Person Re-identification Using Multi-branch Cooperative Network

Yongchao Xu(✉), Fengyuan Zhang, and Yao Hu

College of Information Science and Engineering, Northeastern University, Shenyang l10819, China
1626377654@qq.com

Abstract. Person Re-identification (Re-ID) is to match the images of the given person across multiple non-overlapping cameras. To solve the problems of occlusions and unconstrained poses, we propose a multi-branch cooperative network for person Re-ID. First, attention branch and multi-scale branch are designed respectively. In the attention branch, we design shade module, random erasing module and stepped module, and guide each module to learn discriminative features of different regions through the consistent activation penalty function. In the multi-scale branch, we design a global and a local module to learn deep features to improve the performance of person Re-ID. Finally, the two branches are cascaded to concatenate multi-branch features. Extensive experiments on Market-1501, DukeMTMC-reID and CUHK03 demonstrate that the proposed method outperforms the state-of-the-art methods.

Keywords: Person re-identification · Multi-branch network · Attention branches · Robustness

1 Introduction

Person re-identification (Re-ID) identifying a target person in different scenarios, is an important subject in the field of computer vision [1]. It is widely used in video surveillance and autonomous driving. Early research efforts mainly focus on hand-crafted construction. With the vigorous development of deep learning, convolutional neural networks (CNNs) have become the predominant choices for Re-ID [2], achieving better recognition performances than traditional methods. However, due to various complicated factors, such as body poses and occlusions, learning robust and discriminative features is still a difficult and challenging task.

Re-ID based on CNNs can be summarized into 1) spliting images or feature maps into some horizontal grids. The PCB model [3] implicitly divides images into horizontal grids of multiple scales directly, ignoring the relations between body parts. 2) utilizing a pose estimator to extract a pose map. Wei [4] uses key points positioning to predict and estimate human body poses, which effectively solves the difficulty of person feature alignment, but requires a large amount of additionally labeled data for model training and prediction, and the retrieval accuracy is largely limited by the performance of the

Y. Wang et al. (Eds.): IGTA 2022, CCIS 1611, pp. 181–192, 2022.
https://doi.org/10.1007/978-981-19-5096-4_14

model. 3) leveraging generative adversarial networks (GANs) to generate more images. Zheng [5] uses GAN to generate more simulated data for data enhancement and improve the generalization ability of the model. However, it is easy to generate noisy samples, which significantly affects the accuracy and performance of the model. 4) computing attention maps to focus on a few key parts, but the extracted regions may not contain discriminative body parts, missing some important data.

To solve the above problems, in this paper, we propose a cooperative network based on multi-branch, which can effectively extract more discriminative features. In the shade module, the framework focuses on extracting features from low-response parts. In the stepped module, it focuses on reducing complex and noisy background cutter. Besides, combined with random erasing module, we use consistency activation penalty (CAP) function to ensure that the high activation regions of three networks do not overlap. In the multi-scale branch, We propose a branch to extract different levels of characteristics, which effectively preserves the integrity of pedestrian features. Finally, multiple branches are combined to form a complete multi-branch cooperative network, which can effectively deal with problems such as occlusions, poses changes.

In summary, our contributions can be summarized as follows:

(1) We propose a branch that integrates multiple attention mechanisms. Through erasing high-response regions, we can generate more complex occlusion samples. We also use random erasing module to simulate low-quality samples in the real world. Besides, relations between different parts of pedestrians can be effectively extracted with the help of stepped module. We effectively combine the three modules through the consistent activation penalty function, so as to improve the model's feature extraction ability for samples with less information and solve the problem of low model accuracy.
(2) We use a multi-scale branch to extract local features of persons, and combine them with global features to learn the relation between person parts and mine non-significant information. Consequently, the recognition ability and accuracy of the algorithm are improved significantly.
(3) Experimental results on three large-scale person Re-ID datasets including Market-1501, DukeMTMC-reID, and CUHK03 prove that the proposed methods exceeds state-of-the-art methods.

