:

1. To solve the problems of current pedestrian detection research, such as complex or similar background and prospects, diverse perspectives, diverse pedestrian attitudes, and difficulty in detecting small-scale pedestrians, a data augmentation-based YOLOv3 pedestrian detection method is proposed, which involves the following steps: first, a white edge is added to the original image to form a square, which is divided into 9 blocks to obtain 9 sub-pictures; the original picture is then put into the YOLOv3 detector, and the detected image is adopted. The multi-scale fusion method is used for the detection, which significantly improves the detection of small objects. An optimal pedestrian detection model is constructed from the proposed data augmentation-based YOLOv3 detection method, which solves the problems of diverse perspectives, diverse pedestrian attitudes, and difficulty in detecting small-scale pedestrians.
2. In the research of pedestrian re-identification, there are issues with the appearance of the same pedestrian in multi-lens and the difference in image resolution of different cameras. The existing pedestrian recognition models cannot extract the pedestrian characteristics with strong expression, and cannot distinguish between target pedestrians and non-target pedestrians. We propose a pedestrian recognition model based on lightweight CNN (LCNN), which has been successfully applied to face recognition. The model adopts NIN network, which reduces the parameters and ensures the recognition of

**Table 1** Comparison of object detection algorithm performance on VOC2007 and VOC2012 datasets

Object detection algorithm	Year	$mAP^{0.5}$	FPS
DPM	2008	34.3	0.5
R-CNN	2013	58.5	–
Fast R-CNN	2015	70.0	0.5
Faster R-CNN	2015	73.2	7.0
YOLOv1	2015	63.4	45.0
SSD	2015	72.1	58.0
YOLOv2	2016	78.6	40
YOLOv3	2018	82.29	45.0

**Fig. 1** Working principle of RPN network



the model. The model based on LCNN proposed in this paper improves expression ability of pedestrians and enhances the effect of pedestrian recognition.

- Because the pedestrian re-identification technology is relatively rare in real scenes, and the real scenes data is complex and varied, there are many small-scale pedestrians that affect effective pedestrian feature extraction. The existing pedestrian re-identification models used in real scenes need to be improved. In the proposed method, the data augmentation-based YOLOv3 detector is combined with the pedestrian re-identification model based on the LCNN method, and the image size is scaled-up for pedestrians whose size is less than a specified value. The end-to-end framework proposed has improved robustness and accuracy.

The rest of this paper is arranged as follows: Sect. 2 introduces the related work and motivation for the proposed method; Sect. 3 introduces the end-to-end pedestrian detection and re-identification applied to the real scene; Sect. 4 is the experimental section; and in Sect. 5 we summarize the article.

## 2 Related work

The application of pedestrian detection is extremely extensive and is one of the most important research topics in object detection. The object detection algorithm consists in an object detection algorithm based on the traditional method and one based on deep learning. Felzenszwalb et al. [42] proposed the DPM algorithm in 2008, which is a component-based detection method. The DPM algorithm is generally better than traditional algorithms, but the deep learning-based object detection algorithm is superior in speed and accuracy. At present, there are two main types of deep learning-based object detection algorithms. One of them divides the object detection task into two stages, two-stage methods such as the bounding box algorithm such as Faster

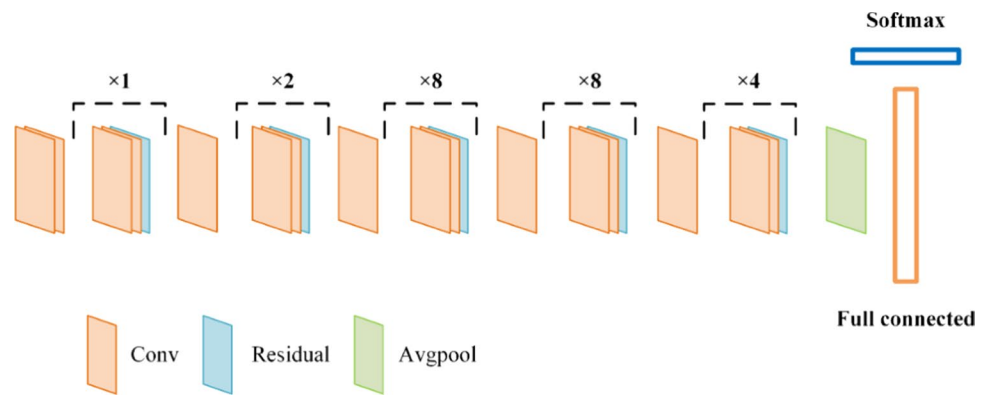
R-CNN, and the other treats the object detection algorithm as an end-to-end task, such as the YOLO series algorithm and the SSD algorithm. The specific performance of the object detection algorithm on the VOC2007 and VOC2012 data sets is shown in Table 1.

It can be seen from Table 1 that the deep learning-based object detection algorithm performance is much higher than the traditional method-based algorithm. At the same time, the performance of the mean average precision (mAP) value based on the deep learning algorithm is at least 24% higher than that of the DPM algorithm<sup>2</sup>. The performance of the YOLOv3 algorithm is more than double the mAP of the DPM algorithm [2]. It can also be seen that the YOLO series algorithm is superior to the SSD algorithm in overall performance. At the same time, The mAP value of the YOLOv3 algorithm is higher than 10% that of the SSD algorithm. Therefore, this section will focus on the Faster R-CNN algorithm and the YOLO series of algorithms.

### 2.1 Faster R-CNN

Faster R-CNN is a series of object detection frameworks and is another major improvement to the object detection model along with R-CNN, SPP-Net, and Fast R-CNN. Girshick et al. [25] used the Selective Search [26] method for regional nomination operations, CNN for feature extraction operations, SVM algorithm for classification, and finally, border regression calculations. This method was trained on the VOC dataset. The mAP value is closer than that of the traditional DPM algorithm, but the training procedure is extremely cumbersome, the training and testing speed is relatively low, and the space is relatively large. He et al. [27] proposed the use of SPP-Net algorithm, but due to its spatial characteristics, the convolution layer could not be fine-tuned. Therefore, Girshick et al. [28] proposed Fast R-CNN in 2014, and the Selective Search method was used for regional nomination operations. After that, the features are extracted by CNN and classified by the Softmax algorithm.

**Fig. 2** Basic Darknet53 network structure



**Table 2** Pedestrian test results of different methods on INRIA, VOC mixed data set

Method	Input	Transmission/f·s <sup>-1</sup>	mAP/%
Fast R-CNN(VGG16)	512×512	0.57	63.65
Faster R-CNN(VGG16)	512×512	12.63	76.87
YOLO	512×512	68.45	59.53
YOLOv2	512×512	115.06	70.12
YOLOv3	512×512	49.73	90.95

This method introduces the RoI Pooling Layer, which enables the previous convolution layer to be fine-tuned.

Instead of using the original Selective Search for regional nomination operations, Faster R-CNN has re-proposed a new regional nomination network, called region proposal networks (RPN), which further enhances the model’s effectiveness. The RPN structure is shown in Fig. 1.

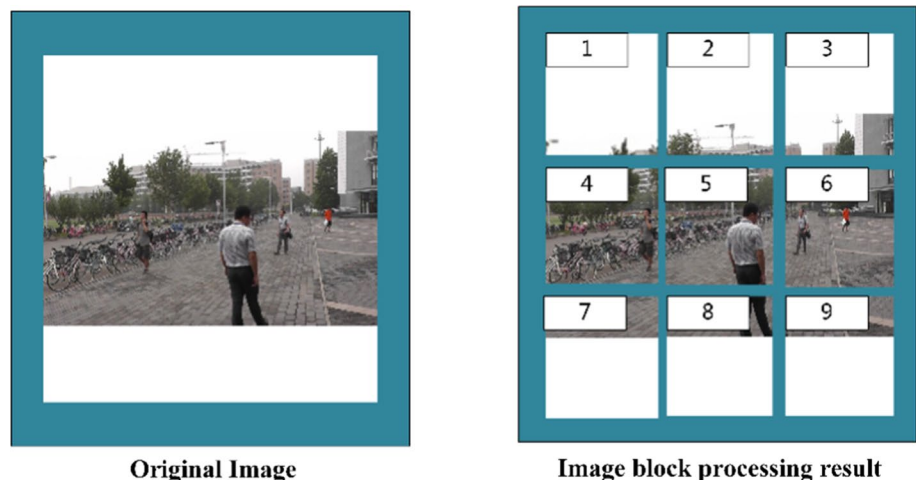
### 2.2 YOLOv3

Unlike the region-based CNN and Faster R-CNN described in the previous section, the YOLOv3 is a regression-based

CNN model that relies on the Darknet [29]. However, similar to Faster R-CNN, YOLOv3 is also an improved network model based on YOLOv1 and YOLOv2. Compared with Faster R-CNN, YOLO does not show the process whereby it obtains a region proposal. The extraction network and the Faster R-CNN share the convolutional layer, but, in the model training process, it is necessary to repeatedly train the area extraction network and the object classification network. Although the accuracy is improved, the speed is reduced. However, YOLO is an end-to-end object detection algorithm, so entering a picture can get the confidence of location and category at the same time, which significantly increases the speed.

