A Person Re-Identification Network Based upon Channel Attention and Self-Attention

2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP) | 978-1-6654-3904-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSIP52628.2021.9688968

Mengzhe Sun, Zhaohui Wang
School of Computer Science and Technology
Hainan University
Haikou, China
mengzhe_sun@163.com

***Abstract*—With the development of deep learning, person Re-Identification (Re-ID) technology has made great achievements. However, there are still some problems, such as pedestrian occlusion, different imaging conditions of posture changes, etc, which can lead to large changes in appearance, and it is difficult to obtain sufficient distinguishable features. Therefore, this paper proposes a network based on the fusion of channel attention mechanism and self-attention mechanism. The network learns more discriminative global features and local features from the spatial dimension and channel dimension. ResNet-50 is utilized as the backbone network. The channel attention module can capture the dependence of channel dimensions, obtain the weight of the importance of feature channels, and improve useful local feature to suppress useless feature. The self-attention module captures the context information from the spatial dimension to obtain the weight of each feature, and further obtains the global feature. Experiments on two datasets reveal the proposed model improves the accuracy compared with the state-of-the-art methods.**

***Keywords*—person re-identification, ResNet-50, self-attention, channel attention**

I. INTRODUCTION

Person Re-Identification (Re-ID) [1] aims to recognize a specific pedestrian image across non-overlapping cameras. With the popularity of monitoring equipment, person Re-ID has been widely used in security monitoring and suspect tracking [2]. Although the emergence of deep learning has advanced the frontier person Re-ID greatly, it is still an open challenge. Due to different imaging conditions of pedestrian occlusion and pose variance, the appearance of image changes greatly, which makes it difficult to obtain robust features. As shown in the first row of Fig.1, Due to the spatial misalignment of human body posture changes, the body parts of the two images do not match. Therefore, there are obvious differences in detecting the same area of the feature maps. From the second row of Fig.1, the back of the pedestrian in the left image is blocked by the umbrella and the body of the pedestrian in the image on the right is blocked by the railing, which increases the difficulty of recognition. In response to the above challenges, focusing on learning the robust and distinguishing feature representation has become the key to solving the person Re-ID problem.

In the past few years, many methods based on hand-crafted features are proposed to extract reliable features for person Re-ID, such as extracting color and texture feature [3]-[5] or learning powerful similarity metrics [6]-[8]. However, these methods rely on the prior knowledge of the designer, but in the complex environment, the generalization ability of feature is weak.

Fig. 1. The Challenge of person Re-ID

In recent years, deep learning methods based on convolutional neural network (CNN) have achieved the dominant status in the field of computer vision. Deep learning has been widely used in person Re-ID task. Most methods focus on using CNN to learn the global features of whole-body images. For example, Ahmed et al. [9] proposed to use Siamese network to extract global features to determine whether the input image pairs belong to the same person. Considering that the extraction of global feature is hard to achieve the expected effect when the target image is occlusion, some methods are proposed to jointly learn whole-body images and body parts images. Common ideas for extraction include image segmentation and pose extraction, etc, Sun et al. [10] utilized the horizontal segmentation method to divide the pedestrian image into 6 identical blocks to extract local features. At the same time, the refined part pooling was applied to purify the evenly divided blocks. For the pedestrian misalignment problem, the pose estimation model is utilized in the paper [11] to estimate the key points of the pedestrian, and then the local features are extracted based on the key points for spatial alignment. However, these methods ignore the difference in the

importance of fusing the local features of body parts and lead to recognition failure.

Although the above methods have made tremendous progress, there are still room for improvement. In order to enhance the ability to distinguish features, an attention mechanism is introduced to learn the local features of the target image. Liu et al. [12] proposed an attention model, which dynamically generates local attention features from global images in a cyclic way to generate discriminative features. Li et al. [13] proposed an attention model that combines the hard attention and soft attention to define body parts from full-body images and simultaneously learn multi-scale feature maps. Liu et al. [14] proposed to use a multi-directional attention mechanism to capture multiple attention information with different levels of features. However, these methods have deficiency that they all focus on using full-body images or local regions for attention learning, but to a large extent ignore the local feature information learning in the channel dimension. Therefore, when the circumstances of large pose change and partial occlusion, the sub-optimal of person ReID performance is obtained.

Therefore, this paper proposes a dual attention fusion network, which can fully capture global and local features from the perspective of spatial and channel dimensions to learn the discriminating features. The core components of the network are the self-attention module and the channel attention SE module . The self-attention module focuses on the non-local features and the captures spatial features by acquiring the contextual information. The SE module extracts serviceable channel features by modeling the interdependence of channels. To effectively optimize the proposed network, the commonly used cross-entropy loss function is combined with the hard sample mining triplet loss function for person Re-ID. Experimental results reveal the proposed method effectively improves the recognition performance of person Re-ID on two public datasets Market1501 [15] and DukeMTMC-reID [16].