2 Multi-branch Cooperative Network

We propose a person Re-ID method based on multiple attention mechanisms and multi-scale branches. The multiple attention mechanisms are used to improve the adaptability of the model to occlusions, pose changes, illumination, low resolution and other factors, while the multi-scale branch is used to improve the ability to fuse and extract global and local features. The complete multi-branch cooperative network structure is shown in Fig. 1.

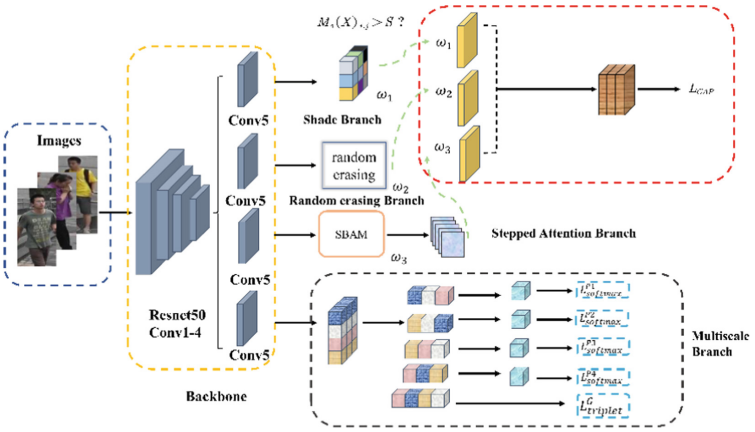

Fig. 1. The proposed network structure

2.1 Shade Module

In the actual application of the algorithm, it is inevitable to encounter the phenomenon of occlusions in pedestrian pictures. Occlusions will cause partial loss of pedestrian features and affect the integrity of features. When distinguishing features are lost, the recognition performance of the model will decline significantly. In order to extract more discriminative pedestrian features, we block the areas with high spatial attention response and retain the feature maps with low response, so as to generate more difficult samples.

The spatial attention model used in this paper is calculated as follows:

$$\mathrm{N_s(Y)} = \mathrm{BN}\left\{\mathrm{C}_2^{2\times1}\left(\mathrm{C}_1^{3\times3}\left(\mathrm{C}_0^{1\times1}(\mathrm{Y})\right)\right)\right\}. \tag{1}$$

where *BN* is the data normalization operation, *C* is the convolution operation, the upper right corner is the size of the convolution kernel, and *Ns(Y)* is the spatial attention map.

After the spatial attention response map is obtained, the high response region is set to 0 and the low response region remains unchanged, so as to obtain the mask $\tilde{N}(Y)$, which is calculated as follows:

$$\tilde{\mathrm{N}}(\mathrm{Y}) = \begin{cases} 0, \mathrm{Ns(Y)i,j} > \mathrm{S} \\ \mathrm{Ns(Y)i,\ j, others} \end{cases}. \tag{2}$$

where $N_s(Y)_{i,\,j}$ represents the value of the spatial attention map at positions *i* and *j*, and S represents the threshold, that is, when the response value is greater than S, it is set to 0, otherwise it remains unchanged.

By setting the high response region to 0, the effect of forcing the model to learn distinctive features from the low response region is realized, and the recognition performance of the model is improved.

2.2 Random Erasing Module

We set the original image as *M*, the image size as *W* and *H*, the image area as *U*, the erasure probability *P*, the random initialization erasure area as *Us*, the value range of

Us/U is set as *(U1, U2)*, the erasure aspect ratio is *Qs*, the value range of *Qs* is set as *(q1,1/q1)*, *Ms* is the random erasure rectangular box, and the random initialization P1*(Xs,Ys)* is the coordinate point randomly selected in the image *M*.

$$O(x,y) = \text{random}(x, y). \tag{3}$$

where *random()* is the random number generation function.

