Application Research on End-to-End Deep Person Re-Identification

Qingzhen Xu · Han Qiao · Shuang Liu · Shouqiang Liu

Received: date / Accepted: date

Abstract Pedestrian detection refers to the technology of predicting and locating the location of pedestrians in video or image. However, the recognition accuracy of existing re-identification methods needs to be improved. In this paper, the deep learning method is adopted, and the pedestrian detection based on YOLOV3-spp is combined with the pedestrian re-identification based on Resnet50, to construct the deep pedestrian re-identification system. In the training part, the COCO training set containing only pedestrian images is used to train pedestrian detection. WGAN-GP was used to expand the pedestrian re-identification training set of Market1501, and pedestrian re-recognition training was carried out on the expanded dataset to reduce the interference of pose diversity in the learning process. In the detection part, the system first uses the pedestrian detection to locate the original image and cut out the corresponding image, then input the cut image and the target image into the person re-identification network to learn the image features, and finally the classifier judges the

image containing the target. In the experiment, the identification accuracy of the system is 87.5% on the Market-1501 test set, and the identification speed of each picture is 0.12s.

Keywords Pedestrian detection · Person re-identification · YOLOv3 network · ResNet50 · WGAN-GP countermeasure generation network

1 Introduction

The purpose of pedestrian detection and person re-identification is to achieve intelligent pedestrian location and identification. The traditional person re-identification is mainly a study of both distance metric learning and feature extraction [1]. However, the low resolution of surveillance camera images, changing viewpoints, different light, and pedestrian blocking are the main challenges faced by person re-identification, making the two different pedestrian appearance features are more similar, so the accuracy of the traditional method of person re-identification is not high, and can't be well applied in practice.

Although the research and development of person re-identification has been going on at a high speed in recent years, there are still many problems to be solved. The problem of posture variability is caused by cross-viewing angles. Different cameras, different angles, and different appearance characteristics of the pedestrian posture at the same time are not the same.

The focus of our research is twofold. The first is how to realize a re-identification system application that can be applied to real scenarios. If person re-identification is used directly to extract features from the original image, a lot of noise will be introduced.

Qingzhen Xu
School of Computer, South China Normal University,
Guangzhou, China
E-mail: xqz1997@126.com

Han Qiao
School of Computer, South China Normal University,
Guangzhou, China
E-mail: qiaohan@m.scnu.edu.cn

Shuang Liu
School of Computer, South China Normal University,
Guangzhou, China
E-mail: 2818434355@qq.com

Shouqiang Liu
School of Physics and Telecommunications Engineering, South
China Normal University, Guangzhou, China
E-mail: liusq@m.scnu.edu.cn

Therefore, person re-identification and pedestrian detection are used together to combine the detection task and identification in real scenes. For pedestrian detection, YOLOv3 is used for person re-identification and pedestrian detection. In order to achieve real-time, ResNet50 discards multiple local features and uses one global feature, which greatly reduces the number of feature vectors in FC layer, reduces the resources consumed for training, and makes the training and testing speed of re-identification increase. The accuracy of the model is improved by applying various training tricks.

Second, for the posture diversity problem, the pedestrians have different postures under different cameras at different points in time, and the re-identification system may misjudge, making the re-identification accuracy decrease. This paper attempts to solve this problem by using GAN-generated adversarial network to generate the pedestrian map with various postures. We train the adversarial network WGAN-GP on the Market-1501 dataset, and use the trained network to generate the pedestrian images with various postures and retain the identity information of the images. The generated images are mixed with the original dataset to enhance the diversity of the person re-identification training set.

The rest of this paper is organized as follows. Section 2 is the related work. It introduces the research background and the current situation. Section 3 is the application development technology. This part introduces the tools and algorithms used in the application, and carries out the overall design of the application. This section also introduces the dataset and model used for training, and explains the key code. Section 4 is the running results of the person re-identification application. This chapter mainly shows the final implementation after the application runs and analyzes the shortcomings in the application. Section 5 is a summary and prospect which sums up the research process and looks forward to the future.