YOLOv1 is a regression problem-based object detection algorithm, which is pre-trained on pictures with a resolution of 224×224. In formal training, YOLOv1 first normalizes the original input picture to a size of 448×448, and then divides the picture into S×S grid cells. The division is logical, i.e., the image itself is not divided. For cropping, if the center of the touch object falls in this grid, then this grid is responsible for predicting the target object. Each grid is responsible for predicting B bounding boxes. For each bounding box, confidence regression is performed as well

**Fig. 3** Data augmentation processing



as position regression, and  $C$  category probabilities are predicted. The predicted position information is the offset of the center position of the object relative to its original position and its width and height. The value of the confidence is not only the probability of the object to be detected in the bounding box, but also the probability of the object to be detected in the bounding box multiplied by the value of the bounding box and the real position, which is mathematically expressed as follows:

$$IOU_{pred}^{truth} = \frac{\text{area}(\text{box}_{pred} \cap \text{box}_{truth})}{\text{area}(\text{box}_{pred} \cup \text{box}_{truth})} \quad (1)$$

where  $\text{box}_{pred}$  is the prediction box and  $\text{box}_{truth}$  is the ground truth box. Therefore, the confidence reflects the accuracy of the position in the bounding box and the position of the containing object, and is expressed as:

$$\begin{aligned} \text{Confidence} &= P_r(\text{class}|\text{object}) \times P_r(\text{Object}) \times IOU_{pred}^{truth} \\ &= P_r \times IOU_{pred}^{truth}, P_r(\text{class}) \in \{0,1\} \end{aligned} \quad (2)$$

$P_r(\text{object})$  represents the probability of an object in the current box.  $P_r(\text{class}|\text{object})$  is the conditional probability of the predicted  $C$  categories. YOLOv3 no longer uses the mean square error function to calculate the loss during the training process but instead uses the binary cross entropy function.  $y$  is the ground truth and  $o$  is the predicted value. The formula is then:

$$L_{MSE} = \frac{1}{2}(o - y)^2 \quad (3)$$

YOLOv3 uses the multi-scale fusion method for prediction. It uses up-sampling and a fusion method similar to feature pyramid networks (FPN) for object detection on multiple scale feature maps. In terms of network structure, YOLOv3 adds a hybrid mode of residual network to the network structure, changing the layer-by-layer training to phase-by-stage training, and dividing the deep neural network into several sub-phases, each of which contains a shallow network, and then uses the shortcut method to connect each sub-phase, which then learn a part of the total difference, and finally reaches the overall small loss. This method can also avoid gradient diffusion or explosion, which is not conducive to training. The network is expanded to 53 layers (darknet-53). Its network structure is shown in Fig. 2.

Many researchers have used YOLOv3 for pedestrian recognition. In [43–45], the pedestrian recognition methods proposed by the authors are all improved based on YOLOv3. Table 2 shows the quantitative pedestrian test results of different detection methods on INRIA, VOC mixed data set. It can be seen that YOLOv3 works very well in pedestrian detection. Therefore, we use yolov3 as the detection part of the pedestrian re-identification method.

### 2.3 Data augmentation-based YOLOv3

At present, the most common reason for the miss detection of pedestrians is the insensitivity of the detector to small objects. There are also cases of false detection due to the detection of non-human objects with structures similar to the human body. To reduce this issue, we perform data augmentation to the original image. Although the YOLOv3 detector is very sensitive to small objects, when the object is smaller than a certain size, the object is obscured, and in a low-illumination condition, the detector may not successfully detect the object.

#### 2.3.1 Non-overlapping image blocking data augmentation method

In this study, we employ a non-overlapped image blocking method to enhance the images. The segmentation of the image can compensate for uneven illumination and enlarge the original distant object, thereby enhancing the sensitivity of the detector.

Suppose the original image  $I$  is an  $M \times N$  image matrix, which is divided into a  $p \times q$  image block matrix. The results of the division are as follows:

$$I = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1q} \\ I_{21} & I_{22} & \cdots & I_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ I_{p1} & I_{p2} & \cdots & I_{pq} \end{bmatrix} \quad (4)$$

The results of image augmentation processing obtained in the experiments are shown in Fig. 3. The specific data augmentation processing is as follows:

**Algorithm 1:** Non-overlapping Image Blocking Data Augmentation Algorithm**Input:** pedestrian detection image  $I_0$ **Output:** Nine  $3 \times 3$  sub-images obtained from the original image  $I_0$ 

1. Read pedestrian detection image  $I_0$
2. Get the size of the pedestrian detection image:  $w$  and  $h$
3. Fill the original image with white edges to obtain a square image
4. The larger  $w$  and  $h$  are selected as the size of the final image, i.e., if  $w > h$ , the final size is  $w \times w$ , otherwise, the size is  $h \times h$
5. Crop the image according to the size of the final image, and get 9 sub-images of size  $n = \frac{l}{3}$

### 2.3.2 Data augmentation post processing based on non-maximum suppression

Because the pedestrians detected in the image may not belong to the same sub-image, e.g., in the first and third block in image block processing result of Fig. 3, the upper body and legs of the pedestrian are respectively assigned, such as sub-image 5 and sub-image 8. However, the upper body and legs of the pedestrians will still be categorized as pedestrians by the detector, so we used the detection results of the data augmentation-based YOLOv3 and the detection results obtained using YOLOv3 as the detector, and non-maximum suppression (NMS) is used to obtain the final detection result. Non-maximal suppression is a very common and effective method in the field of object detection and recognition. NMS offers the following three solutions:

1. Select the intersection of a number of bounding boxes, i.e., select the common area of these boxes as the final object area.
2. Select the multiple rectangular boxes, and the ratio of the intersection of the rectangular boxes to the minimum detection box. If the ratio is greater than a set threshold,

select the smallest outer-cut rectangle of all the selected rectangular boxes as the final object area.

3. As the candidate box for intersection, select the one with the highest confidence level as the final object detection box.

In this study, the third solution is selected, and the confidence of the detected object box and the value of the IOU between the two boxes are considered to select the final candidate box.

### 2.3.3 Data augmented results

The specific process of data augmentation is the following:

1. Input all the object frames detected by the two methods in one image and compare IOU between any two boxes.
2. If the IOU value exceeds a certain threshold and the confidence of box 1 is greater than the box 2, select the candidate box with the highest confidence as the final detection box.

The final data augmentation effect obtained from the process of image blocking described above is shown in Fig. 3.

**Table 3** Performance evaluation of face recognition models

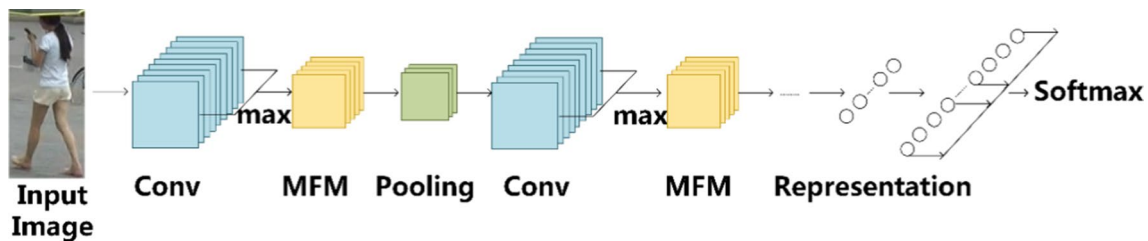
Face recognition technology	Accuracy (LFW data-sets [36])	Training data set	Speed
DeepFace [30]	97.35%	Social Face Classification Dataset (SFC)	220 ms
DeepID1	97.45%	CASIA-Webface [37]	–
DeepID2	99.15%	CASIA-Webface	–
DeepID2+ [31–33]	99.47%	CASIA-Webface	35 ms (GPU)
FaceNet [34]	99.63%	MS-Celeb-1M	–
LCNN_TRAIN [35]	99.33%	CASIA-Webface	67 ms