The random erasing algorithm can be described as follows: Input the pedestrian image, and the random initialization probability is P_1, if P_1 is greater than the erasure probability P, the original image is directly output. Otherwise, the erasing area and aspect ratio are randomly initialized according to the erasing area and image length and width range. The coordinate point P*(Xs, Ys)* is initialized randomly. When the erasing area is set to be smaller than the image size, the random erasing area is randomly selected *(0,255)* for assignment to achieve the effect of random erasing. The specific process is shown in the following pseudocode (Table 1):

Table 1. The process of random erasing algorithm

Algorithm: Random Erasing Algorithm

Input: $M, W, H, P, U, U_1, U_2, q$

Output: I'

Initialize $P_1 = O(0,1)$
if $P_1 \le P$ **then**
  $U_s = O(U_1, U_2) \times U$
  $Q_s = O(q_1, 1/q_1)$
  $H_s = \sqrt{U_s \times Q_s}$
  $W_s = \sqrt{U_s \div Q_s}$
  $X_s = O(0, W)$
  $Y_s = O(0, H)$
end if
while $X_s + Y_s \le W, Y_s + H_s \le H$ **do**
  $I_s = (X_s, Y_s, X_s + W_s, Y_s + H_s)$
  $I(I_s) = O(0,255)$
end while
  $I' = I$

2.3 Stepped Module

The traditional feature segmentation method adopts horizontal segmentation, which pays more attention to different regions of pedestrians, but it is easy to ignore the local relations between pedestrians and the information that may exist at the edge of the block.

As shown in Fig. 2, the PCB algorithm divides each row into six horizontal slices, the handbag and umbrella of pedestrians are separated into different blocks between different blocks, which will cause the loss of important information such as edge information and local relation of the blocks, making it impossible to obtain the ideal effect when analyzing each block individually.

We use the method of dividing 8 slices in a stepped manner, as shown in Fig. 2, starting from the first block, every four blocks are taken as a relatively complete local area, which moves down continuously, and finally five block areas are obtained, as shown in the following figure a, b, c, d, e. We observed that the blocks d and e retain the complete information of handbag and umbrella. At the same time, because the cut is smaller than the original feature map, the noise and background cutter are reduced, so the recognition accuracy and performance are significantly improved.

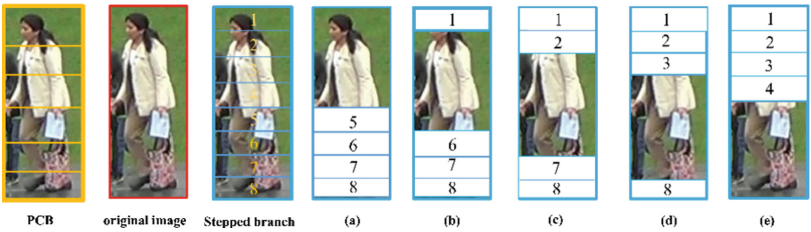

Fig. 2. Ladder block method

In the previous methods based on feature space segmentation, each horizontal slice enjoys the same weight, and the details such as umbrella and handbag can not be highlighted effectively. In this paper, we assign relatively larger weights to the blocks containing more important information, so that the model can focus on the parts with strong resolution. For this branch, an image is input through conv5_x to get the feature map $F \in R^{C\times H\times W}$, and then the feature map F is input into SBAM module at branch 2 and branch 3. In the SBAM module, firstly, step block is carried out, and F is divided into 8 horizontal parts. Every four parts are grouped to obtain a local region. The starting block of the local region moves downward in step 1 from the first block, and finally five local regions are extracted, of which each feature region is $F_i \in R^{C\times\left(\frac{H}{2}\right)\times W}(i = 1, 2, 3, 4, 5)$, Then focusing on each F_i, first it is compressed in the spatial dimension, and the calculation method is as follows:

$$\mathrm{F_i^{'}} = \mathrm{FC_2(FC_1\left(avg_s(F_i)\right) + FC_2(FC_1(max_s(F_i))).} \tag{4}$$

where avg_s and max_s are the average pooling and maximum pooling of the input data in the spatial dimension respectively, and two one-dimensional vectors are obtained after compression.