2 Related work

2.1 Current status of pedestrian detection research

Current pedestrian detection methods include traditional detection methods and deep learning detection methods [2]. The traditional methods include background-based modeling and statistical learning. The major disadvantage of background-based modeling is that the performance is very poor when there is a big difference in image styles such as color difference and dither screen caused by illumination. The statistical learning-based approach is trained using data with

existing known information to derive more accurate detectors. This approach is represented by the HOG detector introduced by Tris in 2005 [3], followed by the Deformable Component Model (DPM) proposed by Felzenszwalb in 2012 based on the HOG detector [4], which is a training model with variability that reduces the impact of problems such as viewpoint variability and morphological differences.

With the important idea of selective search [5] proposed by R-CNN [6], the series of R-CNN such as R-FCN [7], Faster R-CNN [8] and some algorithms based on candidate boxes rapidly appeared. Then there are the popular target detection methods of SSD [9], YOLO [10] and other end-to-end training methods. These methods are faster, maintain high accuracy, reduce model capacity, and make system development easier.

2.2 Current status of person re-identification research

There are four current methods for person re-identification based on deep learning [1]. The first one, based on representation learning, the second one, based on metric learning for ReID, the third one, based on video sequences, and the fourth one, based on GAN [11], an important use of GAN is image generation. CycleGAN [12], DualGAN [13], and DiscoGAN [14] implement image style conversions. CycleGAN enables the migration of camera styles and reduces the bias between cameras. Due to the existence of scene domain bias between different datasets, a training model from one dataset performs poorly in another dataset. PTGAN [15], which was developed from CycleGAN, is able to change the style of one data domain to the style of the target data domain by changing the background of the pedestrian to the background of the target data domain style. The models trained by the above four methods, when used in proper combination, are better than those trained by a single method.

3 Proposed method

3.1 Application development platforms and tools

The development environment used in this paper is Windows 10. The development language and deep learning framework are Python 3.5 and Pytorch 0.4. We used COCO train2017 and Market-1501 datasets in our experiments. On this basis, Yolov3-spp model and ResNet-IBN-a model neural network model were used in this paper.

3.2 GAN-generated adversarial networks

Through unsupervised learning, the generative adversarial network learns the features of the input image, the generator generates the image, and then the discriminator discriminates between the generated image and the fake image, the whole network is tuned in a dynamic adversarial process, and the goal of the generator is to generate most of the images that can fool the discriminator. In order to solve the cross-camera problem [16], Zhun et al. use CycleGAN to process datasets and convert images of the same pedestrians in one lens into the style of images in other lenses to improve the network's adaptation to different perspectives and styles, while Liang et al. solve the cross-domain problem [17], based on the CyCleGAN. Added a SiaNet on top, improved the loss function, and built SPGAN, this enables image-level style migration.

To understand the generator distribution p_g over data x , we define a prior input noise variable $p_z(z)$, and then express the mapping to the data space as $G(z; \theta_g)$, where G is a differentiable function represented by a multilayer perceptron with parameter θ_g . We also define a second level of perceptron $D(x; \theta_d)$ output a single scalar. $D(x)$ is the probability that x comes from the data and not from p_g . We train D to maximize the probability of assigning the correct label to the training sample and to the G sample. We also train G to minimize $\log(1 - D(G(z)))$:

In other words, D and G play the following two-player minimax game with value function $V(G, D)$:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

We first consider the optimal discriminator D for any given generator G . For G fixation, the optimal discriminator D is

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \quad (2)$$

The training criterion for discriminator D given any generator G is to maximize the quantity $V(G, D)$.

$$\begin{aligned} V(G, D) &= \int_x p_{data}(x) \log(D(x)) dx + \int_z p_z(z) \log(1 - D(G(z))) dz \\ &= \int_x p_{data}(x) \log(D(x)) + p_g(x) \log(1 - D(x)) dx \end{aligned} \quad (3)$$