**Table 4** LCNN Network structure

Name	Filter Size/Stride, Pad	Output Size	#param
Input	–	144 × 144 × 1	–
Crop	–	128 × 128 × 1	–
conv1_1	5 × 5/1,2	128 × 128 × 48	1.2 K
conv1_2	5 × 5/1,2	128 × 128 × 48	1.2 K
mfm1	–	128 × 128 × 48	–
pool1	5 × 5/2	64 × 64 × 48	–
Conv2_a	1 × 1/1	64 × 64 × 48	0.04 K
Conv2_1	3 × 3/1,1	64 × 64 × 96	0.8 K
Conv2_2	3 × 3/1,1	64 × 64 × 96	0.8 K
mfm2	–	64 × 64 × 96	–
pool2	2 × 2/2	32 × 32 × 96	–
Conv3_a	1 × 1/1	32 × 32 × 96	0.09 K
Conv3_1	3 × 3/1,1	32 × 32 × 192	1.7 K
Conv3_2	3 × 3/1,1	32 × 32 × 192	1.7 K
mfm3	–	32 × 32 × 192	–
pool3	2 × 2/2	16 × 16 × 192	–
Conv4_a	1 × 1/1	16 × 16 × 192	0.19 K
Conv4_1	3 × 3/1,1	16 × 16 × 128	1.1 K
Conv4_2	3 × 3/1,1	16 × 16 × 128	1.1 K
mfm4	–	16 × 16 × 128	–
pool4	2 × 2/2	8 × 8 × 128	–
Conv5_a	1 × 1/1	8 × 8 × 128	0.12 K
Conv5_1	3 × 3/1,1	8 × 8 × 128	1.1 K
Conv5_2	3 × 3/1,1	8 × 8 × 128	1.1 K
mfm5	–	8 × 8 × 128	–
Pool5	2 × 2/2	4 × 4 × 128	–
fc1	–	256	524 K
fc2	–	10,575	2707 K
Loss	–	10,575	–

## 2.4 LCNN-based pedestrian re-identification

In Sect. 2.3 we introduced the data augmentation-based YOLOv3 pedestrian detection method. Pedestrian recognition technology relies on pedestrian detection technology in the early stage. The performance of pedestrian detection technology will directly affect the performance of pedestrian recognition technology. Therefore, in this section we

**Fig. 4** LCNN framework

discuss a LCNN-based pedestrian recognition method based on Sect. 2.3.

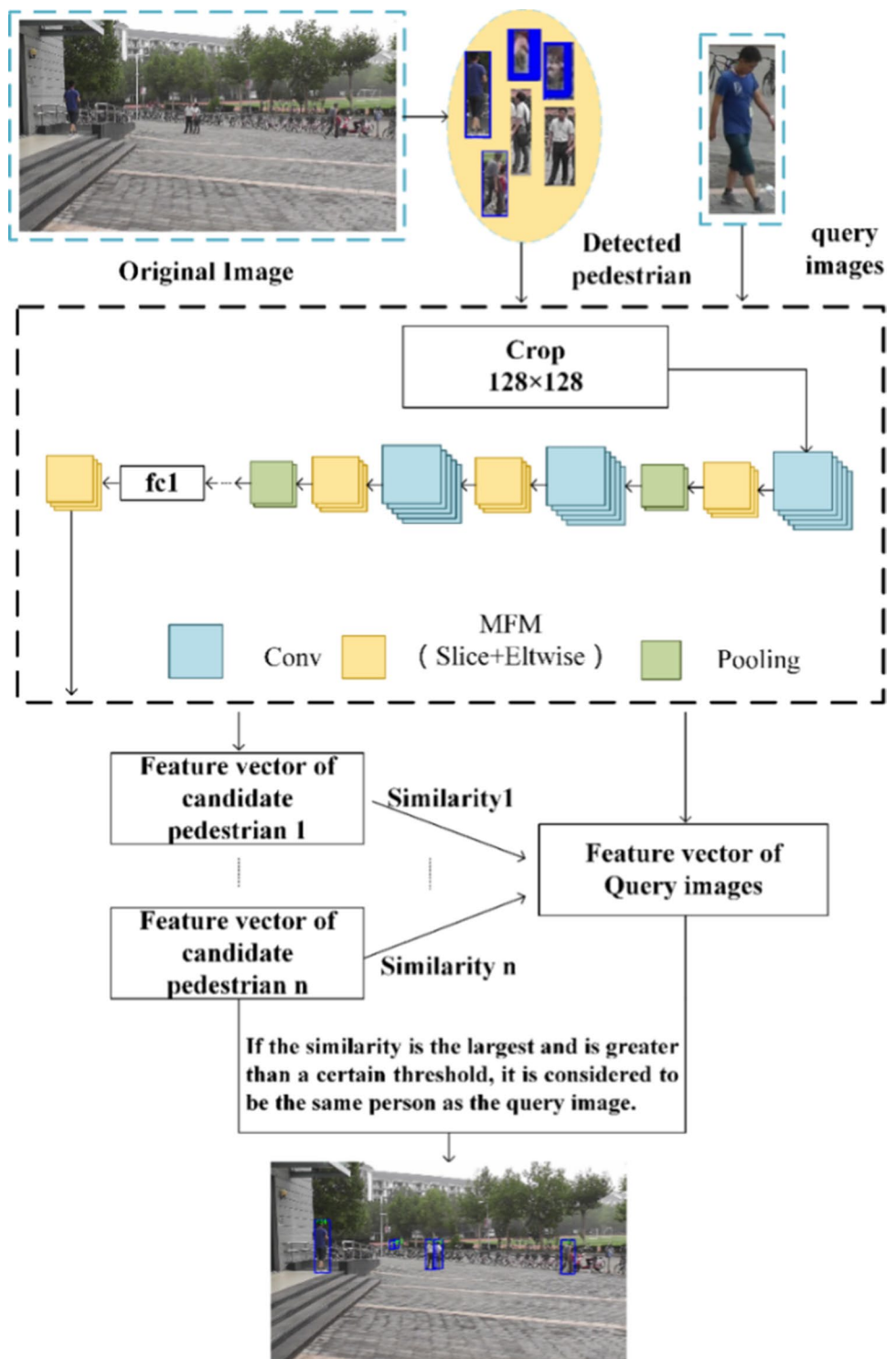
### 2.4.1 CNN structure and performance analysis

Since AlexNet was proposed by Krizhevsky et al. [5] in 2012, CNN have found a wide range of applications, such as in object detection and recognition and image classification, but as people's requirements for performance gradually increase, better performing CNNs have emerged, such as ZF-Net, VGGNet, GoogLeNet, and ResNet. However, while these models achieve better results, the number of layers in the network has increased, from 7 of the AlexNet to 16 of the VGGnet, 22 of the GooLeNet, and 152 of the ResNet, and some even reach thousands of layers of networks, such as DenseNet. The increase in the number of layers makes the network's performance better, but it takes up more storage space and become slower.

However, the LCNN can compress the model while ensuring its stability, thereby reducing the model parameters and the number of network layers, reducing the space occupancy rate and improving the efficiency. At present, LCNNs have also been successfully applied to face recognition. A specific performance evaluation of LCNNs and other models on large data sets of face recognition is shown in Table 3.

On the labeled faces in the wild home (LFW) data set, the performance of LCNN\_TRAIN ranks third, with an accuracy rate reaching 99.33%, which is only 0.3% lower than the that of the FaceNet algorithm, which ranks first. However, the training data set used by the FaceNet network has 10 million photos and 100,000 human faces, while the LCNN\_TRAIN algorithm uses 13,000 photos and 5,000 human faces. Furthermore, LCNN\_TRAIN has a small network structure, relatively few parameters, low space occupancy, and higher speed compared to other algorithms. The LCNN\_TRAIN algorithm has been successfully applied to face recognition problems. First, the face classification task is used to train the CNN to extract the face features, and the classifier is then used to determine whether the current and the target pedestrian are the same person, which is very similar to the process of the pedestrian recognition we perform. The structural features of the face are similar to those of the

**Fig. 5** Pedestrian re-identification process



pedestrian. Therefore, we decided to select the network in the LCNN\_TRAIN algorithm as the feature extractor.

### 2.4.2 LCNN

The previous section introduced the LCNN. We mentioned that the LCNN reduces the size of the space occupancy

and improves the recognition efficiency. In this section, we introduce the LCNN-based pedestrian re-identification framework.