The remaining of this paper is organized as follows. Section II describes the proposed approach in detail. Section III discusses the experimental results. Section IV summarizes the paper.

II. PROPOSED METHOD

A. The Proposed Network

The proposed network consists of a ResNet-50 network, SE module and self-attention module. The network structure is shown in Fig. 2.

We choose the classic ResNet-50 network as the backbone network because of their deep network structure, which avoids the disappearance and degradation of gradients in the deep network to obtain typical features. The SE module is embedded in the ResNet-50 network to effectively extract the local features of the channel. The ResNet-50 network outputs a complete feature map and puts it into the self-attention module to extract spatial features. Then global average pooling (GAP) is performed on the final output feature map to further extract global features. Finally, the Fully Connected (FC) layer is mapped to the classification space.

In the stage of data preprocessing, in order to solve the occlusion problem, random erasure enhancement (REA) [17] is applied to enhance the training image. The combination of cross-entropy loss with label smoothing regularization [18] and the triplet loss of difficult sample sampling [19] is used as the final loss for training our model.

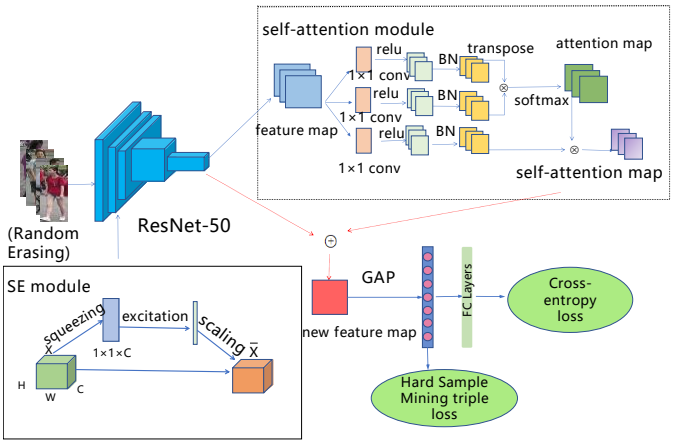

Fig. 2. Framework of the proposed network.

B. Squeeze-and-Excitation(SE) Module

Hu et al.[20] proposed a " Squeeze-and-Excitation(SE)" network based on the channel dimension. As shown in the solid line box in Figure 2, the feature attention unit in SE module consists of three parts, squeezing, excitation and scaling. Squeezing is to perform GAP on the feature map obtained by H×W convolution, and generate a 1×1×c vector. Excitation refers to the formation of a Bottleneck structure through two fully connected layers to model the correlation between channels. The number of channels is first compressed, then the number of channels is reconstructed, so that the number of parameters of the model can be reduced, and the complexity between model channels can be reduced. Finally, it is converted into a normalized weight of 0~1 through the Sigmoid function. Scaling means that the normalized weights obtained are weighted to the features of each channel.

C. Self-Attention Model

Self-attention mechanism [21] is usually applied to text representation. But recent studies have shown that self-attention has great potential for building image recognition models. The attention function can be seen as an operation that maps a set of queries and key-value pairs to the output, where query, key, value, and output are all vectors. The output is a weighted sum of values. Input includes query, key, and value are exploited to dot product operation. First, the calculation of the dot product of the query and all keys divided by the square root of the dimension, and then use the softmax function to get the weight. Note the attention function is defined as shown in formula (1).

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

where d_k is the dimension of Q and K, d_V is the dimension of V, and softmax calculates the weight.

As shown in the dotted box of Fig. 2, this abstract proposes a novel self-attention module based on the above mechanism.

The features extracted by ResNet-50 are transferred to the self-attention model. First, the 1×1 convolution layer is linearly combined with feature maps of different scales and different spaces. After the batch normalization layer (BN), the normalized features are obtained. BN is made the distribution of the input feature map relatively stable and the model learning speed is accelerated. The first two branches are transposed and multiplied to obtain the attention score, then softmax is used in the column direction perform normalization to get the attention matrix. The matrix elements are shown in formula (2). The last attention matrix and the last branch are matrix multiplied on the self-attention feature map.

$$\gamma_{j,i} = \frac{\exp(A_{ij})}{\sum_{i=1}^{M} \exp(A_{ij})} \tag{2}$$

where $\gamma_{i,j}$ is each element in the obtained attention map matrix; A_{ij} is each element in the scale feature map matrix; M is the number of elements in the scale feature map matrix.

In addition, the self-attention feature map is multiplied by a weight, added it to the input feature map, and finally returned to the new feature map. Therefore, the final output is as shown in the formula (3).

$$y = \partial o + x \tag{3}$$

Where x is the original input feature map, o is the self-attention feature map, α is the weight of the self-attention feature map, and y is the final feature map.