Note that D the training goal of maximum can be explained by the estimated conditional probability $P(Y = y|x)$ of the logarithmic likelihood, including $Y = x$ is from $p_{data}(y = 1)$ or from $p_g(y = 0)$. The minimax game in Eq. 1 can now be reformulated as:

$$\begin{aligned} C(G) &= \max_D V(G, D) \\ &= E_{x \sim p_{data}} [\log D_G^*(x)] + E_{z \sim p_z} [\log(1 - D_G^*(G(z)))] \\ &= E_{x \sim p_{data}} [\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + E_{x \sim p_g} [\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}] \end{aligned} \quad (4)$$

Through the above types, we can get:

$$\begin{aligned} C(G) &= -\log(4) + KL(p_{data} || \frac{p_{data} + p_g}{2}) + KL(p_g || \frac{p_{data} + p_g}{2}) \end{aligned} \quad (5)$$

Where KL is the Kullback–Leibler divergence. We recognize in the previous expression the Jensen–Shannon divergence between the distribution of the model and the data generation process:

$$C(G) = -\log(4) + 2 \times JSD(p_{data} || p_g) \quad (6)$$

Since the Jensen–Shannon divergence between two distributions is always non-negative and is zero only if they are equal [18].

3.3 Application function module design and functional descriptions

Two main parts are divided by this application. The first part is pedestrian detection. The training model is used to obtain the result of predicting pedestrian location from a single image. The pedestrian detection part uses COCO train2017 as the training dataset, and there are 118287 images in the training set, of which 64115 are pedestrian images, which is a large dataset to train a good performance model. YOLOv3-spp is a deep convolutional neural network target detection algorithm, which has the advantage of fast detection speed and is suitable for real-time systems. YOLOv3 with the spp module is more accurate than the native YOLOv3 detection and saves computational cost.

The second part is the re-identification part. The pedestrians appearing in the realistic task scenario have

more complex backgrounds and the accuracy of re-identification will be reduced. For model training, the network model for person re-identification is ResNet-IBN-a, which is a fusion of IN and BN added on the basis of ResNet50 to improve the network's generalization performance. The training dataset uses the Market-1501 dataset, which was made public in 2015 and consists of six cameras with a total of 1501 pedestrians.

The application enters a picture of the target pedestrian to be queried and a set of pictures or videos containing the target. The application design flowchart is shown in Fig. 1.

3.4 YOLOv3-based pedestrian detection

3.4.1 Datasets

For the pedestrian detection part, the COCO train2017 dataset was selected for this application. The COCO dataset is a large dataset proposed by Microsoft that targets scene identification, taken from everyday scenes, and pinpoints the location of pedestrians in images. The images include 330,000 images from 91 categories.

The dataset contains 64115 images of pedestrians out of 118287 images, and each image has a corresponding tag file that records the exact location of the pedestrian in that image. There are three tag types: object instance, object key points, and image, which are stored in JSON files. Since the study is about pedestrians, the COCO dataset is preprocessed as follows.

In this experiment, the images of pedestrians in the COCO train2017 dataset are first selected and a training set containing only the pedestrian gender is compiled. The training set contains 64115 images of pedestrians with different backgrounds and poses, and then the dataset is divided according to the ratio of 8:1:1. The specific situation is shown in Table 1.

3.4.2 YOLOv3-spp network model

YOLOv3 [19] is an algorithm for target identification and localization based on convolutional neural networks. YOLOv3 uses a model called darknet-53, which is characterized by the absence of pooling and full connection layers. Darknet-53 uses downsampling to extract deep features. There are five downsamples, so the height and width of the output picture are 1/32 of the height and width of the input picture, so the length and width of the input image are generally multiples of 32.

The spp module is added to YOLOv3 to solve the problem of varying input image sizes, and the addition

of the spp module enriches the expression capability of the feature map by fusing local features with global features, which is beneficial to the case where the size of the object to be detected varies greatly, and improves the detection accuracy.

