Person re-identification method based on multi-scale residual pooling

Zhenxiang He[*a], Xiaorong Liu[b], He Zhu[b]
[a]School of Intelligent Science and Technology, Tianfu College of Southwest University of Finance and Economics, Mianyang 621000, Sichuan, China; [b]School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, Sichuan, China

ABSTRACT

Person re-identification is a cross camera pedestrian retrieval problem. The data retrieved by pedestrians can be images, videos, and text. The current person re-identification methods are insufficient in expression of pedestrian features and poor robustness, resulting in low model accuracy. This paper proposes a Multi-scale Residual Pooling for person re-identification. ResNet50 is used as the basic network to obtain the multi-scale features. Global average pooling and maximum average pooling are performed on the input features at different network levels. Each group of average pooled and maximum pooled features is subtracted to remove the influence of image background clutter. The subtracted difference features are added to the maximum pooled features to obtain a more discriminative residual pooled fusion feature, making the network focus on the whole body contour of pedestrians and the difference between pedestrians and background. On this basis, triplet loss and cross-entropy loss are combined to optimize the model, and reordering technology is used to optimize the network. The experimental results showed that the Rank1 index of this paper's method tested on the Market1501 and Duke MTMC-reID datasets reaches 96.41% and 91.43%, respectively, and mAP (mean Average Precision) reaches 94.52% and 89.30%, respectively. which is better than the current mainstream algorithms.

Keywords: Person re-identification, multi-scale features, residual pooling, feature fusion, deep learning (DL)

1. INTRODUCTION

In the field of computer vision, the pedestrian re-identification task usually first gives the monitoring image of a specific pedestrian and uses the pedestrian re-identification technology to find the image taken by the pedestrian under other cameras in the database[1]. Due to the different position and angle of view of the camera, and the influence of pedestrian posture, occlusion, illumination changes, and other factors, the images of the same pedestrian are quite different, which makes pedestrian re-identification change into a hot research topic.

In the early days, researchers realized the task of pedestrian re-identification uses traditional methods. With the development of convolutional neural networks, the method based on deep learning is applied to the task of pedestrian re-identification. At present, the mainstream methods of pedestrian re-identification based on deep learning mostly adopt average pooling, max pooling, or a combination of the two methods. Combined with the features of average pooling and max pooling. We propose a pedestrian re-identification method based on Multi-scale Residual Pooling. According to the information contained in different scale features in the network, a new residual pooling module is used to combine the advantages of average pooling and max pooling to extract more comprehensive and more discriminative residual pooling features to represent pedestrians, and improve the performance of pedestrian re-identification network.

2. RELATED WORK

The conventional pedestrian recognition methods are mainly included feature based and distance measurement method. The feature representation method mainly extracts features such as color, LBP[2], and SIFT[3]. Due to the limitations of a single feature in pedestrian target representation, researchers have proposed many other methods: Reference[4] uses a cumulative color histogram to represent global features, and then extracts local features; Reference[5] introduced LOMO. Its idea is that the distance of the same pedestrian target should be less than the different pedestrian targets. KISSME[6] and the LMNN[7] algorithm are used to learn the best similarity measure.

[*] 122333539@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066I
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661830

The traditional pedestrian re recognition method has limited ability to extract features, which is not effective for the recognition task in the actual scene. Numerous scholars apply advanced deep learning to solve the pedestrian re-identification task nowadays. At present, most research on deep learning are devoted to extracting global and local features to obtain discriminative pedestrian feature expression. GLAD[8] uses global and local thinning methods to extract features; Reference[9] proposed an evenly partitioned PCB model. After the obtained features are equally divided, the image blocks are aligned through the RPP network, and then the local features of each image block are extracted.

When constructing convolution neural network, the general processing is to insert a pooling layer followed the convolutional layer. Its function is to reduce feature dimension of convolution output and suppress noise and prevent overfitting Average pooling in a convolutional neural network can transfer feature information completely, but it is easily affected by background noise; Max pooling can extract features with better recognition but pay more attention to local information. Table 1 shows the pooling methods of mainstream networks. Most pedestrian re-identification methods based on convolutional neural networks only use average pooling or max pooling, or simply fuse the output features of the two pooling.