III. Experimental

A. Datasets

The effectiveness of the proposed model is evaluated on two mainstream public datasets, Market1501 [15] and DukeMTMC-reID [16]. Market1501 dataset contains 32,668 pedestrian pictures of 1501 pedestrians taken by 6 cameras, the training set has 751 people, including 12,936 pictures, and the gallery set has 750 people, including 19,732 pictures. DukeMTMC-reID dataset, which is a subset of the DukeMTMC dataset, is collected from Duke University. These images are from 8 cameras. The dataset contains 36411 pictures of 1812 pedestrians. The training set contains 16522 pictures of 702 pedestrians. The gallery set contains 2228 pictures of 702 pedestrians, and the remaining 408 people is used as distractors.

B. Evaluation Metrics

Two evaluation indicators are applied to evaluate the performance of our person Re-ID, accuracy of Rank-n and mean average precision(mAP). Rank-n represents the accuracy rate of the top N images in image recognition, for example Rank-1 represents the accuracy rate of the image with the highest similarity,while mAP represents the average accuracy of all categories according to the precision rate and recall rate, and can measure the performance of the network.

C. Training Details

The proposed model is based on the ResNet-50 network. The network initialization weight comes from the ImageNet pre-trained model. All the input images are 384×128. The training of the model is optimized by Adam. The learning rate is 0.00035. Random rotation, translation and REA are exploited as data enhancement for datasets. 500 epochs are trained, the batchsize is set to 32, and the marginal parameter of triplet loss is 0.3. The experiments are performed on a computer equipped with the graphics process unit(GPU) of NVIDIA GeForce GTX 1080 Ti.

D. Experiments Results

The proposed method is compared with other four algorithms to test the recognition performance of the proposed network. Table 1 shows that the proposed network has the highest performance on Rank-1 and mAP. On the Market-1501 dataset, Rank-1 and mAP are 1.1% and 6% higher than the fourth algorithm named PCB. On the DukeMTMC-reID dataset, Rank-1 and mAP are 1.1% and 0.4% higher. It is worth noting that the PCB is divided into 6 blocks according to the local characteristics of pedestrians, and each horizontal block is classified. However, the importance of local features of body parts is ignored, which leads to the failure of occluded pedestrian image recognition. Compared with these models, the proposed model reies on channel attention and spatial attention to obtain distinguishable global and local features to improve recognition performance.

TABLE I. The Accuracy (%) Of Comparision with Other Methods on DukeMTMC-ReID, Market1501

Methods	Market1501		DukMTMC-reID	
	Rank-1	***mAP***	***Rank-1***	***mAP***
SVDNet[22]	82.30	62.10	76.70	56.80
APR[23]	84.29	64.67	70.69	51.88
DaRe[24]	89.00	76.00	80.20	59.00
PCB[8]	92.30	77.40	81.80	66.10
Ours	**93.40**	**83.40**	**82.30**	**66.50**

E. Ablation Studies

In order to verify the enhancement effect of each proposed attention module on the ResNet-50 network, different ablation experiments were designed on the Market-1501 and DukeMTMC-reID datasets. The experimental parameters are identical.

As shown in Table 2, SE represents the channel attention module and SA represents the self-attention module. in order to verify the effectiveness of the SE module, the first combination is to embed the SE module into the ResNet-50 network. Compared with ResNet-50 network, on the Market1501 dataset, Rank-1 and mAP have improved by 5% and 7.7%, on the DukeMTMC-reID dataset, Rank-1 and mAP increased by 5.2% and 6.8%, respectively. The reason is that the SE module models the correlation of each channel and mines channel information of beneficial features to improve the discrimination of local features. In order to test the influence of SA module on the performance of the model, the second combination is to add SA to the ResNet-50 network. The Rank-1 and mAP of the Market1501 dataset added by 3% and 4.5%, respectively, and the Rank-1 and mAP of the DukeMTMC-reID dataset increased by 2.3% and 6.8%, respectively. Becasue the SA module can obtain context information by learning the dependence of pixel, obtain the global information of the feature map, and finally normalize the feature map to improve the generalization ability of the network. The last combination integrates all the modules together. Experimental results show that it achieves the most advanced performance in two datasets. The Rank-1 and mAP on the Market1501 dataset are promoted by 7.67% and 6.6%, respectively, and the Rank-1 and mAP on the DukeMTMC-reID dataset are respectively improved by 6% and 7.0%, which shows that the proposed network can effectively capture the local characteristics of the channel and the spatial global characteristics for improving the recognition effect.