LCNN is a network for face recognition proposed by Wu Xiang in 2016. The model is a lightweight structure, and, although the model is small, a very good recognition rate can be obtained. The forward calculation is fast and can be



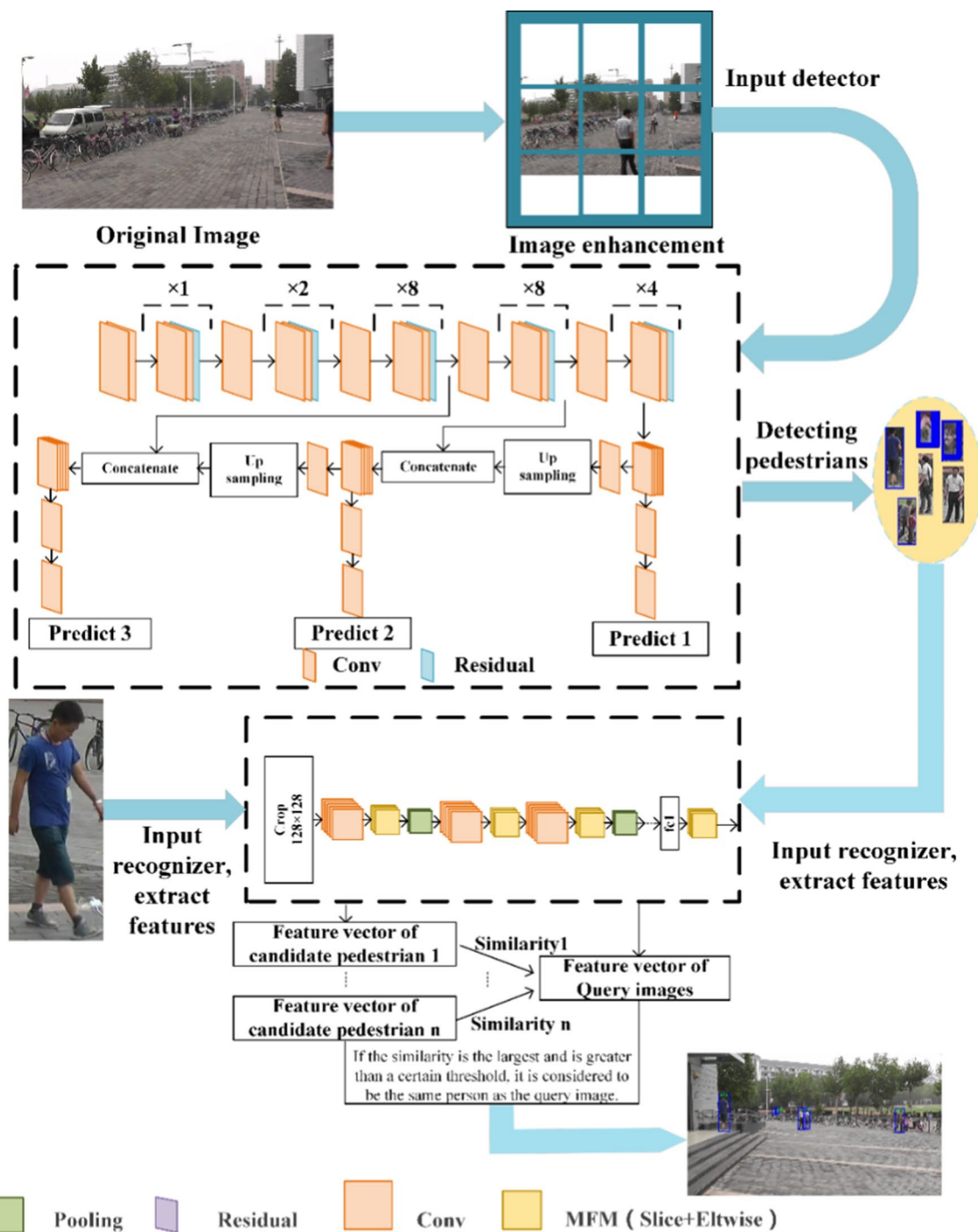


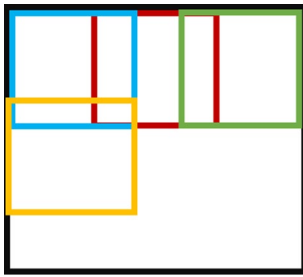
Fig. 6 End-to-end pedestrian detection and re-identification process

used on embedded devices, both time and space are well optimized, and the face features extracted by the model are well differentiated. The new activation function Max-Feature-Map (MFM) is applied to the convolutional layer of the network. Given an input convolution layer  $C \in R^{h \times w \times 2n}$ . The MFM activation function is based on maxout networks, and is defined as:

$$f_{ij}^k = \max_{1 \leq k \leq n} (c_{ij}^k, c_{ij}^{k+n}) \tag{5}$$

And the gradient of the activation function is:

$$\frac{\partial f}{\partial c^k} = \begin{cases} 1, & \text{if } c_{ij}^k \geq c_{ij}^{k+n} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$



**Fig. 7** Overlapping blocking method

where,  $1 \leq k' \leq n$ , and

$$k = \begin{cases} k' & 1 \leq k' \leq n \\ k' - n & n + 1 \leq k' \leq 2n \end{cases} \quad (7)$$

Thus, half of the active layer has a gradient of 0, and the MFM activation function can get a sparse gradient. Compared with the ReLU function, which obtains the features of sparse high-dimensional features, the MFM function can obtain compact features, and can perform feature selection and dimensionality reduction.

The LCNN network occupies less space and has a higher degree of refinement, and it has successfully been applied in face and vehicle recognition. The network structure of LCNN is shown in Table 4. The LCNN has 29 layers. After inputting, the picture is normalized to  $128 \times 128$  px, and then processed by convolution and MFM activation function. The size is selected and reduced, and the image goes through the pooling layer to reduce the over-fitting. One this process is carried out 5 times, the image enters the two fully connected layers. The specific process is shown in Fig. 4.

The small structure and high resolution characteristic of the LCNN can produce a low-dimensional feature expression vector that can be used for fast calculations, so that the identity of the pedestrian can be well recognized.

The network-in-network (NIN) sub-network is also designed in the LCNN model, and the linear relationship between the convolutional layer and the pooled layer becomes a nonlinear relationship of the  $1 \times 1$  convolutional layer. The NIN can not only fit any function, because it is essentially a small and fully connected neural network. Although, the NIN model uses only one-tenth of the AlexNet network parameters, the final result is superior.

NIN uses a multi-layer perceptron to improve the generalization ability and the ability to extract abstract features, and it replaces the most connected fully connected layer with the global average pooling layer, which reduces the size of the network and its effectiveness. To avoid model overfitting, the global average pooling layer performs global mean pooling on each feature map, so that an output for each feature map can be obtained. The integrated processing of

cross-character maps in NIN can help the model to obtain complex and practical cross-characteristic features.

### 2.4.3 LCNN-based pedestrian re-identification

Pedestrian re-identification consists of mainly two steps. One is extracting the pedestrian features with good robustness and strong representation ability from the image, and the second is measuring the similarity between a pedestrian's and the target pedestrian's features in the image library. If a certain condition is met, the current pedestrian and the target pedestrian are considered to be the same person, otherwise, they are considered different. The process of pedestrian recognition is shown in Fig. 5.

The pedestrian re-identification model involves the following steps: inputting the picture of the candidate pedestrian and the image of the object figure detected from the original image via the detector to the LCNN. In the feature extractor, the pedestrian image is first normalized to a size of  $128 \times 128$  px, then passed through the convolution layer, then into the Max-Feature-Map (MFM) activation function layer, and the pooling layer. After several such convolution, activation, and pooling operations, the image enters the fully connected layer and the MFM layer, and gets assigned the eigenvectors of the candidate pedestrians and the target pedestrians by the feature extractor. The eigenvectors of the candidate pedestrians and the target pedestrians are then calculated for similarity. The similarity metric algorithm used is the cosine distance. After all the similarity values are obtained, and the candidate pedestrian with the largest similarity value is selected, if the similarity is greater than a set value, the pedestrian is considered the same as the target pedestrian.

## 3 Proposed method

Sections 2.1 and 2.2 of this paper described the deep learning-based pedestrian detection and Sect. 2.4 described the LCNN-based pedestrian recognition method and verifies the YOLOv3 detector based on data augmentation and the LCNN. To re-identify the validity of the model, in this section, we integrate the pedestrian detection and pedestrian recognition tasks, construct an end-to-end pedestrian detection and re-identification model, and use the real data set to verify the validity of the model. Furthermore, we use data augmentation methods to improve the end-to-end pedestrian detection and re-identification models.

**Table 5** Real scenes image scaling processing

Original image								
Image after size scaling								

### 3.1 End-to-end pedestrian detection and re-identification overall process in real scenes

In this paper, the data augmentation-based YOLOv3 detector and the lightweight re-convolution neural network proposed in Sect. 2 are used to develop an end-to-end pedestrian detection and re-identification model applied to real scene is proposed. The model is shown in Fig. 6.