Table 1. Pooling methods of mainstream network.

Network	Max pooling method	Average pooling method
LeNet5[10]	√ (Early)	√ (Late)
AlexNet[11]	√	
VGGNet[12]	√	
NIN[13]	√	√
GoogloeNet[14]	√	√
ResNet[15]	√	√

3. PERSON RE-IDENTIFICATION METHOD BASED ON MULTI-SCALE RESIDUAL POOLING

The pedestrian re-identification network structure includes: multi-scale feature extraction and residual feature acquisition. Figure 1 shows the network structure.

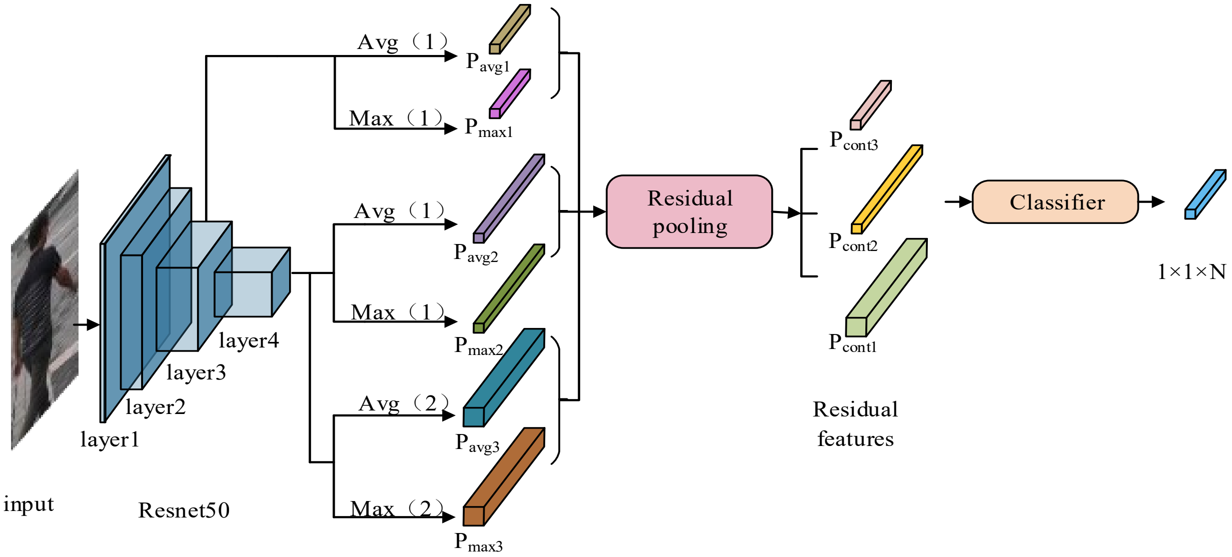


Figure 1. Person re-identification network structure based on multi-scale residual pooling feature fusion.

3.1 Multi-scale feature extraction

Different levels of the convolutional neural network will produce different spatial resolution feature maps, and the feature maps obtained through different convolution layers contain different information. High-level features focus more on semantic information and less on image details., while the low-level features may contain more details and chaotic background information. Therefore, most researchers combine the features of multiple scales to complement the features of different levels in this simple and effective way. The pedestrian re-identification network structure designed in our paper removes the last full connection layer of the ResNet50 network, plus the average pooling layer and the max pooling layer, as shown in Figure 1. Avg (m) and Max (m) represent average pooling and max pooling respectively, and a feature mAP with width and height of m is obtained. In this paper, the output features of layer 3 in the ResNet50 network are global average pooling and global max pooling respectively, and Pavg1 and Pmax1 feature maps with output dimensions of 1×1×1024 are obtained. Similarly, the output features of layer4 in the ResNet50 network are global average and global max respectively, and Pavg2 and Pmax2 feature maps with output dimensions of 1×1×2048 are obtained. To reduce the impact of pooling on information loss, the stripes of average pooling and max pooling are adjusted to obtain richer feature information. Pavg3 and Pmax3 feature maps with output dimensions of 2×2×2048 are obtained. The multi-scale features Pavg1, Pmax1, Pavg2, Pmax2, Pavg3, and Pmax3 of the pedestrian images are sent to the residual pooling module to obtain the corresponding residual features Pcont1, Pcont2, and Pcont3, which are transformed into a unified dimension for fusion, the fused features are sent to the classifier for classification, and finally, the pedestrian re-identification results are obtained.

