Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.017926
Article

Tech Science Press

Person Re-Identification Based on Joint Loss and Multiple Attention Mechanism

Yong Li and Xipeng Wang*

College of Information Engineering, Engineering University of PAP, Xi'an, 710086, China
*Corresponding Author: Xipeng Wang. Email: wanghp369@sina.com
Received: 17 February 2021; Accepted: 15 April 2021

Abstract: Person re-identification (ReID) is the use of computer vision and machine learning techniques to determine whether the pedestrians in the two images under different cameras are the same person. It can also be regarded as a matching retrieval task for person targets in different scenes. The research focuses on how to obtain effective person features from images with occlusion, angle change, and target attitude change. Based on the present difficulties and challenges in ReID, the paper proposes a ReID method based on joint loss and multi-attention network. It improves the person re-identification algorithm based on global characteristics, introduces spatial attention and channel attention and constructs joint loss function, improving the characteristic extraction ability of the network and improving the model performance of person re-identification. It analyzes the validity and necessity of each module of the algorithm through the ablation experiment. In addition, it carries out training and evaluation on the two person re-recognition data sets Market1501 and MSMT17, and verifies the advantages of the proposed algorithm in contrast to other advanced algorithms.

Keywords: Person re-identification; attention mechanism; loss function; global feature

1 Introduction

In recent years, the maturity of face recognition technology has enabled algorithms to surpass humans in the accuracy of identifying other people's faces, and has also been widely used in the construction of "smart cities" and "safe cities". However, in actual application scenarios, the camera cannot capture a clear face under any circumstances, and one camera often cannot cover all areas. Therefore, it becomes very necessary to use the person's whole body information to lock and find people. By taking the overall person characteristics as an important supplement to the face, the cross-camera tracking of persons can be realized. Person re-recognition is the use of computer vision and machine learning techniques to determine whether the persons in the two images under different cameras are the same person. It can also be regarded as a matching retrieval task for person targets in different scenes.

With the successful application of convolutional neural network in image task, deep features from the deep learning net is increasingly used for feature representation of person image, rather than manual feature [1]. Person re-identification can be divided into two categories according to different feature learning

strategies: global features and local features [2]. The global feature extracts the global information of each person image, and the local feature extracts the feature of a certain area in the image, and then merges multiple local features. Tang et al. [3] extracted color and texture of each region of the human body as local features. And local features were fused with gait energy image features. Ahmed et al. [4] proposed a deep learning framework to solve the problem of person re-identification by entering a pair of images and outputing similarity to indicate whether they belong to the same person, based on the feature learning and similarity measurement methods of person re-identification. Wang et al. [5] combined the single-image feature matching and cross-image feature two-class classification methods in person re-recognition, and the losses of the two were added. Qian et al. [6] proposed a multi-scale deep learning model that can learn features at different scales and automatically determine the most suitable scale for matching. Zhou et al. [7] proposed a full-scale network OSNet, which can not only capture different spatial scales, but also encapsulate any combination of multiple scales. The local feature learning of person re-recognition usually focused on the aggregated features of the local area, so that the model had better robustness. And the local area of the person was usually generated by slicing, pose, segmentation and meshing. Sun et al. [8] proposed the PCB algorithm. It divided each input image into 6 parts evenly, classified these 6 layers of features, and merged the 6 features together to calculate the similarity during testing.

Recently, visual attention mechanism has been widely used in various visual tasks and has also been successfully applied in ReID [9,10]. Chen et al. [11] proposed the higher-order attention module, which improved the discriminability of attention module through the model and higher-order statistical information, so as to capture the subtle differences between persons. Jiao et al. [12] proposed JA-ReID, which combined a soft pixel-level attention mechanism and a hard region-level attention mechanism. The feature extraction capability was improved from three aspects: global scale, regional scale, and fine-grained scale.

Given the deficiency of feature extraction ability in the existing network for ReID, the paper compares the performance improvement effects of several attention mechanisms and loss functions, and proposes a simple and efficient ReID model that combines attention mechanism and circle loss.