FC1 and FC2 are used as shared parts to compress and restore the two vectors on the channel. Finally, they are added and fused to get $F_i^{'} \in R^{C\times 1\times 1}$. In order to give weight to each local region, $F_i^{'}$ is compressed in the channel dimension and expressed as

$$\mathrm{s_i} = \mathrm{Sum_c(F_i^{'}).} \tag{5}$$

$$\mathrm{m_i} = \mathrm{Max_c(F_i^{'}).} \tag{6}$$

where Sum_c and Max_c are respectively the sum and maximum value of the input data in the channel dimension. Eventually, we can get $s_i \in R^{C\times 1\times 1}$ and $m_i \in R^{C\times 1\times 1}$.

According to the calculation of s_i and m_i of each local area, the proportion can be calculated. The calculation formula is:

$$V_i = \lambda(F_{sum}(s_i) + F_{max}(m_i)). \tag{7}$$

where λ is set to 6 according to the calculation, and finally the proportion value is adjusted between 0 and 1 through the sigmoid function. The original local area F_i is multiplied by the adjusted proportion value to obtain the update result of the local area $S_i \in R^{C\times\left(\frac{H}{2}\right)\times W}$, namely:

$$S_i = F_i \times \text{sigmoid}(V_i). \tag{8}$$

2.4 Multi-scale Branch

For fine-grained pedestrian feature extraction, the existing method can obtain fine pedestrian features by horizontally segmenting the features. However, due to the local misalignment and occlusion of pedestrian images, it is easy to produce wrong matching.

At the same time, due to the separate existence of each segment, complete pedestrian characteristics cannot be perceived. In contrast, in the multi-scale branch, fine-grained global module and local module are designed to refine the representation of pedestrian features, and achieve the effective feature extraction of "global + local" (Fig. 3).

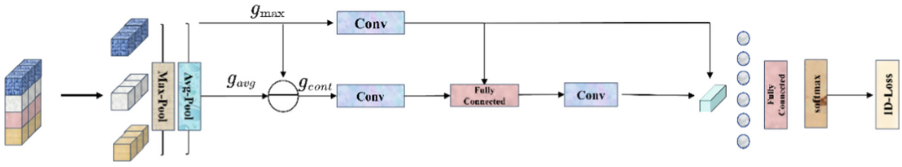

Fig. 3. Multiscale local branch

For the fine-grained global module, the size of the feature map obtained in the convolutional neural networks (CNNs) is $R^{C\times H\times W}$ (C represents the number of channels, H represents the height, and W represents the width). In the H dimension, we divide the feature map into N parts, and perform maximum pooling and average pooling operations on each part respectively to obtain feature vectors $g_{maxi}(i = 1, 2, 3\cdots n)$ and $g_{avgi}(i = 1, 2, 3\cdots n)$. Afterwards, the max-pooling and average-pooling results of different parts are concatenated respectively to obtain the description vectors G_{max} and G_{avg} of the fine-grained global branch. In this paper, triplet loss is used to train G_{max} and G_{avg}, which can realize the integrity of information while considering local correlation, and achieve effective identification of similar parts of different persons.

For fine-grained local modules, different local blocks are considered separately. Global average pooling is easy to introduce background or local noise information, while global maximum pooling can overcome noise interference, but cannot consider all information in the same layer.

Therefore, it is possible to better extract the information between the same layers by performing the difference operation for the two. The specific calculation is as follows:

$$g_{cont} = g_{aug} - g_{max}. \tag{9}$$

After that, we perform convolution dimension reduction for g_{max} and g_{cont} respectively to obtain g'_{max} and g'_{cont}, and then we perform cascade and convolution dimension reduction to obtain g_{inter}. Finally, the global max pooling result g'_{max} and g'_{cont} are recombined as the layer feature $\tilde{L}_1$ of the same layer, which is calculated as follows:

$$\widetilde{L_1} = g'_{max} + g'_{inter}. \tag{10}$$

Finally, we connect through the full connection layer and conduct joint training through softmax and ID-Loss. By analyzing the correlation between non-adjacent parts, more significant potential information can be mined. In addition, through the analysis of discarded local information, the actual situation of local occlusion is effectively simulated, which enhances the robustness and discrimination of the model.