3.4.3 Model training

The first step is to build the network according to YOLOv3-spp's network structure, there are 53 layers in total. According to the name of each layer in the cfg file, we will build the corresponding layer, such as the name of the convolutional layer to create a convolutional layer, and the name of the pooling layer to create a pooling layer. The pooling and shortcut layers are also similarly designed. After building the model, Darknet() is called to initialize the model variables.

After initializing the model, we need to get the images and labels of the COCO dataset, and according to the path on the previously obtained training list, the program will find the corresponding images and label information. Although images can also be put into memory to speed up training, the memory in the environment where the experiment is conducted is not sufficient to put all the images into memory due to the large size of the images data. After the dataset variables are initialized, two more data loaders are defined in the code: dataloader and testloader, both of which are dataloaders.

The epoch parameter in the program determines the number of training rounds for all the data, in this application, the epoch is set to 273, to enter 273 training rounds. In each training round, the YOLOv3 model first acquires the pre-processed images and labels, and then outputs three feature maps after the YOLOv3 network, the results of the three feature maps and the labels of the target are calculated to get a loss. Perform a BP(reverse pass) and then save the final parameters of the round.

Save the model after each training round and record the current number of training rounds and accuracy.

3.4.4 Pedestrian detection results

Once the model has been trained, it can be used to perform pedestrian detection, enter a picture or a video, and output the predicted location of the pedestrian in it.

Load the model with the training weights, then load the test set, go through the YOLOv3 network and draw a box selection map that frames each predicted pedestrian location and gives the confidence level of each box.

Fig. 1: Application design flowchart

Table 1: YOLO training set

Dataset	Total number of images	Training set	Test set	Number of Categories
COCO train2017	165482	118287	81434	80
Ours	64115	51292	6411	1

3.5 ResNet50-based person re-identification

3.5.1 Datasets

The dataset used for person re-identification is the Market-1501 dataset. Among the query images, there are 3368 manually drawn pedestrian detection boxes, and the rest of the pedestrian detection boxes are generated using automatic DMP detection. The training set of ResNet50 is shown in Table 2.

3.5.2 WGAN-GP

The difficulty in training and the inability of loss of generator and discriminator to indicate the training process are the problems that the original GAN cannot solve. Improvements were made to the original GAN, and improvements like DCGAN [20] appeared, but the

problems still could not be fundamentally solved. The improvement of WGAN is as follows: remove sigmoid from the last layer of discriminator; loss of generator and discriminator does not take log; the problem of difficulty and instability in GAN training is solved by forcing the weight to be truncated and updated within a certain range. WGAN-GP [21] is an improvement on WGAN [22], and improves the condition of continuity restriction, so that possible gradient disappearance and gradient explosion problems can be solved, and the convergence speed of model training becomes faster.

WGAN-GP was used to generate images with different attitudes to expand the dataset. First, 200 epochs were trained on the Market-1501 dataset as the training set of WGAN-GP, so that the generator could learn the posture characteristics of different pedestrians. After that, the trained generator is used to process

Table 2: Training set of ResNet50

Dataset	Total number of images	Training set	Test set	Number of training set IDs	Number of test set IDs
Market-1501	32668	12936	19732	751	750

the original dataset and generate various pedestrian images. The generated images are named according to the naming format of the original images. Finally, the generated images are added to the expanded person re-identification training set.

3.5.3 Network model based on ResNet50

BNNeck is added on the basis of ResNet50-IBN-a model used in this experiment. The feature obtained after the fully pooling layer of the network directly connects triplet loss, and then the feature finally reaches the fully connected layer after passing through a BN layer. By adding a BN layer, the feature is normalized to the hyper sphere, so that triplet loss can constrain the feature and improve the performance in the free European space [23].

3.5.4 Model training

In the first stage, WGAN-GP was trained on Market-1501, the trained generator was used to process the original Market-1501 dataset, and the diversified images generated by WGAN-GP were added to the training image set. The identity label was consistent with the original image, which was equivalent to the expansion of the number of different pose images of the same pedestrian in the dataset.