2 Related Works

2.1 Deep Learning Global Features

The ReID task aims at associating images taken by the same person from different cameras. The ReID dataset is divided into four parts, namely the verification set, the training set, the query set, and the person gallery. After inputting a probe, the model can find images belonging to the same target from gallery and sort them according to the similarity score.

The training process of ReID model is shown in Fig. 1. First, the model is trained on the training set to obtain the representation of the image in the feature space. After that, the deep learning model is used to extract features from the images in query and gallery and calculate the similarity. For each query, the first N images similar to the image are found in the gallery.

The realization of person recognition can be roughly divided into four parts:

a) Obtain the features of the image through the deep learning network.
b) Extract the features of all images in gallery.
c) Calculate distance (such as Euclidean distance) by probe and Gallery features.
d) Sort according to the calculated distance. The higher the ranking, the higher the similarity.

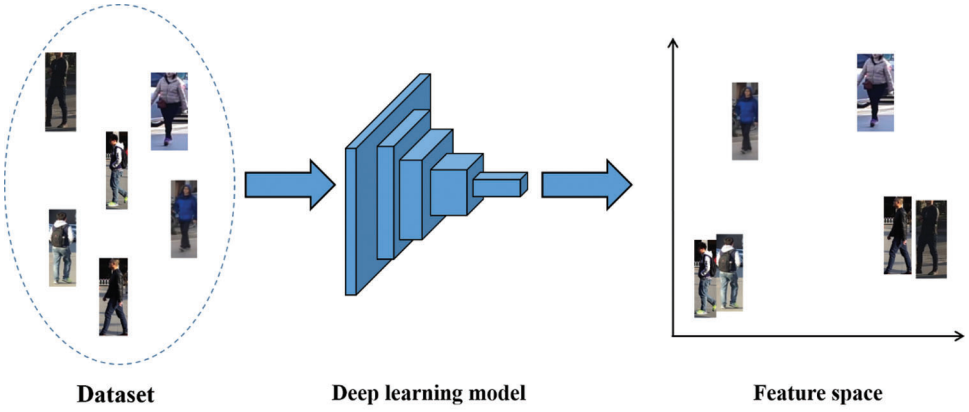

Figure 1: Training process of ReID model

2.2 Attention Module

The Attention Mechanism derives from the study of human vision. Humans, in their cognitive processes, selectively focus on a portion of all information while ignoring the rest. In order to reasonably utilize the limited resources of visual information processing, human beings need to select a specific part of the visual region to focus on. Attention mechanism is widely used in natural language processing and computer vision [13–15]. The convolutional layer of a convolutional neural network (CNN) obtains the output features through the linear combination of the convolution kernel and the original features. The convolution kernel has locality and only combines pixels from a small local area to form the output. The feature vector in the feature map is calculated from the input of receptive field of the previous layer. The larger the receptive field is, the larger the scope of the image that the network can reach, and the more global and semantic features will be included in the output.

The attention mechanism has two main functions: deciding which part needs attention, and allocating limited information processing resources to important parts. The attention mechanism can be regarded as the resource allocation mechanism, which can be understood as the redistribution of resources according to the importance of the attention object. In the structural design of deep neural network, the attention mechanism allocates the network weight.

Wang et al. [16] demonstrated the role of visual attention mechanism and proposed a residual attention network, which was formed by stacking multiple attention modules, each of which included a mask branch and a trunk branch. The mask branch was generated by the feature map, the trunk branch was a convolutional neural network model, and the feature map of the two branches was dotted to obtain the final output feature map. SENet [17] assigned weights to different channels through feature compression (squeeze) and extraction (excitation) operations, so that the model focused more on important channel features while suppressing unimportant channel features. The squeeze part compressed the original feature map H×W×C into 1×1×C through the global pooling layer, which made the receptive area wider. In the excitation part, features 1×1×C were input into a full convoluted layer. The importance of each channel was predicted for the different channel weights. Finally, the channel weights were multiplied by the original feature map to obtain the final feature.