2.5 CAP Network

For the cascade of three branches of different attention mechanisms, in order to make different branches focus on different regions of the image and enhance the diversity and comprehensiveness of local feature extraction, CAP network is introduced in this paper to coordinate different attention branches and make each branch focus on different regions with different characteristics. Different weights are assigned to different branches through LAN, and Hellinger distance [7] is used to measure the consistency of output weights of different branches:

$$H(\omega_i, \omega_j) = \frac{1}{\sqrt{2}} \left\| \sqrt{\omega_i} - \sqrt{\omega_j} \right\|_2. \tag{11}$$

where the sum of the elements of ω_i and ω_j is 1, then the square of the above formula can be obtained:

$$H^2(\omega_i, \omega_j) = 1 - \sum \sqrt{\omega_i \omega_j}. \tag{12}$$

In order to ensure that the high activation regions of different attention models do not overlap, it is necessary to maximize the distance between ω_i and ω_j, that is, to minimize the value of $\sum \sqrt{\omega_i \omega_j}$, then the CAP loss can be defined as follows:

$$L_{CAP} = \sum \sqrt{\omega_i \omega_j}. \tag{13}$$

Through the above formula, it can be optimized to diversify the local feature extraction and enhance the representation ability of the model.

3 Experiments

3.1 Datasets

Experiments are conducted on three commonly-used large-scale person Re-ID datasets. The Market-1501 (Zheng et al., 2015) [8] dataset contains 32,668 images of 1,501 persons captured by 6 cameras. In the experiment, a total of 12,936 images of 751 persons are used as the training set, and a total of 19,732 images of 750 persons are used as the test set. The DukeMTMC-reID (Ristani et al., 2016) [9] dataset contains 36,411 images of 1,404 persons captured by 8 cameras. In the experiment, a total of 16,522 images of 702 persons are used as the training set, and a total of 17,66l images of 702 persons are used as the test set. The CUHK03 (Li et al., 2014) [10] dataset contains 14,097 images of 1,467 persons captured by 10 cameras. The experiment uses 767 persons samples as the training set and 700 persons samples as the test set.

3.2 Implementation Details

We use the Pytorch framework based on deep learning, and use the GPU RTX2080Ti server for training. During the training process, the input image size is adjusted to 384 * 128, and data enhancement methods such as random flipping and random cropping are used. In the experiment, the training batch is 32, where P is 8 and K is 4. ResNet50 pre-trained on ImageNet is used as the backbone network. In this paper, SGD is used as the optimizer, and the learning rate is set to 8e-4. The weight decay is 5e-4, the momentum is set to 0.9, and the number of training iterations (epoches) of the whole network is set to 300. This paper uses the cumulative matching characteristic curve (CMC) and the mean average precision (mAP) to analyze and evaluate the performance of the algorithm. Among them, Rank-1 represents the ratio of finding the person to be queried in the first search results. mAP is the average of the area under the accuracy-recall curve of all query samples, which reflects the overall performance of person Re-ID methods.

3.3 Experimental Results

In order to verify the effectiveness of the method proposed in this paper, the method proposed in this paper is compared with existing pedestrian re-identification methods, including PCB [3], MGN [11], PCB + RPP [3], Harmonized Attention Convolutional Neural Network (HA-CNN) [12], Second-Order Non-Local Attention (SONA) [13], AlignedReID [14], HONet [15], GCP [16], CDNet [17], PAT [18]. Table 2 shows the comparison of experimental results on the three datasets.