3.2 Residual pooling module

Figure 2 shows the average pooling and max pooling commonly used in convolutional neural networks. Average pooling is to average the features in the neighborhood, and max pooling is to maximize the features in the neighborhood. Although the average pooling features can transfer the global information of the image more completely, their calculation methods are easily affected by background clutter and occlusion, and cannot highlight the difference between pedestrians and the background. Compared with the average pooling, the max pooling can reduce the impact of background clutter, but the max pooling pays more attention to extracting the local salient features of pedestrian images and the contour information of pedestrians. The pooling features cannot completely contain the whole body information of pedestrians. In the actual recognition environment, due to the change of camera angle and external illumination, it is necessary to remove the influence of background clutter while preserving the whole body information of pedestrians and highlighting the difference between pedestrians and background. On this basis, this paper proposes a residual pooling module, as shown in Figure 3. By combining the advantages of max pooling and average pooling, this module can make up for the shortcomings of average pooling and max pooling, and on the basis of preserving the whole body information, highlight the pedestrian contour and focus on the difference between the pedestrian and the background, to make the final feature expression of pedestrian images more comprehensive and discriminatory and improve the accuracy of pedestrian re-identification.

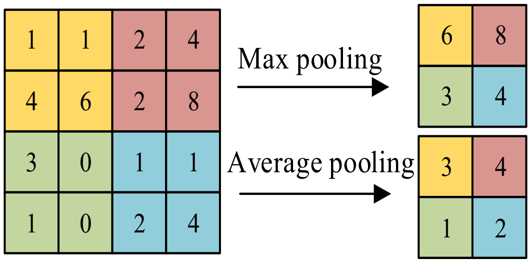


Figure 2. Schematic diagram of pooling layer.

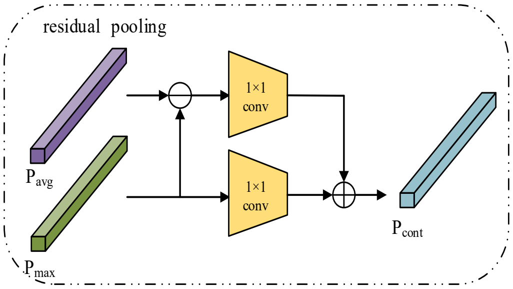


Figure 3. Structure of residual pooling module.

As can be seen from Figure 3, the residual pooling module subtracts the multi-scale feature Pavg of a pedestrian image extracted by Resnet50 from Pmax and uses a convolutional kernel of 1×1 to obtain the difference between Pavg and Pmax. The Pmax feature obtained through max pooling also uses a convolutional kernel of 1×1, and adds the difference feature between Pavg and Pmax to obtain the residual feature Pcont. The residual pooling module integrates the merits of max pooling and average pooling. The obtained residual feature Pcont not only covers the whole body of pedestrians and deepens the contour of pedestrians, but also reduces the impact of background clutter, and pays more attention to the difference between pedestrians and background. The residual feature Pcont is calculated as shown in equation (1):

$$P_{\mathrm{cont}} = \delta_{1\times1}(P_{\mathrm{max}}) + \delta_{1\times1}(P_{\mathrm{avg}} \text{-} P_{\mathrm{max}}) \tag{1}$$

where: Pavg and Pmax are the features obtained after average pooling and max pooling respectively; $\delta_{1\times1}(\boldsymbol{x})$ are used for 1×1.