In the second stage, the augmented dataset is used for person re-identification training. First, import torch.nn and Model_zoo, load pytorch network module and load pre-training weights according to urls.

Next comes the residuals module definition. Three convolutional layers are set up to use different dimensions to compress, convolutional process, and restore dimensions. Inplane variable is the channels' input number in every layer, plane is the output number of channels per layer after processing. Expansion is a parameter used to multiply the output number of channels, so the total output number of channels is the product of plane * expansion.

ResNet50 Layer4 convolution layer after a gap (global average pooling layer), and then reshape, instead of full connection layer, the global characteristics, used as Tiptetloss function input. Then the global feature is connected to the BNNeck structure to obtain the normalized feature which is accessed to the

fully connected classifier (classification output layer). Finally, the network returns a classified output layer and a global feature.

With the above models in place and the TripletLoss loss function as well as the SGD Optimizer, you can train and save the weights in the weights folder.

3.5.5 Combination of person re-identification and pedestrian detection

The input of this application is a target image Query. There are also some datasets containing target images, such as a photo set or a video. If it is a video, the method of frame skipping detection is adopted to split the video into one image. The combination of person re-identification and pedestrian detection flowchart is shown in Fig. 2.

First, load the person re-identification model. After that, the pedestrian detection model is loaded, and the pedestrian detection traverses the dataset, and the pedestrians in each picture are predicted and pruned according to the prediction box: the features of the cut pictures are extracted by using the person re-identification model, and the Gallery_feats is named.

After that, the distance measurement algorithm of the obtained Query_feats and Gallery_experience is used to calculate the similarity of the two extracted features. If the distance is less than the threshold value of 1.0, the same person is considered to be found and the prearranged box is punched in the original picture and saved in the folder.

4 Experiments

4.1 Analysis of experimental results

The application can predict the pedestrian position in pictures or videos, and find out the picture of the pedestrian to be searched or the location of the pedestrian to be searched in the video. The person found will be checked in the box and prompted to find. If the image is found in the image set, output the image to the Output folder. If the object is a video, output the re-identification video to the folder. The output of the identification system is as shown in Fig. 3 and Fig. 4.

In this paper, the mAP index is used to evaluate the recognition performance, which represents the average

Fig. 2: Combination of person re-identification and pedestrian detection

Fig. 3: The pedestrian to be retrieved on the left and one of the retrieved image outputs on the right

accuracy of the recognition task. It can objectively evaluate the experimental results by the position of positive and negative samples in the retrieval results. AP (Average Precision) represents the Average accuracy of each query, which is calculated as shown in Eq(7):

$$AP = \frac{\sum_{k=1}^n (p(k) * rel(k))}{N} \quad (7)$$

Among them, N recognition results and check belong to the same kinds of results, n represents the number of all results returned by recognition, k index for identification results in a sequence of position, $p(k)$ for the former k recognition accuracy, $rel(k)$ said the recall the first k recognition results and whether the query with the same label. When $n = R$, $mAP@R$ represents the first R results with the highest similarity sorted by similarity, and the average of R results is taken. When n is all samples, mAP represents an

average of all samples for the current query. mAP is the average of the corresponding AP values for multiple queries:

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \quad (8)$$

In terms of pedestrian detection, the dataset selects the COCO dataset processed by pedestrian extraction as the training set. The model conducts a total of 290 epoches training, and the final accuracy mAP reaches 78.4

In terms of person re-identification, the images generated by Market-1501 and WGAN-GP were selected as the training set, which included 25,872 images and 751 test sets. The accuracy mAP of 120 models after the epoch was trained to the dataset of Market-1501 reached 87.5

Fig. 4: The pedestrian to be retrieved on the left and the processed video output on the right

In order to prove the performance improvement of the model with WGAN-GP generated data for training, the dataset containing only the original Market-1501 was also trained. After that, the two training models were tested on the original Market-1501 test set. The above tests were also performed on the DukeMTMC-reID dataset. The test results are shown in Table 3 and Table 4.