Table 2. Comparison of experimental results of different algorithms on data sets

Methods	CUHK03-Labeled		CUHK03-Deteced		Market1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB(ECCV2018)	-	-	54.2	61.3	77.3	92.4	63.5	81.9
MGN(MM2018)	67.4	68.0	66.0	66.8	86.9	95.7	78.4	88.7
PCB + RPP(ECCV2018)	-	-	57.5	63.7	81.0	93.1	68.5	82.9
HA-CNN(CVPR2018)	41.0	44.4	38.6	41.7	75.7	95.6	63.8	80.5
SONA(ICCV2019)	79.2	81.9	76.4	79.1	88.7	95.7	78.1	89.3
AlignedReID(PR2019)	-	-	59.6	61.5	79.1	91.8	69.7	82.1
HONet(CVPR2020)	-	-	-	-	84.9	94.2	75.6	86.9
GCP(AAAI2020)	75.6	77.9	69.6	74.4	88.9	95.2	78.6	89.7
CDNet(CVPR2021)	-	-	-	-	86.0	95.1	76.8	88.6
PAT(CVPR2021)	-	-	-	-	88.0	95.4	78.2	88.8
Ours	**82.4**	**84.6**	**81.5**	**82.7**	**90.1**	**96.2**	**81.1**	**90.2**

As shown in Table 2, compared with the PCB + RPP method, our method has significant improvements on Market-1501, DukeMTMC-reID and CUHK03 datasets. The Rank-1 indicators increased by 3.8%, 8.3%, and 21.4% respectively. And the mAP indicators increased by 12.8%, 17.6%, and 27.3% respectively. The reason is that the multi-branch method considers the relationship between different parts of the human body, better characterizing person information through the joint collaboration of multiple branches, so the experimental results are significantly improved. Compared with the GCP method, we achieve significant improvements on all three datasets. The main reason is that although the GCP method adopts the relation analysis module, it does not pay enough attention to the global features. In this paper, the global and local training are combined on the multi-scale branch, and the extracted features are more discriminative. Thereby a better effect is achieved. Compared with the MGN method, the branch set in this paper is more reasonable. Compared with the current SONA with the highest accuracy, all indicators in this paper have been significantly improved on the three datasets. The model in this paper has achieved best results compared with the existing algorithms with better effects, which verifies the robustness of the model and effectively improves the accuracy of the pedestrian re-identification algorithm.

3.4 Ablation Experiments

In order to further verify the effectiveness of the multi-branch proposed in this paper, we analyze the model from both qualitative and quantitative aspects. First, we test the experimental effect of the baseline network of ResNet50 pre-trained on ImageNet. After that, the experiments of multiple branches and mutual combination methods are added respectively. The specific comparison results are shown in the following table.

Table 3. Ablation study on three datasets

Methods	CUHK03-Labeled		CUHK03-Detected		Market1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	60.2	63.3	54.7	60.1	81.1	92.7	71.0	82.3
Baseline+SB	67.1	68.4	66.4	66.8	83.3	93.4	74.4	83.4
Baseline+RE	66.3	70.5	62.3	63.5	82.5	93.1	73.9	82.8
Baseline+SBAM	72.5	74.1	71.5	74.4	86.4	95.4	77.8	88.2
Baseline+SB+RE	73.4	76.9	70.6	72.3	84.8	94.1	76.5	85.5
Baseline+SB+SBAM	76.7	78.2	74.4	76.0	87.5	95.8	79.2	89.1
Baseline+RE+SBAM	75.6	77.3	73.8	75.5	86.8	95.5	78.6	88.4
Baseline+SB+RE+SBAM	78.8	78.6	77.1	78.7	89.1	95.9	80.1	89.3
Baseline+SB+RE+SBAM+CAP	80.1	81.5	80.2	81.5	89.9	96.1	80.7	89.9
Baseline+MB	79.9	80.1	78.4	77.2	88.3	94.4	77.4	84.4
Baseline+SB+RE+SBAM+CAP+MB	**82.4**	**84.6**	**81.5**	**82.7**	**90.1**	**96.2**	**81.1**	**90.2**

The construction of the multi-branch cooperative network is to extract the features of pedestrians more effectively, improving the recognition performance of the model and achieving good recognition results in more complex environments and conditions. Table 3 shows the comparison between the baseline method and the model proposed in this paper. It can be seen from the table that the recognition accuracy of the baseline method is the lowest, and multiple branches are better than the baseline method whether used alone or in combination. For the attention branch, the introduction of the CAP network can significantly improve the results, which verifies the rationality of the network structure. At the same time, the results of the network including all branches are better than that of each branch working alone, and Rank-1 and mAP are significantly improved. It shows that there is a complementary and cooperative relationship between different branches, and person features with different levels of discrimination can be extracted respectively, which verifies the effectiveness of the model design in this paper.