3.3. Loss Function

We use the joint optimization model of triple margin loss and cross-entropy loss to train the model. The loss function is shown in equation (2):

$$L_{\mathrm{total\_loss}} = \frac{1}{6}\sum_{i=1}^{6} L_{\mathrm{Triplet}}^{i} + \frac{1}{3}\sum_{j=1}^{3} L_{\mathrm{CE}}^{j} \tag{2}$$

where: $L_{\mathrm{total\_loss}}$ is a total loss. L_{Triplet}^{i} is triple loss. L_{CE}^{j} is the cross-entropy loss. Triple loss makes the distance between positive sample pairs shorter and the distance between positive and negative samples larger. Cross entropy loss focuses on the closeness between the actual output and the expected output, where, $i \in [1,6]$ in L_{Triplet}^{i}, *i* represents the *i*th feature among the six basic pedestrian image features Pavg1, Pmax1, Pavg2, Pmax2, Pavg3, and Pmax3 extracted after passing through layer3 and layer4 of the ResNet50 network. *j* represents the *j*th feature among the three comparison features Pcont1, Pcont2 and Pcont3 extracted through the residual pooling module. The loss function designed in this paper adopts the joint optimization model of triple loss and cross-entropy loss. The convergence of the model is accelerated by calculating multiple losses.

4. EXPERIMENT

To achieve objective comparative analysis, this experiment includes the following four training strategies: 1) The learning rate uses Warmup mode in the training stage; 2) Random erasure the data of the training set with a probability of 0.5; 3) Label smoothing is used to improve the generalization performance of the model; 4) BNNeck is used to normalize features. In addition, we use the mean value of three repeated experiments as the experimental results to avoid randomness and ensure the accuracy of the experimental results.

4.1 Experimental dataset

This paper's comparative laboratory is based on the market1501[16] and Duke MTMC Reid[17] datasets.

Market1501[16] includs 1501 pedestrians captured by 6 cameras. Training set: 12936 images of 751 pedestrians. test set: 19732 images of an additional 750 pedestrians. Test set includes: query set and gallery set. MTMC-ReID[17] dataset includes 36411 images of 1404 pedestrians captured by 8 cameras. Training set:16522 images of 702 pedestrians. Test set: 17661 images of an additional 702 pedestrians. Test set includes: query set and gallery set.

4.2 Experimental preparation

The algorithm is implemented on PyTorch framework. We are experimenting with a computing platform with a GPU model NVIDIA GTX1080Ti. CPU model Intel® Core™ i7-7700k @ 4.20 GHz, the memory is 32GB. When training the model, the resolution of the input pedestrian image is set to 288×144 pixels, 32 training batches, and 220 iterations in total. We use SGD optimizer, set the learning rate to 0.03 and the weight decay rate 0.0005. When iteration ordinal number gradually increases to 0.003, it then drops to 0.003, 0.0003, and 0.0003 at 40, 110, and 150 iterations, respectively.

4.3 Experimental evaluation criteria

In this experiment, Rank-1, Rank-5, Rank-10 and mean Average Precision (mAP) in the cumulative matching features (CMC) curve are used as evaluation indicators. During the test, take a query image from the query set and measure the similarity between all the images in the test set and the query image. CMC refers to the probability of successful matching with the query image in the first K candidate images. The values of Rank-1, Rank-5, and Rank-10 are the corresponding accuracy when *K*=1, 5, and 10 in CMC (K). the mAP is the average of the areas under the accuracy recall curve for all samples.

4.4 Analysis of experimental results

To validate the proposed method, a comparative experiment is carried out on Market1501[16] and DukeMTMC-reID[17] datasets. Figure 4 shows the re-identification results. The first column of each row is the query image, the last 10 columns are the top 10 query results. In Figure 4, the solid line border represents the correct query results, and the dotted line border represents the wrong query results.