In order to obtain better performance, a lot of current researches on the attention mechanism have continuously proposed more complex attention modules, which has increased the computational burden. In order to solve the contradiction between performance and complexity, Wang et al. [18] proposed a channel attention module ECANet. It improved SENet, which greatly reduced the number of parameters while maintaining high performance. The SE module used two full convolutional layers to determine the

channel weights while the ECA module generated the channel weights through a fast one-dimensional convolution of size k, which was adaptively determined by a function related to the channel dimension. The structure of the ECA module is shown in Fig. 2. We also need a method to integrate multiple attention mechanisms to simply and efficiently improve the discriminativeness of the network extracted features.

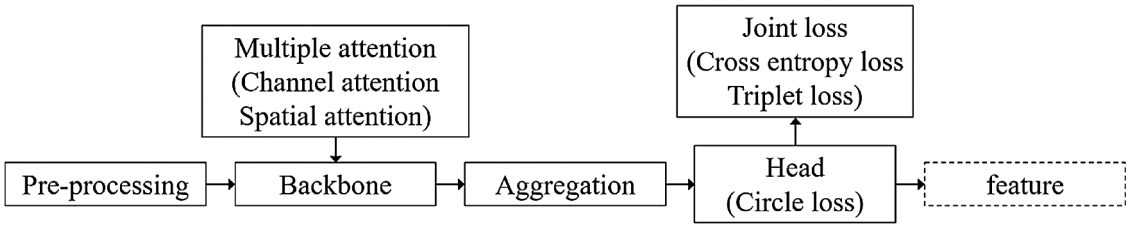

Figure 2: An overview of the proposed algorithm architecture

2.3 Loss Function

The loss function is used to evaluate the degree to which the predicted value of the model is different from the real value.

2.3.1 Cross-entropy Loss

Cross-entropy loss describes the distance between two probability distributions and is usually used in multi-classification tasks. The smaller the cross entropy is, the closer the two samples are. Softmax regression is used to normalize the neural network prediction results to 0~1 and convert them into probability distribution. The calculation formula of cross-entropy loss function in the multi-classification problem is as follows:

$$L_{ce} = -\sum_{i=0}^{C} y_i \log \widehat{y}_i + (1 - y_i) \log(1 - \widehat{y}_i) \tag{1}$$

where C represents the number of categories, $\widehat{y}_i$ is predicted results, $\widehat{y}_i = \frac{e^{W_i^T \mathrm{f}}}{\sum_{i=1}^{C} e^{W_i^T \mathrm{f}}}$, y_i is real results.

2.3.2 Triple Loss

Triplet loss was originally proposed in face recognition [19], as a widely used measure of learning loss. It was used to train samples with small differences. Ding et al. [20] applied triplet loss to ReID task and achieved good results. A triple loss requires the input of three images to make a triple. Anchor images x_i^a, positive examples x_i^p and negative examples x_i^n, x_i^a and x_i^p have the same tag. x_i^a and x_i^n have different tags. Through formula (4), the distance between the samples within the class is smaller than that between the samples.

$$D\left(x_i^a, x_i^p\right) + m < D\left(x_i^a, x_i^n\right) \tag{2}$$

$D(:,:)$ represents the distance between a pair of images, and the triplet loss of N samples is defined as:

$$\sum_{i=1}^{N} \left[m + D\left(g_i^a, g_i^p\right) - D\left(g_i^a, g_i^n\right)\right] \tag{3}$$

where m represents the margin term between a pair of positive and negative samples.

2.3.3 Circle Loss

Sun et al. [21] proposed circle loss, a pin-based similarity optimization method, to maximize the similarity within classes s_p and minimize the similarity between classes s_n. Classification learning and

pairwise learning using positive and negative samples are two basic paradigms for feature learning. The cross-entropy loss uses the class label to optimize the similarity between the sample and the weight vector. Triple loss uses sample pairs to optimize similarity between samples. Circle loss unifies the above two basic paradigms and results in better performance than both methods. x is a single sample in a given eigenspace. Suppose there are K intra-class similarity scores related to X, and L inter-class similarity related to X, referred to as $\left\{s_p^i\right\}(i=1,2,\cdots,K)$ and $\left\{s_n^j\right\}(j=1,2,\cdots,L)$. The circle loss pairs up all s_p and s_n, iterate over the $s_n - s_p$. Going through all the similarities, the difference between them is reduced to obtain the following uniform loss function:

$$\begin{aligned} L_c &= \log\left[1+\sum_{i=1}^{K}\sum_{j=1}^{L}\exp\left(\gamma\left(s_n^j - s_p^i + m\right)\right)\right] \\ &= \log\left[1+\sum_{j=1}^{L}\exp\left(\gamma\left(s_n^j+m\right)\right)\sum_{i=1}^{K}\exp\left(\gamma\left(-s_p^i\right)\right)\right] \end{aligned} \tag{4}$$

The parameter M controls the radius of the decision boundary, and it can be regarded as a relaxation factor and an extension factor.

The paper introduces circle loss into the head module of the net to get the classifier based on circle loss. The features obtained from the backbone are input into the head module, and the processed feature vectors are output, and then the feature vectors are input into the loss function. The loss of cross entropy corresponds to the loss of different ids through classification loss learning, and the loss of triples learns the similarity and differentiation within classes through eigenvectors. In this paper, cross entropy loss and triplet loss are selected simultaneously in the net to construct the joint loss function. The calculation formula is as follows:

$$L = L^{CE} + L^{Tri} \tag{5}$$

3 Methodology

The target of the experiment is to improve the feature extraction ability of ReID network and reduce the influence of occlusion, illumination change, and angle change on the performance. The proposed algorithm improves the benchmark algorithm from two directions: improvement of network with attention mechanism and loss function. Fig. 2 shows the architecture of the algorithm. The paper adds multiple attention mechanisms to the backbone and uses circle loss to design classifier. The Joint loss function is constructed based on cross-entropy loss and triple loss. Finally, the features that can effectively distinguish persons are obtained.

A mature and stable network ResNet [22] is chosen as the backbone network, which has been widely and successfully used in various research fields in recent years, with its proven reliable stability. ResNet is a kind of residual network, which can form a deep network through network stack. Its core lies in residual block with the principle expressed in formula (6) and $F(x_l, W_l)$ as the residual part.

$$x_{l+1} = x_l + F(x_l, W_l) \tag{6}$$

ResNet50 is composed of 6 modules of Conv1, Conv2_X, Conv3_X, Conv4_X, Conv5_X, and pooling layer. ResNet50 is based on residual blocks, with 3+4+6+3=16 as the residual blocks. Conv1 extracts features by 7×7 convolution kernel and reduces image resolution. Conv2_X, 3_x, 4_x and Conv5_X are all made up of residual blocks, with each module expanding the channels for the output feature map to twice the channels for the input, and reducing the length and width of the feature map by half.

In the paper, attention mechanism CBAM is added to the backbone network. CBAM [23] is a kind of attention mechanism module that combines spatial attention and channel attention. The attention module combined and added to the residual block by sequential combination method. The structure of the residual block is shown in Fig. 3. Given the feature map $F \in \mathbb{R}^{C\times H\times W}$, the feature map first passes through the channel attention module and then through the spatial attention module. The process can be expressed as:

$$F' = M_C(F) \otimes F$$
$$F'' = M_S(F') \otimes F' \tag{7}$$

$\otimes$ represents element by element multiplication. In the multiplication process, channel attention will propagate along the spatial dimension and be the output of the attention module.