3.5 Query Results Display

As can be seen from the Fig. 4, the recognition accuracy of the proposed baseline method is not high, and the error rate in top10 recognition results is high. However, the features learned by the attention branch and the multi-scale branch can complement each other, which significantly improves the recognition accuracy.

Fig. 4. Market1501 dataset recognition examples

4 Conclusions

Designing multi-branch networks to learn rich feature representation is one of the important directions in person re-identification (Re-ID). However, when extracting pedestrian features, the regions and non-significant regions where the model has significant recognition ability should be extracted as much as possible to enhance the robustness and recognition performance of the model. Therefore, this paper proposes a joint network based on multi-branch cooperation. Through the occlusion module, the random erasing module and the stepped module, strong person features are jointly extracted, and the CAP network is used to ensure the diversity of local feature extraction. It exploits the potential connections of non-adjacent parts through multi-scale branching, and combines global features to construct a high-precision person Re-ID network. The experimental results on three public person Re-ID datasets show that the multi-branch cooperative network proposed in this paper extract more discriminative and robust pedestrian features.

References

1. Luo, H., Jiang, W., Fan, X., Zhang, S.P.: A survey on deep learning based person re-identification. Acta Automatica Sinica **45**(11), 2032–2049 (2019)

2. Zhu, F.Q., Kong, X.W., Fu, H.Y., Tian, Q.: two-stream complementary symmetrical CNN architecture for person re-identification. J. Image Graph. **23**(7), 1052–1060 (2018)
3. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 501–518. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_30
4. Wei, L.H., Zhang, S.L., Yao, H.T.: GLAD: global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, pp. 420–428 (2017)
5. Zheng, Z.D., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of 2017 IEEE International Conference on Computer Vision, Venice, pp. 3774–3782. IEEE Press (2017)
6. Li, W., Zhu, X., Gong, S.G.: Harmonious attention network for person re-identification. In: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 2285–2294. IEEE Press (2018)
7. Li, S., Bak, S., Carr, P.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 369–378. IEEE Press (2018)
8. Zheng, L., Shen, L.Y., Tian, L., Wang, S.J., Wang, J.D., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of 2015 IEEE International Conference or Computer Vision, Santiago, pp. 1116–1124. IEEE Press (2015)
9. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
10. Li, W., Zhao, R., Xiao, T., Wang, X.G.: DeepRe ID: deep fliter pairing neural network for person eidinition. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, pp. 152–159. IEEE Press (2014)
11. Wang, G.S., Yuan, Y.F., Chen, X., Li, J.W., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, New York, pp. 274–282. IEEE Press (2018)
12. Fu, Y., et al.: Horizontal Pyramid Matching for Person Re-identification (2018)
13. Xia, B., Gong, Y., Zhang, Y.Z., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Washington, pp. 3759–3768. IEEE Press (2019)
14. Luo, H., Jiang, W., Zhang, X., Fang, X., Qian, J.J., Zhang, C.: AlignedReID++: dynamically matching local information for person re-identification. Pattern Recogn. **94**, 53–61 (2019)
15. Wang, G.A., et al.: High-order information matters: learning relation and topology for occluded person re-identification. In: Proceedings of 2020 IEEE/CVF International Conference on Computer Vision, Seattle, pp. 3759–3768. IEEE Press (2020)
16. Park, H., Ham, B.: Relation Network for Person Re-identification (2020)
17. Li, H.J., Wu, G.J., Zheng, W.S.: Combined depth space based architecture search for person re-identification. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. IEEE Press (2021)
18. Li, Y.L., He, J.F., Zhang, T.Z., Liu, X., Zhang, Y.D., Wu, F.: Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. IEEE Press (2021)