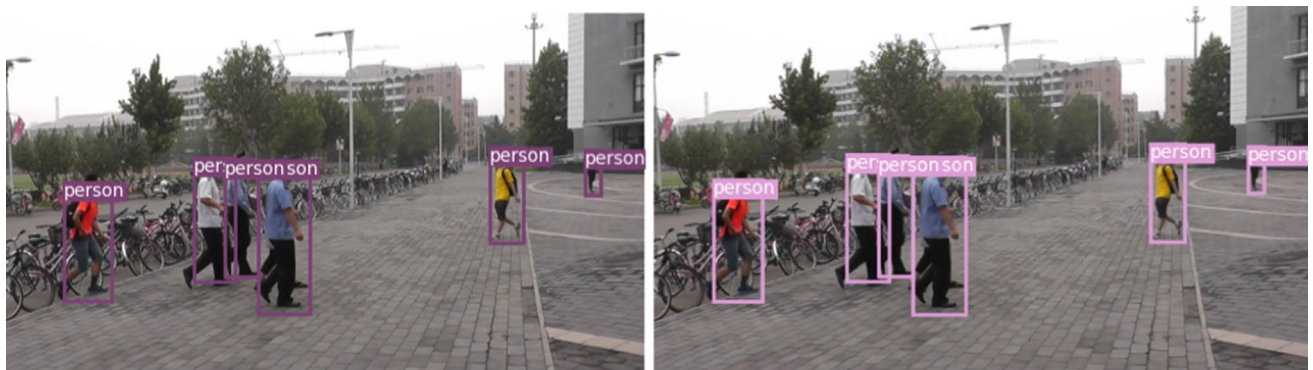
The main steps of the end-to-end pedestrian detection and re-identification model in real scenes are:

1. The original image is subjected to data augmentation processing. The data augmentation method is based on image blocking, which first fits the image into squares depending on their length or width, then divides it into  $3 \times 3$  blocks, and locally enlarges them to increase the sensitivity of the detector to small-sized pedestrians.
2. Input all the augmented sub-images in step 1 to the detector based on the YOLOv3 method, in which three multi-scale predictions are performed and the final prediction result is obtained, but because the detector is

based on the data augmentation of the YOLOv3 detector, it is necessary to first restore the coordinates of the prediction result. Subsequently, because the image augmentation method of image blocking divides a pedestrian into multiple subgraphs, the detector will detect the features separately. Thus, every part of the body is treated as a pedestrian, and we need to use the detection results of the YOLOv3-based detector to solve the problem by NMS suppression. Through a series of operations, we can finally get the detection results by the YOLOv3 detector based on data augmentation.

**Table 6** Performance evaluation of overlapping and non-overlapping blocking method on the PRW data set

Method	map@0.5 (%)
Non_overlapping	79.32
Overlapping (100 px)	79.32
Overlapping (300 px)	79.32



**Fig. 8** Comparison of overlapping and non-overlapping blocking method test results

3. We input the pictures of the pedestrians detected in step 2 and target pedestrians into the LCNN, obtain the feature vectors of the candidate pedestrians and the target pedestrians respectively, and then calculate the similarity using the cosine distance-based metric algorithm. The calculated threshold determines whether the current candidate pedestrian is the target pedestrian. The more detailed algorithm flow is as follows:

adopted a data augmentation method to amplify small-sized pedestrians.

The interpolation algorithm is a basic but extremely important in image scaling. Different interpolation algorithms have different precisions and their performance directly affect the distortion of the image. Interpolation algorithms include nearest neighbor interpolation, bilinear interpolation, cubic convolution interpolation, and

---

**Algorithm 2:** End-to-End Pedestrian Detection and Re-identification Algorithm

**Input:** pedestrian detection image  $G_0$

Target pedestrian image  $D_0$

Similarity judgment threshold  $t$

**Output:**  $N$  is a set of candidate boxes that meet the final criteria.

---

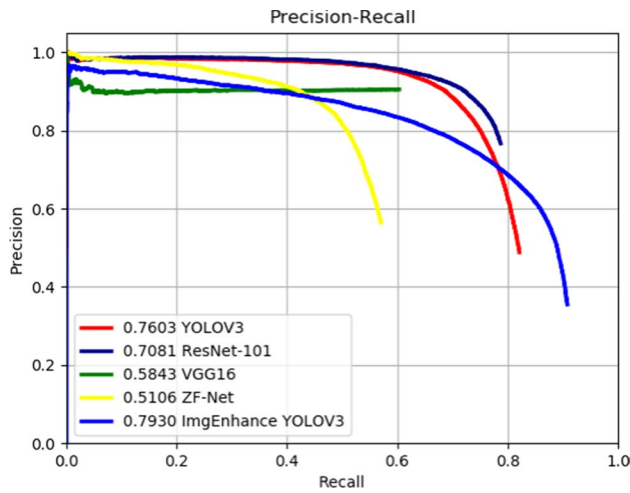
1. Read pedestrian detection image  $G_0$
  2. Get the size of the pedestrian detection image:  $w$  and  $h$
  3. Fill the original image with white edges to obtain a square image, Crop the image according to the size of the final image, and get 9 sub-images  $G_1, G_2, \dots, G_9$ .
  4. build an empty set for Candidate boxes detected by YOLOv3 :  $S \leftarrow \emptyset$ ;
  5. for each image  $i \in [0, 9]$  do
  6. send  $G_i$  to YOLOv3 detector,  $S \leftarrow$  a set of anchor boxes
  7. end for
  8. Use NMS to eliminate the redundant boxes in  $S$  and save the results in the set  $H$
  9. All candidate anchor boxes in set  $H$  and  $P_0$  are cropped to  $128 * 128$ , and sent to LCNN to extract features. Get  $1 + \text{length}(H)$  128 dimensional eigenvectors.
  10. The similarity between  $D_0$  and candidate boxes in set  $H$  is calculated by cosine distance measurement, get the length ( $H$ ) scores and save the results in the set  $Q$
  11. build an empty set for Candidate boxes that meet the requirements :  $N \leftarrow \emptyset$ ;
  12. for each candidate  $q \in Q$  do
  - 13: if  $q \geq t$  then
  - 14: Save the candidate box in the corresponding  $H$  to  $N$
  - 15: end if
  - 16: end for
  17. return  $N$
- 

### 3.2 Pedestrian image scaling in real scenes

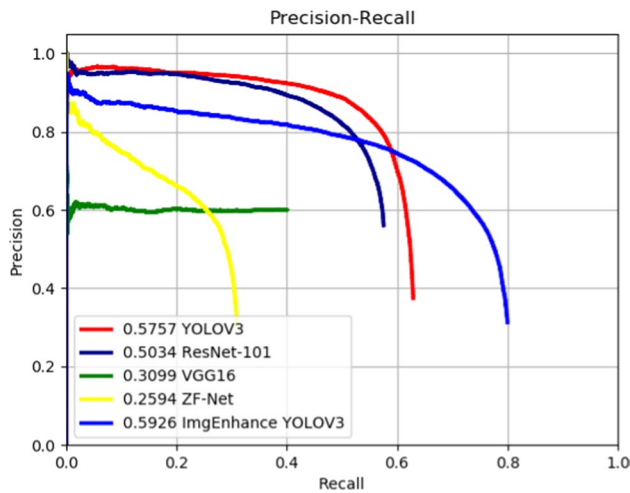
To better extract pedestrian characteristics and better distinguish target pedestrians and non-target pedestrians, we

pixel-relational resampling, with the first three being the most commonly used.

In the process of scaling the image, there is no one-to-one correspondence between the pixels of the output and input



**Fig. 9** Precision-Recall curve of each detector at  $IOU > 0.5$  from the PRW data set



**Fig. 10** Precision-Recall curve of each detector at  $IOU > 0.7$  from the PRW data set

images. Therefore, it is necessary to consider the output image, and use reverse mapping, whereby the pixel points in the input image corresponding to the output image are found from the output image, and, if it is considered the input image, some pixels in the output image may have no gray value.

The nearest neighbor algorithm, also known as zero-order interpolation, is the simplest, and involves mapping the gray value of the closest pixel of the input image to the pixel value of the output image. Although the algorithm is relatively simple, it is easy to cause the image to be jagged.

The bilinear interpolation algorithm, also known as first-order interpolation, is an improvement to the nearest neighbor interpolation algorithm. As the name implies, the

**Table 7** Performance evaluation for each detector on the PRW data set

Detector	$mAP$	
	$IOU > 0.5$ (%)	$IOU > 0.7$ (%)
DPM	70.2	45.1
DPM_Alex	75.5	59.1
ACF_Res	76.3	34.7
ACF_Alex	75.9	33.9
ACF	71.5	33.2
LDCF	76.4	36.2
YOLOv3_darknet-53	76.03	57.57
Faster R-CNN_ZF	51.06	25.94
Faster R-CNN_VGG16	58.43	30.99
Faster R-CNN_ResNet-101	70.81	50.34
Image Augmenta- tion + YOLOV3_darknet-53	79.30	59.26




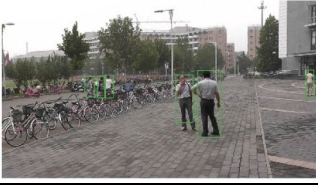





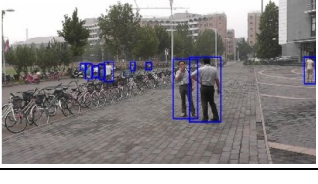
algorithm performs three interpolation calculations to obtain the final result. It is relatively computationally intensive, but the quality of the image after scaling using the interpolation algorithm is relatively high, and there is no pixel discontinuity. However, since the interpolation algorithm contains the characteristics of a low-pass filter, it may cause the image to become blurred.