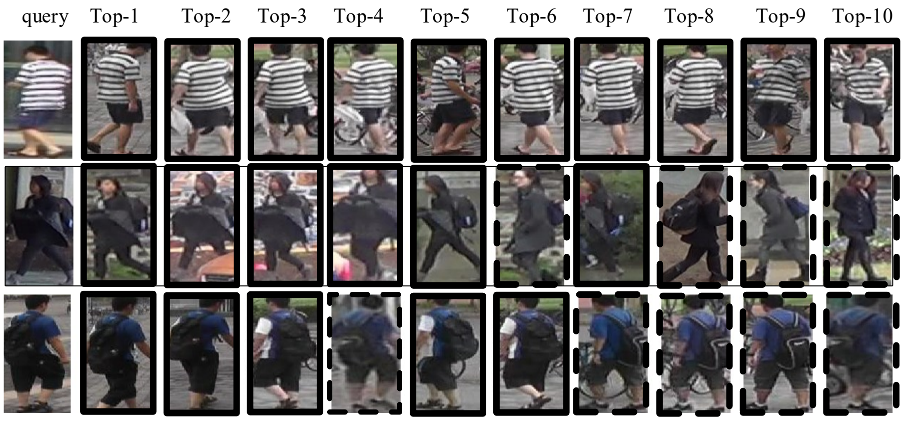
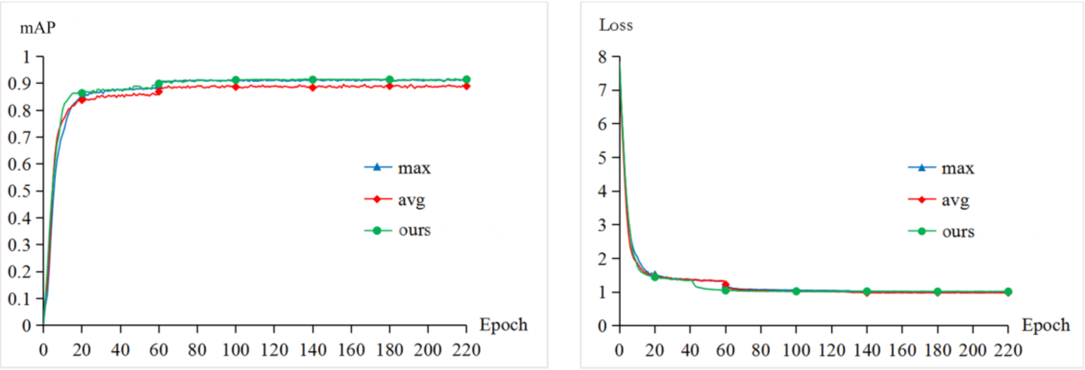


Figure 4. Person re-identification results of the proposed method.

The AVG method is to pool the basic features of images by using our pedestrian re-identification network structure to obtain the feature maps. Max method is to maximize the pool and get the feature graph. Figure 5 shows the comparison of accuracy and loss value between avg method, max method, and residual pool module (residual) method on the Market1501[16] dataset. The accuracy of our method is higher than avg method and max method, and the loss value decreases faster. It is verified that the residual pooling feature obtained based on the residual pooling module can reach better performance in pedestrian re-identification tasks.

In this paper, Grad-CAM avg is used to visualize avg, max method, and residual methods. Max method and residual method are visualized by using the Grad-CAM class activation thermodynamic diagram. Figure 6 shows that the AVG method performs well for the acquisition of whole body information, but it is also easily affected by the background clutter; the max method focuses on to the local pedestrian contour but does not include the whole pedestrian body. Our method combines the advantages of both methods, which can include the whole body of pedestrians and reduce the impact of background clutter.



(a) Accuracy values comparison

(b) Loss values comparison

Figure 5. Accuracy and loss values comparison among different methods.

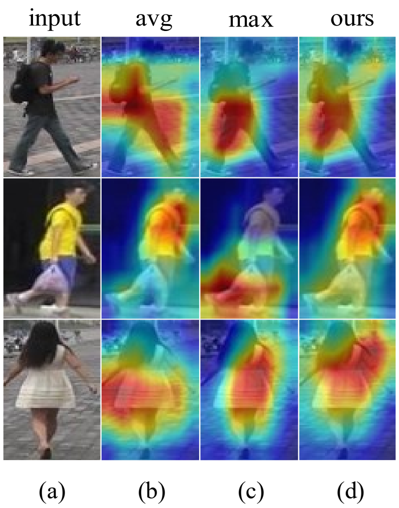


Figure 6. Visualization results of different methods. (a): input images; (b): avg method; (c): max method; (d): residual method.