The model's accuracy trained with mixed dataset is improved on the original dataset, and the accuracy is also greatly improved when the model is tested on the diversified dataset. There are two reasons for the improvement in accuracy. First, the scale of the training dataset is increased. The number of pictures generated by WGAN is the same as the original training set, so the total training set is doubled. The larger the training set, the better the training effect will be. Second, the network can learn more natural diversified pictures, so the network can learn more features of the same pedestrian, so the accuracy rate is also improved.

Table 5 compares the recognition accuracy with that of other classical algorithms on DukemtMC-Reid and Market-1501 dataset, and the data show that the method we proposed can effectively improve the identification accuracy.

It takes a total of 4.12 seconds to search for a pedestrian in Market-1501, with an average of 0.12 seconds for each picture. Overall, the system takes a longer time, because every time testing under real

scene images, it takes the pedestrian detection model of pedestrians in the diagram position detection, then detect the pedestrian position input to the person re-identification model in feature extraction and identification, so after the two models, the time is long. From the perspective of individual models, the complexity of YOLOv3 and ResNet50 models themselves and the detection part of each model has not been optimized and pruned, so the overall speed is relatively slow.

4.2 Experimental reflection

Currently, there are three major problems in the application. Low identification accuracy. The accuracy of the final system is not up to the requirement of real-time detection. Because of the amount of data is less, and a single training dataset, the pedestrian detection if can in COCO datasets to join other joint training dataset, the model will be more accurate, but since more data requires more hardware, our existing hardware cannot meet this idea; In terms of person re-identification, the quality of WGAN-GP graphs is not very good, so the accuracy can be improved, but not much.

In the practical application of the scene, the number of pedestrians is large, there are occlusion problems, or the pedestrian clothes are changeable, the image resolution collected by the camera is too low to extract valuable features, and the generalization ability of the

Table 3: Results of person re-identification model in Market-1501

Dataset	Total number	Original test set (mAP)	Diversified test set (mAP)
Market-1501 training set	12936	86%	65%
Market-1501 + WGAN-GP generating set	25872	87.5%	70%

Table 4: Results of person re-identification model training in DukemtMC-Reid

Dataset	Total number	Original	Diversified
DukeMTMC-reID training set	16522	76.5%	62.5%
DukeMTMC-reID + WGAN-GP generating set	33044	80%	65%

Table 5: Comparison with classical algorithms on Market-1501 and DukemtMC-Reid dataset

Method	Market-1501 (mAP)	DukeMTMC-reID (mAP)
AlignedReID [24]	77.7%	67.4%
SCPNet [25]	75.2%	62.6%
Mancs [26]	82.3%	71.8%
DuATM [27]	76.6%	62.3%
Camstyle [28]	68.7%	53.5%
SVDNet [29]	62.1%	56.8%
AWTL [30]	75.7%	63.4%
Ours	87.5%	80.0%

model is insufficient, resulting in the low identification rate of the system.

Long detection time. Although this application combines the detection part of the two models to improve the detection accuracy compared with the single use of re-identification detection, the detection speed cannot meet the requirements of real-time detection. The reasoning part of the model needs to be optimized, and distillation and pruning operations may speed up the reasoning process.

5 Summary and prospect

5.1 Summarize

This paper used the deep learning method to extract the pedestrian from the complex background, and then input the extracted pedestrian into the person re-identification for feature extraction, and measure the similarity with the target pedestrian, and determine whether the pedestrian is the same person according to the measurement results. The specific work is as follows.