The cubic convolution interpolation algorithm, also known as Bicubic interpolation, is an improvement on

**Table 8** Performance comparison of each detector and recognizer combination on the PRW data set. The results are obtained by selecting the optimal value under the same conditions

Detector	Recognizer	$mAP^{0.5}$
DPM	IDE	13.7
DPM	IDE <sub>det</sub>	18.8
DPM-Alex	IDE	14.8
DPM-Alex	IDE <sub>det</sub>	20.3
DPM-Alex	IDE <sub>det</sub> + CWS	20.5
ACF	IDE <sub>det</sub>	17.5
ACF-Res	IDE	12.5
ACF-Alex	IDE <sub>det</sub>	17.5
ACF-Alex	IDE <sub>det</sub> + CWS	17.8
LDCF	IDE	34.4
LDCF	IDE <sub>det</sub>	18.3
LDCF	IDE <sub>det</sub> + CWS	18.3
ZF	Lightened CNN + cosine	27.99
VGG16	Lightened CNN + cosine	34.83
ResNet-101	Lightened CNN + cosine	52.87
Ground Truth	Lightened CNN + cosine	52.86
YOLOv3	Lightened CNN + cosine	52.90
Data augmentation-based YOLOv3	Lightened CNN + cosine	<b>53.21</b>

**Table 9** Comparison of manual labeling results and other detectors on the PRW data set

Original image			
Ground Truth			
Faster R-CNN	ResNet-101		
YOLOv3	Darknet-53		
Data augmentation-based YOLOv3	Darknet-53		

the bilinear interpolation algorithm. It solves the issues of the two abovementioned algorithms and its accuracy is extremely high, but the calculation is still large. In addition to considering the gray values of the four adjacent pixels, the algorithm also considers the rate of change of the gray values of these pixels.

In this study, the higher precision cubic convolution interpolation algorithm is used to scale the image. The results of image augmentation are shown in Table 5.

## 4 Experimental methods

### 4.1 Experiment setup

For the experiments we used the Linux Ubuntu 16.04 operating system, the Tesla K80 graphic card, a 128 GB memory, Python and C++ as programming languages, and the deep learning common framework Darknet and Caffe to construct the CNN structure. The pedestrian detection and re-identification data sets mainly include PRW large public data sets and pedestrian data in real scenes that we

have photographed. The PRW data set is collected by Liang Zheng et al., which uses 6 cameras. At Tsinghua University, 5 of the 6 cameras were 1080 × 1920 HD, and the remaining one was a 576 × 720 SD. The ID numbers marked in the data are 1 ~ 932 and - 2. The pedestrians marked as - 2 will be used in the pedestrian detection part of the previous period, and the pedestrian re-identification part will be eliminated in the later stage. There are 42 target pedestrians in the real scenes data, and 174 complete image frames containing the real scenes. The pedestrians in the collected data are captured at different angles and the scenes contain different levels of illumination.

We use common metrics for general classification, namely Precision, Recall, and Mean Average Precision, which are calculated as follows:

For each category  $y_j$ , the accuracy  $P^j$  and recall  $R^j$  of the test data set, where  $j \in \{1, \dots, N\}$ , are expressed as:

Precision( $y_j$ ) indicates the number of categories  $y_j$  that are correctly predicted,  $Prediction(y_j)$  indicates the total number of categories  $y_j$  predicted, and  $Ground(y_j)$  indicates the number of categories  $y_j$  that are manually labeled in the data set. By averaging the accuracy and recall rates over all

**Table 10** Comparison of the recognition results of the models *ImgEn\_YL* and *Real\_ImgEn\_YL*

Target person		<i>ImgEn_YL</i>	<i>Real_ImgEn_YL</i>
Back			
		Correct identification	Correct identification
Front			
		Unrecognized	Correct identification
Side			
		Misidentification	Correct identification
Occlusion			
		Correct identification	Correct identification
Smaller person			
		Misidentification	Correct identification

categories, the average accuracy rate  $P$  and the average recall rate  $R$  are obtained.



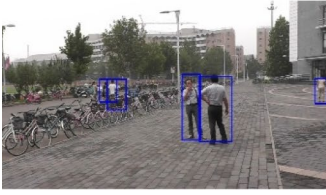
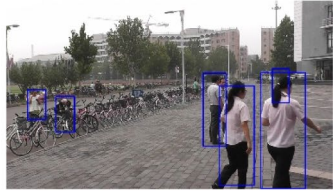
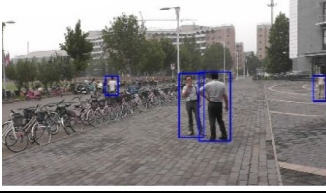
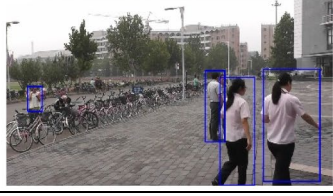
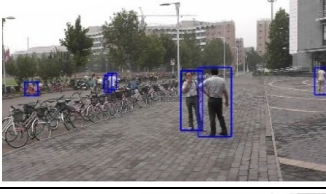
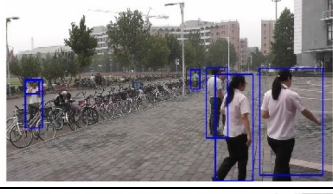
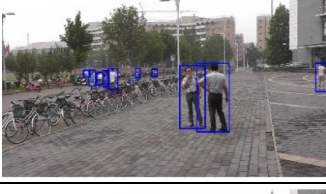
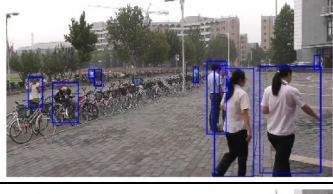
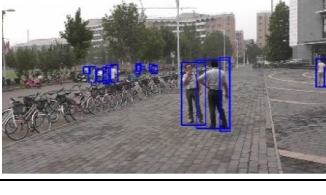
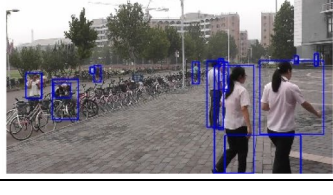
$$P = \frac{1}{M} \sum_{j=1}^M P^j \tag{10}$$

$$R = \frac{1}{M} \sum_{j=1}^M R^j \tag{11}$$

Ideally, the higher is the accuracy of the test results, the higher the recall rate will be, but, in reality, the two are somewhat contradictory. In extreme cases, there will be an accuracy rate of 100% with an extremely low recall rate, or with all targets retrieved, or there might be a recall rate of 100% with an extremely low accuracy rate. Therefore, in the experimental study, we need to use the Precision-Recall curve to analyze the performance of the model.

The area under the Precision-Recall curve is the average precision (AP) value, from which the average recall rate can be obtained. A Precision-Recall curve can be drawn for

**Table 11** Test results of different detectors on the PRW data set

Original image			
Faster R-CNN	ZF-Net		
	VGG16-Net		
	ResNet-101		
YOLOv3	Darknet-53		
Data augmentation-based YOLOv3	Darknet-53		

every category, from which the corresponding AP value can be obtained. The AP ranges from 0 to 1 and higher AP values, lead to a better detection classification of the model. We are pursuing an ideal model where the AP value is 1. By averaging the sum of the AP values of the categories and the number of categories, a mean average precision (mAP) value can be obtained. The mAP value can better characterize the performance of detectors and recognizers in pedestrian detection and re-identification, and it can be calculated as follows:

$$mAP = \int_0^1 P(R)dR \tag{12}$$

## 4.2 Estimation results

### 4.2.1 Validity of the data augmentation-based YOLOv3 pedestrian detection method

We used the overlapping image blocking augmentation method for the experiment. The Fig. 7 shows the blocking method we used:

We made an experimental comparison on the overlapping and non-overlapping image blocking augmentation method, and the experimental results are shown in Table 6.

We set multiple overlapping pixel values, and the experimental results show that the result of setting 0 on the data set



**Table 12** Test results of different detectors using partial enlargement method on the PRW data set

Original image			
Partial enlargement			
Faster R-CNN	ZF-Net		
	VGG16 Net		
	ResNet-101		
YOLOv3	Darknet-53		
Data augmentation-based YOLOv3 based	Darknet-53		







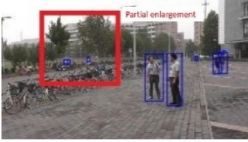








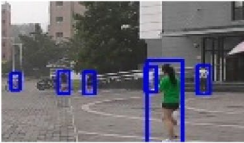
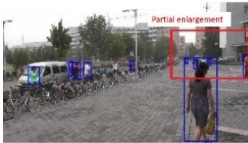

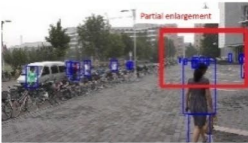
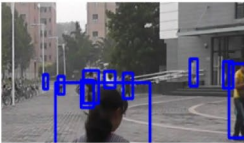
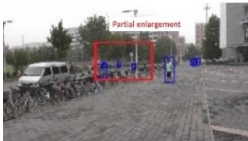

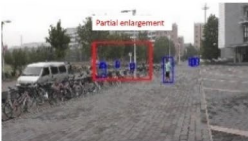

PRW is similar to that of other values. The impact of using this overlapping method on this dataset is not particularly significant. The test result graph selected from the test data are in Fig. 8.