Table 2 shows the comparison of several mainstream methods' indicators on the Market150[16] and DukeMTMC-reID[17] datasets. Our residual +re ranking method optimizes the network performance by re-ranking[18] technology, ResNet50_baseline method directly utilizes the features of layer 4 output of the ResNet50 network to optimize the network performance. The mAP of the residual +re-ranking methodon Market1501 and DukeMTMC-reID datasets is 94.52% and 89.30% respectively. This method's indicators can be significantly improved with resnet50 as the backbone network.

Table 2. Performance indicators comparison 1 among different methods (%).

Methods	Market1501				DukeMTMC-reID			
	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10
ResNet50_baseline	64.67	79.63	91.86	93.88	32.54	51.79	67.43	71.89
Avg	65.97	85.36	94.12	96.41	45.51	70.83	84.07	88.38
Max	74.18	89.82	96.26	97.80	52.10	71.10	83.62	87.88
Residual	87.68	95.01	98.34	99.17	77.70	88.64	94.57	96.90
Residual + re-ranking	94.52	96.41	98.22	98.90	89.30	91.43	95.20	96.77

Table 3 shows the indicators of Residual, Residual +re-ranking and the representative methods (SVDNet, GLAD[8], PCB[9] PCB+RPP[9] BEF, etc.), The indicators of the comparison methods are all quoted from the original text, where "—" means that there is no such experimental result in the original text. At the same time, the indicators after using re-ranking technology to optimize our method are given. Rank1 and mAP of Residual +re-ranking method are better than the current advanced methods, especially in the mAP indicators. On the Market1501[16] dataset, after reordering, the proposed method (residual + re-ranking) is 2.61% higher than the PCB+RPP method based on local features in Rank1 and about 13% higher in mAP. Compared with the BEF method, Rank1 and mAP are increased by 1.1% and 7.8%, respectively; Compared with the DG-Net method, Rank1 and mAP are increased by 1.61% and 8.52%, respectively; It is 2.2% higher than that of CtF method in Rank1 and 9.62% points higher in mAP. On the DukeMTMC-reID[17] dataset, compared with the BEF, the Rank1 of this method (residual) is lower, but the mAP is increased by 1.7%. After reordering, the Rank1 and mAP of Residual +re-ranking are higher than the representative methods.

Table 4 shows the comparison of parameter quantity, calculation quantity, and inference time between this method and the comparison method. Compared with most representative methods, the method in Residual has more model Parameters, but the method in Residual has less computation (FLOPs) and consumes less time for reasoning in a single image.

Table 3. Performance indicators comparison 1 among different methods (%).

Methods	Market1501				DukeMTMC-reID			
	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10
SVDNet[19]	62.10	82.30	92.30	95.20	56.80	76.70	86.40	89.90
GLAD	73.90	89.90	—	—	62.20	80.00	—	—
PCB	77.40	92.30	97.20	98.20	66.10	81.70	89.70	91.90
PCB+RPP	81.60	93.80	97.50	98.50	69.20	83.30	90.50	92.50
BEF[20]	86.70	95.30	—	—	76.00	89.00	—	—
APR[21]	66.89	87.04	95.10	96.42	55.56	73.92	—	—
DG-Net[22]	86.00	94.80	—	—	74.80	86.60	—	—
CtF[23]	84.90	94.20	—	—	75.60	86.90	—	—
Residual	87.68	95.01	98.34	99.17	77.70	88.64	94.57	96.90
Residual +re-ranking	94.52	96.41	98.22	98.90	89.30	91.43	95.20	96.77

Table 4. Parameters, calculation, and inference time comparison among different methods.

Methods	Parameters/M	FLOPs/G	Inference time/ms
SVDNet	26.02	10.81	4.15
GLAD	26.07	17.74	5.61
PCB	23.72	11.91	4.38
PCB+RPP	23.74	11.92	4.38
BEF	27.59	12.93	4.59
APR	27.74	12.94	4.59
DG-Net	31.54	12.94	4.60
CtF	37.54	12.95	4.60
Residual	30.55	10.82	4.15

5. CONCLUSION

In order to solve the problem of low feature expression ability in person re-identification network, combined with the advantages of max pooling and average pooling, we propose a multi-scale residual pooling pedestrian re-identification method, which makes up for the deficiency of max pooling and average pooling by extracting different scale features in the network and inputting the residual pooling module, and can make the network focus on the difference between pedestrians and background in the image. Experiments on Market1501 and DukeMTMC-reID datasets show that, our method can significantly increase the accuracy, compared with SVDNet, GLAD, and PCB methods, and its first hit rate and average accuracy on Market1501 datasets are 96.41% and 94.52% respectively. In the future, more prominent features will be extracted by using deformable convolution or introducing an attention mechanism, to further improve the performance indicators of pedestrian re-identification task.