Detailed introduction of each model from dataset, network architecture to training process and some important code implementation. The first step is to train pedestrian detector on COCO train2017 dataset based on YOLOv3 target detection algorithm. In the

second step, data preparation was carried out and WGAN-GP was used to expand the dataset of Market-1501, which enriched the diversity of pedestrian posture in the training set. In the third step, the ResNet50 model with BNNeck layer is used to carry out re-identification network training on the expanded dataset of Market-1501. The fourth step is to combine the pedestrian detection with the actual detection of person re-identification.

The design of this application has basically completed the expected function and can detect and identify pedestrians in the photo collection or video. At the same time, this application also has many disadvantages.

- (1) The scale of training set is not large enough.
- (2) The system accuracy needs to be improved.
- (3) Slow system prediction speed.

The system is still a long way from the requirements of real-time applications and needs to be improved.

5.2 Prospect

In the future, there will be higher and higher requirements on the performance and speed of person re-identification. Due to the complexity of practical application scenes, the following scale changes, rotation, shielding, lighting and other problems will occur in

the pedestrian images taken [31], which are all great challenges. Feature extraction and measurement learning optimization are now the main research directions. Meanwhile, person re-identification cannot rely on additional body marks or clear biological information like face identification, so the latest theoretical knowledge and computer vision technology are needed to solve the re-identification problem in complex scenes.

The accuracy of the detection and re-identification algorithm used in this paper can be further improved in the following aspects: First, further optimization of the person re-identification model. Different datasets have different data distribution. At the same time, the training set data obtained from the camera is small in scale and the model training accuracy is low. For these problems, the generative adversarial networks can be used for style migration to reduce the differences between datasets and improve the migration ability of models [30]. Or, from another perspective, techniques such as posture prediction can be used for training and detection with additional body labeling information. Second, continue to pay attention to the latest research results of the academic community and think about improving the model. At the present stage, we are not familiar with all kinds of models outside this paper, so we can strengthen our knowledge in this aspect. For example, the newly launched YOLOV5 model, compared to YOLOV3 model, has a smaller network structure, faster speed and higher efficiency. YOLOV5 is divided into 4 models, namely YOLOV5, YOLOV5M, YOLOV5L and YOLOV5X. The suitable model can be selected according to different types of tasks, which is quite flexible. Next, the author will try to make further exploration based on the model of YOLOV5. Third, the influence of person re-identification interference factors still exists, and the solutions to various interference factors emerge endlessly at the present stage, but they can only mitigate to a certain extent the influence of one or more kinds of noises in the complex scene [1]. Therefore, how to learn from the existing excellent network architecture design ideas and study a more comprehensive model that can resist a variety of complex noises is the goal of continuous learning and efforts in the future. Fourth, the dataset used in academic research is very different from the data of the actual scene, because the images of the actual scene will have problems such as low resolution and insufficient light [32]. If we want to achieve the accuracy in the actual scene, such as in the experimental dataset, we should continue to improve the existing method. The goal is to brush the accuracy in the research dataset to 100%, and then enter the study of the actual scene to make this technique more practical. Fifth,

throughout the present re-identification algorithms, most of them are supervised learning, which requires a large amount of marked data and takes time and effort to collect samples manually and classify. Therefore, semi-supervised learning or unsupervised learning will be the mainstream in the future [1].

Acknowledgements The Project was supported by Guangzhou Science and Technology Plan Project(No.201903010103),the “13th Five-Year Plan” for the development of Philosophy and Social Sciences in Guangzhou (No.2018GZYSB36), Science Foundation of Guangdong Provincial Communications Department, China (No.N2015-02-064),The Ministry of Education’s 2018 first batch of Industry-University Cooperation Collaborative Education Information Security curriculum system construction projects (201801087012).

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest The authors declare that they have no conflict of interest.