It can be seen from the experimental results in Table 6 that the overlapping image blocking method has no effect on the results, so we use the non-overlapping image blocking data augmentation method.

The Precision-Recall curves and mAP values of the corresponding detectors at  $IOU > 0.5$  and  $IOU > 0.7$  are shown in Figs. 9 and 10.

Combining the results in Table 5 with the Precision-Recall curves, it can be seen that when the unified network Faster R-CNN model is used and combined with the different network structures ZF-Net, VGG16, and ResNet-101, at both  $IOU > 0.5$  and  $IOU > 0.7$ , ResNet-101 showed the best performance, followed by the VGG16 network, and the

**Table 13** Comparison of detection results between the YOLOv3 detector and the data augmentation-based YOLOv3 detector

YOLOv3-based detector		Data augmentation-based YOLOv3	
Test results	Partial enlargement	Test results	Partial enlargement
			
			
			
			
			
			

ZF-Net performed the worst on the PRW dataset. However, when using the YOLOv3 model and the Darknet-53 network structure, the mAP can reach 76.03%, nearly 6% higher than with ResNet-101, and far higher than with ZF-Net and VGG16 networks, while the recall rate reaches 82%. The last method based on the performance of data augmentation-based YOLOv3 has a mAP of 79.30% with  $IOU > 0.5$ , which is higher than 3% than when using YOLOv3 alone. With  $IOU > 0.7$ , its mAP value reaches 59.26%. This is more than 2% higher than when using YOLOv3 alone. This shows the effectiveness of the data augmentation-based YOLOv3 detector.

We also compared these five methods with DPM, ACF, and LDCF. As it can be seen from Table 7, although DPM, ACF, and LDCF are all excellent detectors, they perform relatively poorly compared to YOLOv3, especially compared to the data augmentation-based YOLOv3 detector. When  $IOU > 0.5$ , the best detection performance is by the LDCF, whose mAP value is 76.4%, which is approximately

3% lower than that of the data augmentation-based YOLOv3 detector. When  $IOU > 0.7$ , the best detection performance is by the DPM\_Alex, whose mAP value is 59.1%, which is still lower than that of the data augmentation-based YOLOv3 detector.

#### 4.2.2 Effectiveness of the LCNN-based pedestrian re-identification method

It can be seen from Table 8, when the LDCF is used as the detector and the IDE (Identification Embedding) as the recognizer, best performance is achieved, with a  $mAP^{0.5}$  of 34.4%. The performance of the combination of the YOLOv3 as the detector and the weighted CNN as the feature extractor and the cosine distance as the metric algorithm is much better than that of the other re-identification models, with a  $mAP^{0.5}$  of 53.21%. In general, the better is the performance of the detector, the better is the re-identification, but the final evaluation result from manual labeling as detector and

**Table 14** Data augmentation-based YOLOv3 detector and LCNN-based Recognizer on the PRW data set

Target Person		Pedestrian detection and map matching with different angles and across shots	
ID=484			
ID=488			
ID=539			
ID=540			
ID=627			

LCNN as the identifier is lower than that of the ResNet-101, YOLOv3, the data augmentation-based YOLOv3 as the detector and the LCNN as the recognizer. Because manually annotated pictures cannot effectively mark all pedestrians and target pedestrians in the picture. Their positions are all in close proximity, and the other three detectors indicate the results that are not manually labeled in the picture, as shown in Table 9. In this case, the performance of the detector is incorrectly evaluated.

#### 4.2.3 Effectiveness of end-to-end pedestrian detection and re-identification model based on image scaling data augmentation method

In this paper, the pedestrian detection and re-identification model based on the data augmentation-based YOLOv3 detector and the LCNN-based recognizer is called ImgEn\_YL, and the model based on the real-life pedestrian-based data augmentation method will be combined with ImgEn\_YL, and it will be called Real\_ImgEn\_YL. The  $mAP^{0.5}$

**Table 15** End-to-end pedestrian detection and re-recognition experiment results based on real-world data augmentation methods

Target Person		Pedestrian detection and recognition results in different states	
Back			
Front			
Side			
Occlusion			
Smaller person			

values on the real scene datasets of *ImgEn\_YL* and *Real\_ImgEn\_YL* are 53.11% and 55.49%, respectively.

From the first two experiments, we conclude that for pedestrian re-identification, the pedestrian detection and re-identification model consisting of the data augmentation-based YOLOv3 detector and the lightweight-based CNN recognizer have the best performed on the PRW dataset compared to other models. In this section we mainly use real-world scene datasets to further validate and improve the performance of the model. As we have mentioned before, the pedestrian in the foreground causes the overall image of the pedestrian to be blurred, which will affect the detection

and re-identification process. We found that when small-scale pedestrians are amplified and the lightweight-based CNN is used for feature extraction and similarity calculation, they can increase the similarity of the same pedestrian, and reduce the similarity of the different pedestrian. It can be seen from the above results that, after adding the pedestrian real scene-based data augmentation, the  $mAP^{0.5}$  is improved by 2.38% compared to the original value. The identification results of *ImgEn\_YL* and *Real\_ImgEn\_YL* are shown in Table 8. In the second and third columns, the pedestrians are marked with the ID number. When an unrelated pedestrian is detected, it will be marked.

We selected the back, the front, and the side of the target pedestrians, as well as occluded and small-sized target pedestrians. As shown in Table 10, for the back and occluded target characters in the picture, both models have selected target pedestrians successfully. But for the side and smaller target pedestrians, *ImgEn\_YL* marked unrelated pedestrian as the target, while *Real\_ImgEn\_YL* correctly identified the target. Furthermore, for the positive target pedestrian, *ImgEn\_YL* marked all the pedestrians as unrelated pedestrians, and the *Real\_ImgEn\_YL* model increased the similarity between the same pedestrian features, identifying the target pedestrians correctly. In summary, the end-to-end pedestrian detection and re-identification based data-augmented on real scene data proposed are effective.

### 4.3 Test results

#### 4.3.1 Results of the data augmentation-based YOLOv3 pedestrian detection method

Table 11 shows the results of the different detectors, which include the YOLOv3 based on the Darknet framework, the Faster R-CNN based on the Caffe framework, the Faster R-CNN based on the Tensorflow framework, and the YOLOv3 based on data augmentation, which use the networks darknet-53, ZFNet, VGG16Net, and ResNet-101, respectively. Test results of different detectors using partial enlargement method are shown in Table 12. The comparison of the detection effects of the YOLOv3-based pedestrian detector and the data augmentation-based YOLOv3 pedestrian detector is shown in Table 13.

It can be seen from Tables 11 and 12 that the missed detection phenomenon is more serious, which is based on the ZF network and the VGG16 network of the Faster R-CNN model. For the partial enlarged area in the first original image, the detector based on the ResNet-101 network of the Faster R-CNN model not only missed a pedestrian, but also marked a pink electric vehicle as a pedestrian. For the YOLOv3 detector and the data augmentation-based YOLOv3 detector, there is virtually no missed detection or false detection in the partially enlargement area of the first original image. For the partial enlargement area in the second original image, the data augmentation-based YOLOv3 detector detects more pedestrians than the YOLOv3 detector, and the pedestrian that is detected is small, which indicate the effectiveness of the data augmentation-based YOLOv3 detector we proposed.

It can be seen from Table 13, that the YOLOv3 detector and the data augmentation-based YOLOv3 detection method are highly sensitive to small objects, and have accurate positioning for small objects. In the two-column partial enlargement diagram in Table 13, it can be seen that the data augmentation-based YOLOv3 detector has improved

inspection of small-sized pedestrians performance in the six partial enlargement maps. In the second partial enlargement, while the YOLOv3 detector did not detect any pedestrians, the data augmentation-based YOLOv3 detector correctly detected 3 pedestrians. In summary, the data augmentation-based YOLOv3 detector has better accuracy and recall rate, which improves the robustness of the model.

#### 4.3.2 Results display of LCNN-based pedestrian recognition method

We selected three representative target pedestrians with ID numbers 484, 488, 539, 540, and 627, and selected two images for each object character to show the results of re-identification. If the detected person is the target person, it is marked with the corresponding ID number, otherwise, it is marked with a -2. We used the data augmentation-based YOLOv3 detector and LCNN-based recognition. The results are shown in Table 14.