REFERENCES

[1] Song, W., Zhao, Q., Chen, C., et al., "Survey on pedestrian re-identification research," CAAI Transactions on Intelligent Systems 12(6), 770-780 (2017). (in Chinese)

[2] Ojala, T., Pietikainen, M. and Maenpaa, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971-987 (2002).

[3] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision 60(2), 91-110 (2004).

[4] Bazzani, L., Crisyani, M., Perina, A., et al., "Multiple-shot person re-identification by chromatic and epitomic analyses," Pattern Recognition Letters 33(7), 898-903 (2012).

[5] Liao, S., Hu, Y., Zhu, X., et al., "Person re-identification by local maximal occurrence representation and metric learning," IEEE Conf. on Computer Vision and Pattern Recognition, 2197-2206 (2015).

[6] Koestinger, M., Hirzer, M., Wohlhart, P., et al., "Large scale metric learning from equivalence constraints," IEEE Conference on Computer Vision and Pattern Recognition, 2288-2295 (2012).

[7] Weinberger, K. Q. and Saul, L. K., "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research 10(2), 207-244 (2009).

[8] Wei, L. H., Zhang, S. L., Yao, H. T., et al., "GLAD: Global local-alignment descriptor for scalable person re-identification," IEEE Transactions on Multimedia 21(4), 986-999 (2018).

[9] Sun, Y. F., Zheng, L., Yang, Y., et al., "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," European Conf. on Computer Vision, 480-496 (2018).

[10] LeCun, Y., Bottou, L., Bengio, Y., et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86(11), 2278-2324 (1998).

[11] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 25, 1097-1105 (2012).

[12] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556v6, (2014).

[13] Lin, M., Chen, Q. and Yan, S., "Network in network," arXiv:1312.4400v3, (2013).

[14] Szegedy, C., Liu, W., Jia, Y., et al., "Going deeper with convolutions," IEEE Conf. on Computer Vision and Pattern Recognition, 1-9 (2015).

[15] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 770-778 (2016).

[16] Zheng, L., Shen, L. Y., Tian, L., et al., "Scalable person re-identification: A benchmark," IEEE Inter. Conf. on Computer Vision, 1116-1124 (2015).

[17] Zheng, Z. D., Zheng, L. and Yang, Y., "Unlabeled sample generated by gan improve the person re-identification baseline in vitro," IEEE Inter. Conf. on Computer Vision, 3774-3782 (2017).

[18] Zhong, Z., Zheng, L., Cao, D. L., et al., "Re-ranking person re-identification with k-reciprocal encoding," IEEE Conf. on Computer Vision and Pattern Recognition, 1318-1327 (2017).

[19] Sun, Y. F., Zheng, L., Deng, W. J., et al., "SVDNet for pedestrian retrieval," IEEE Inter. Conf. on Computer Vision, 3800-3808 (2017).

[20] Dai, Z. Z., Chen, M. V. Q., Gu, X. D., et al., "Batch feature erasing for person re-identification and beyond," arXiv:1811.07130v2, (2019).

[21] Lin, Y. T., Zheng, L., Zheng, Z. D., et al., "Improving person re-identification by attribute and identity learning," Pattern Recognition 95(C), 151-161 (2019).

[22] Zheng, Z. D., Yang, X. D., Yu, Z. D., et al., "Joint discriminative and generative learning for person re-identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2138-2147 (2019).

[23] Wang, G. A., Gong, S. G., Cheng, J., et al., "Faster person re-identification," European Conf. on Computer Vision, 275-292 (2020).