Informed Consent This article does not contain any studies with animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

Authors’ contributions All authors contributed to the research, experiment and manuscript. Qingzhen Xu and Han Qiao were responsible for the design of the algorithm and the preparation of the experiment. The experiment and related discussion were performed by Qingzhen Xu, Han Qiao, Shuang Liu, and Shouqiang Liu. Qingzhen Xu, Han Qiao and Shouqiang Liu wrote the manuscript. Shuang Liu and Shouqiang Liu were responsible for the final optimization. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Wanru Song, Qing-qing Zhao, Chang-hong Chen, Zong-liang Gan, Feng Liu. A review of pedestrian recognition research [J]. Journal of Intelligent Systems, 2017,12(06):770-780.
2. Rahul Rama Varior, Mrinal Haloi, Gang Wang. Gated-siamese convolutional neural network architecture for human re-identification[C]. Euopean Conference on Computer Vision. Springer, 2016:791-808.
3. Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:886-88-93.
4. Felzenszwalb P F, Girshick R B, McAllester D, et al. Object Detection with Discriminatively Trained Part-Based Models[C]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645

5. Uijlings, Jasper RR, et al. "Selective search for object recognition." [J]International journal of computer vision, 104(2) (2013): 154-171.
6. GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Ohio, USA: IEEE, 580-587.
7. Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks[C]. Proc of Conference on Advances in Neural Information Processing Systems.2016: 379-387.
8. REN S, HE K, GIRSHICK R, et al. Faster R-CNN:Towards real-time object detection with region proposal networks[C]//International Conference on Neural Information Processing Systems.Cambridge:MIT Press,2015:91-99.
9. Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision. 2016:21-37.
10. Redmon J, Divvala S, Girshick R and Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 779-788.
11. Goodfellow I, Pouget Abadie J, Mirza M, Xu B, Warde Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the Advances in Neural Information Processing Systems (NIPS)[J], 2014. 2672-2680.
12. Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017.2242-2251.
13. Yi Z L, Zhang H, Tan P, Gong M L. DualGAN: unsupervised dual learning for image-to-image translation[C]. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2868–2876.
14. Kim T, Cha M, Kim H, Lee J K, Kim J. Learning to discover cross-domain relations with generative adversarial networks[J]. arXiv: 1703.05192, 2017.
15. Wei L H, Zhang S L, Gao W, Tian Q. Person transfer GAN to bridge domain gap for person re-identification[J]. arXiv:1711.08565, 2018.
16. Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, Yi Yang. Camera Style Adaptation for Person Re-identification[J].2018.
17. Deng W, Zheng L, Ye Q, et al. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain Dissimilarity for Person Re-identification[J].2017.
18. Goodfellow I J , Pouget-Abadie J , Mirza M , et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.
19. Redmon, Joseph, Farhadi, Ali. YOLOv3: An Incremental Improvement[J].arXiv:1804.02767, 2018.
20. Li J , Jia J , Xu D . Unsupervised Representation Learning of Image-Based Plant Disease with Deep Convolutional Generative Adversarial Networks[C]. 2018.
21. Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin. etc. Improved Training of Wasserstein GANs[J].arXiv:170.00028, 2017.
22. Arjovsky, Martin, Chintala, Soumith, Bottou, Léon. Wasserstein GAN[J]. 2017.
23. Hao Luo Youzhi Gu Xingyu Liao Shenqi Lai Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification[C]. CVPRW, 2019.00190.
24. Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person reidentification. arXiv preprint arXiv:1711.08184, 2017.
25. Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. ArXiv preprint arXiv:1810.06996, 2018.
26. Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggong Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 365–381, 2018.
27. Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5363–5372, 2018.
28. Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. IEEE Transactions on Image Processing, 28(3):1176–1190, 2019.
29. Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, pages 3800–3808, 2017.
30. Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6036–6046, 2018.
31. Feng Yang, Yu Xu, Mengxiao Yin, Jiacheng Fu, Bing Huang, Fang-xuan Liang. A Review of Pedestrian recognition based on deep learning [J/OL]. Computer application:1-12[2020-04-26]. <http://kns.cnki.net/kcms/detail/51.1307.TP.20191223.0927.002.html>.
32. Huaxiang Zhang, Li Liu. Research review of Pedestrian Recognition [J]. Journal of Shandong Normal University (Natural Science edition), 2018,33(04):379-387.