In Table 14, all target pedestrians, from the back side, front side, side view, distant view, and close range, are recognized accurately when the data augmentation-based YOLOv3 is used as the detector and the LCNN is used as the recognizer.

#### 4.3.3 End-to-end pedestrian detection and re-recognition model based on image scaling data augmentation method

Table 15 shows the experimental results of the end-to-end pedestrian detection and re-identification based on data augmentation of real scenes. We first selected a pedestrian in the real scenes as the target pedestrian, while the other are unrelated pedestrians, and then use the data augmentation-based YOLOv3 detector to detect the pedestrians in the picture. Before using the recognizer to calculate the cosine similarity between the detected and target pedestrians, we need to use the cubic convolution interpolation algorithm for image size scaling to resolve the pedestrians whose size is less than a set threshold, and then re-identify the target pedestrian by the LCNN-based re-recognition model. We selected the target pedestrians from the back, the front, and the side, and occluded and small size pedestrians. For these five types of targets, the target pedestrian in the image are accurately identified, which proves that the effectiveness of the end-to-end pedestrian detection and re-identification models we proposed.

## 5 Conclusion

In this paper, we proposed an end-to-end pedestrian detection and re-recognition model applied to real scenes, to solve the problems of insufficient pedestrian expression ability, occlusion, pedestrian attitude, and small-scale pedestrian detection. The method first processes the original image with non-overlapped image blocking data augmentation method, and then inputs it into the YOLOv3 detector to obtain the target position information. Second, the LCNN-based pedestrian re-identification model is used to extract the features of the target. The eigenvectors of the target and the detected pedestrians and the similarity between the two are then calculated. Depending on the similarity value, the pedestrian may be marked as the target pedestrian. Finally, considering the characteristics of the scene, we process small-sized pedestrians whose size is less than a specified value using image size scaling, and an end-to-end pedestrian detection and pedestrian re-identification framework are obtained, which can be applied to the real scenes. Experiments show that the end-to-end pedestrian detection and re-recognition model developed in this study can improve the detection of small-scale pedestrians, improving pedestrian occlusion, diverse viewing angles, and the similarity between the foreground and background, while enhancing standard pedestrian recognition.

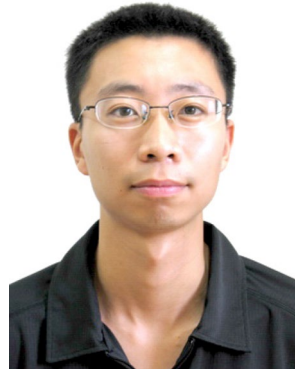
**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China under Grant 61972097, Grant 61502105, in part by the Technology Guidance Project of Fujian Province under Grant 2017H0015, in part by the Natural Science Foundation of Fujian Province under Grant 2018J1798, in part by the University Production Project of Fujian Province under Grant 2017H6008, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Fujian Engineering Research Center of Big Data Analysis and Processing. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

## References

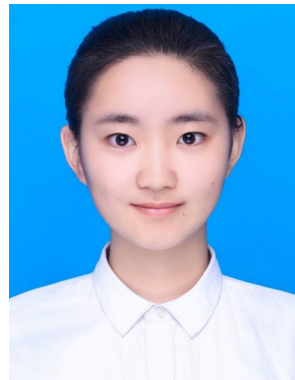
- Oren, M., Papageoriou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 193–199 (1997)
- Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision. Springer, Berlin, Heidelberg (2006)
- Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 886–893, vol. 1 (2005)
- Dollár, P., Appel, R., Belongie, S.: Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(8), 1532–1545 (2014)
- Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105 (2012)
- Girshick, R., Donahue, J., Darrell, T.: Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(1), 142–158 (2015)
- Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, 1440–1448 (2015)
- Ren, S., He, K., Girshick, R.: Faster R-CNN: towards real-time object detection with region proposal networks. In: International Conference on Neural Information Processing Systems. MIT Press, 91–99 (2015)
- Liu, W., Anguelov, D., Erhan, D.: SSD: single shot MultiBox detector. In: European Conference on Computer Vision. Springer International Publishing, 21–37 (2016)
- Redmon, J., Divvala, S., Girshick, R.: You only look once: unified, real-time object detection. In: IEEE Computer Vision and Pattern Recognition, 779–788 (2016)
- Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517–6525 (2016)
- He, K., Gkioxari, G., Dollar, P.: Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **99**, 1–1 (2017)
- Gheissari, N., Sebastian, T. B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1528–1535 (2006)
- Liao, S., Hu, Y., Zhu, X.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, 2197–2206 (2015)
- Yi, D., Lei, Z., Liao, S.: Deep metric learning for person re-identification. In: International Conference on Pattern Recognition. IEEE, 34–39 (2014)
- Li, W., Zhao, R., Xiao, T.: DeepReID: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, 152–159 (2014)
- Xiao, T., Li, H., Ouyang, W.: Learning deep feature representations with domain guided dropout for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 1249–1258 (2016)
- Zheng, L., Zhang, H., Sun, S.: Person re-identification in the wild. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3346–3355 (2017)
- Zheng, L., Bie, Z., Sun, Y.: MARS: a video benchmark for large-scale person re-identification. In: European Conference on Computer Vision. Springer International Publishing, 868–884 (2016)
- Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
- Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. *IEEE Conference on Computer Vision and Pattern Recognition* (2016)
- Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **10**(1), 207–244 (2009)
- Roth, P. M., Wohlhart, P., Hirzer, M.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, 2288–2295 (2012)
- Pedagadi, S., Orwell, J., Velastin, S.: Local fisher discriminant analysis for pedestrian re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, 3318–3325 (2013)
- Girshick, R., Donahue, J., Darrell, T.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, 580–587 (2014)
- Uijlings, J.R., Sande, K.E.A.V.D., Gevers, T.: Selective search for object recognition. *Int. J. Comput. Vis.* **104**(2), 154–171 (2013)

27. He, K., Zhang, X., Ren, S.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015)
28. Girshick, R.: Fast R-CNN. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1440–1448 (2015)
29. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. (2018). [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
30. Taigman, Y., Yang, M., Ranzato, M. A., Deepface: closing the gap to human-level performance in face verification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708 (2014)
31. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1891–1898 (2014)
32. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. *Adv. Neural Inf. Process. Syst.* **27**, 1988–1996 (2014)
33. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 2892–2900 (2015). <https://doi.org/10.1109/CVPR.2015.7298907>
34. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 815–823 (2015)
35. Wu, X., He, R., Sun, Z.: A lightened CNN for deep face representation. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 4 (2015)
36. Huang, G. B., Ramesh, M., Berg, T.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, 07–49, University of Massachusetts, Amherst
37. Munasinghe, S., Fookes, C., Sridharan, S.: Human-level face verification with intra-personal factor analysis and deep face representation. *IET Biometrics* **7**(5), 467–473 (2018)
38. Göngör, F., Tutsoy, Ö.: Design and implementation of a facial character analysis algorithm for humanoid robots. *Robotica* **37**, 1850–1866 (2019)
39. Göngör, F., Tutsoy, O.: Eigenface based emotion analysis algorithm and implementation to humanoid robot. In: *International Science and Academic Congress* (2018)
40. Cheng, K., Xu, F., Tao, F., et al.: Data-driven pedestrian re-identification based on hierarchical semantic representation. *Concurr. Comput. Pract. Exp.* e4403 (2017)
41. Sathish, P.K., Balaji, S.: A complete person re-identification model using Kernel-PCA-based Gabor-filtered hybrid descriptors. *Int. J. Multimed. Inf. Retrieval* **1**, 1–9 (2018)
42. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010). <https://doi.org/10.1109/tpami.2009.167>
43. Luo, J., Wang, Y.: Real-time pedestrian detection method based on improved YOLOv3. *J. Phys. Conf. Ser.* **1453**, 012149 (2020)
44. Yi, Z., Yongliang, S., Jun, Z.: An improved tiny-yolov3 pedestrian detection algorithm. *Optik Int. J. Light Electron Opt.* (2019)
45. Qu, H., Yuan, T., Sheng, Z., et al.: A pedestrian detection method based on YOLOv3 model and image enhanced by retinex. In: *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (2018)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Xiao Ke** received his PhD degree in artificial intelligence from Xiamen University of China in 2011. He is an associate professor with the College of Mathematics and Computer Science, Fuzhou University, China. His research interests relate to computer vision, pattern recognition, machine learning and their relations with innovative technologies. He has published over 60 scientific papers in the journals and conferences. He is a member of IEEE and CCF.



**Xinru Lin** is a graduate student in the college of mathematics and computer science, Fuzhou University. Her research interests relate to computer vision and deep learning.



**Liyun Qin** received her M.S. degree from the college of mathematics and computer science, Fuzhou University, in 2019. Her research interests relate to computer vision and machine learning.