TABLE II. THE ACCURACY (%) OF DIFFERENT COMBINATIONS IN THE BASIC NETWORK

Methods	Market1501		DukMTMC-reID	
	Rank-1	*mAP*	*Rank-1*	*mAP*
ResNet-50	85.76	76.80	76.30	59.50
ResNet-50+SE	92.50	83.10	81.50	66.30
ResNet-50+SA	90.50	81.30	78.00	64.30
ResNet-50+SE+SA	**93.40**	**83.40**	**82.30**	**66.50**

Fig. 3. The top five query visual results on the Market1501 dataset

The visualization results of this paper based on the Market-1501 dataset are shown in Fig. 3. The blue digital superscript images and the query images to be queried belong to the same pedestrian. The red number superscript images and the query images to be queried do not belong to the same pedestrian. This shows that the proposed model has better discriminative ability than the ResNet-50 network structure.

IV. CONCLUSION

In this paper, a network of channel attention mechanism and self-attention mechanism has been proposed, which aims to learn enough discriminative features from input image to improve the accuracy of person Re-ID. As the crucial component of the network, the proposed network includes channel attention module and self-attention module. The channel attention module models the dependency of feature channels, which enables the network model to selectively obtain more critical image information. The self-attention mechanism module obtains global spatial feature information by obtaining the weight of each feature. Through two attention modules, the fine-grained local features and global features of pedestrians are obtained respectively. Therefore, the more discriminating features are captured. Experimental results demonstrate that the proposed method can obtain more robust features and improve accuracy of recognition.

REFERENCES

[1] M Ye, J Shen, G Lin, T Xiang, SCH Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP(99):1-1.

[2] KW Chen, CC Lai, PJ Lee, CS Chen, YP Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," IEEE Transactions on Multimedia, 2011, 13(4): 625-638.

[3] S Khamis, CH Kuo, VK Singh,VD Shet, LS Davis, "Joint learning for attribute consistent person re-identification," Springer International Publishing, 2014.

[4] WS Zheng, SH Gong, T Xiang, "Reidentification by relative distance comparison," IEEE transactions on pattern analysis and machine intelligence, 2012, 35(3): 653-668.

[5] Y Yang, J Yang, J Yan, S Liao, Y Dong, SZ Li, "Salient color names for person re-identification," 2014:536-551.

[6] X Yang, M Wang, D Tao, "Person re-identification with metric learning using privileged information," IEEE Transactions on Image Processing, 2018.

[7] X Wang, WS Zheng, L Xiang, J Zhang, "Cross-scenario transfer person reidentification" IEEE Transactions on Circuits & Systems for Video Technology, 2016, 26(8):1447-1460.

[8] YC Chen, WS Zheng, J Lai, "Mirror representation for modeling view-specific transform in person re-identification," International Joint Conference of Artificial Intelligence (IJCAI). AAAI Press, 2015.

[9] E Ahmed, M Jones, TK Marks, "An improved deep learning architecture for person re-identification," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[10] Y SUN, L Zheng, Q Tian, S Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," Springer, Cham, 2017.

[11] S Chi, J Li, J Xing, T Qi, "Pose-driven deep convolutional model for person re-identification," 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[12] Liu, S Gong, C Loy, X Lin, "Person re-identification: what features are important?," Springer, Berlin, Heidelberg, 2012.

[13] C Wang, Q Zhang, C Huang, W Liu, X Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," Proceedings of the European Conference on Computer Vision (ECCV). 2018: 365-381.

[14] X Liu, H Zhao, M Tian, L Sheng, J Shao, S Yi,et al, “Hydraplus-net: attentive deep features for pedestrian analysis,” ICCV (2017), pp. 350-359.

[15] L Zheng, L Shen, T Lu, S Wang, T Qi, “Scalable person re-identification: A benchmark,” 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.

[16] E Ristani, F Solera, R Zou, RS Zou, R Cucchiara, C Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” European Conference on Computer Vision. Springer, Cham, 2016.

[17] Z Zhong, L Zheng, G Kang, SZ Li, Y Yang “Random erasing data augmentation,” Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 13001-13008.

[18] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, Z Wojna, “Rethinking the inception architecture for computer vision,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016:2818-2826.

[19] A Hermans, L Beyer, B Leibe, “In defense of the triplet loss for person reidentification”, (2017).

[20] H Jie, S Li, S Gang, S Albanie, “Squeeze-and-excitation networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[21] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gpmez, “Attention is all you need”, Advances in neural information processing systems. 2017: 5998-6008.

[22] Y Sun, L Zheng, W Deng, S Wang, “Svdnet for pedestrian retrieval”, 2017 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2017.

[23] Y Lin, L Zheng, Z Zheng, Y Wu, Z Hu, C Yan, et al, “Improving person re-identification by attribute and identity learning,” Pattern recognition, 2019, 95(C):151-161.

[24] Y Wang, L Wang, Y You, X Zou, V Chen, S Li, et al, “Resource aware person re-identification across multiple resolutions,” IEEE. IEEE, 2018.