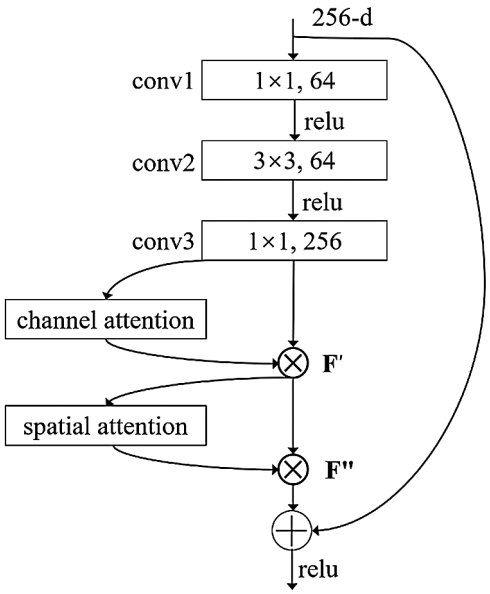

Figure 3: Residual block structure with attention module

The schematic diagram of attention module is shown in Figs. 4 and 5. The attention module uses the maximum pooling and average pooling features to obtain two channel weight matrices 1×1×C. The two weight matrices are input into a multi-layer perceptron composed of two convolution layers, and then the two outputs are added one by one and $M_C(F)$ obtained by the sigmoid activation function.

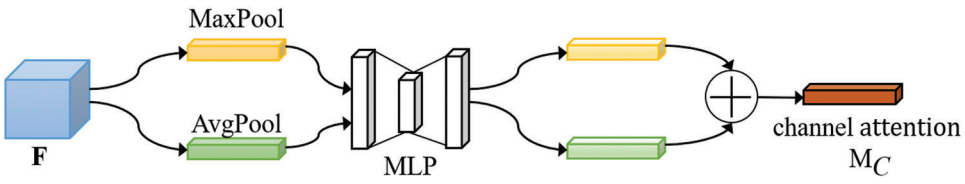

Figure 4: Channel attention module

The spatial attention module also goes through a maximum pooling layer and an average pooling layer. Different from the channel attention module, two feature maps 1×H×W are generated by calculating the

maximum and mean values along the channel dimension, which respectively represent the average pooling features and maximum pooling features in the channel, and then they are spliced together by a convolution layer. $M_S(F)$ is obtained by sigmoid activation function.

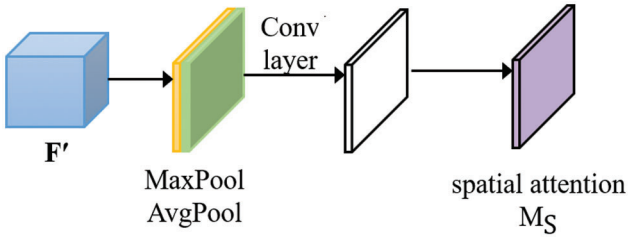

Figure 5: Spatial attention module

4 Experiments

In the experiment, a GTX1080 graphics card is used for calculation, Ubuntu16.04 is used for the operating system, and the program is based on Pytorch1.5 of python3.6.

In order to verify the impact of different attention mechanisms and loss functions on the performance of ReID, ablation experiments are conducted for different attention mechanisms and loss functions. The training and evaluation dataset is Market1501. In order to maintain the consistency of variables, the pre-training model is not used in the ablation experiment, and the input size is set to 256×128, the learning rate is 0.00035, and the batch size is 64. The experimental results are shown in Tab. 1. The Baseline algorithm uses the ResNet50 network and the cross-entropy loss function. Baseline (+Tri) introduces triple loss, Baseline (+Tri Cir) uses a classifier based on circle loss, and the loss function is a joint loss function constructed by cross entropy and triple loss. The remaining algorithms in the table introduce different attention mechanisms on the basis of the baseline algorithm. Through experimental verification, the evaluation result of the proposed algorithm (Ours) is higher than that of other improved algorithms. Rank1 reaches 92.4%, which is 8.73% higher than the baseline algorithm, and mAP reaches 79.7%, which is 14.45% higher than the benchmark algorithm.

Table 1: Experimental results of accuracy verification of different attention mechanisms and loss functions

Methods	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	mAP (%)
Baseline	83.67	93.62	95.78	65.05
Baseline(+Tri)	84.00	94.60	96.53	66.28
Baseline(+SE)	88.06	95.93	97.80	73.43
Baseline(+DANet)	89.52	96.85	97.89	73.43
Baseline(+Non-local)	91.69	96.82	97.92	79.38
Baseline(+ECA)	91.92	96.38	97.98	78.31
Baseline(+Tri Cir)	91.69	97.24	98.28	78.26
Ours	92.4	96.82	98.16	79.70

Tabs. 2, 3 and 4 show the performance comparison of the proposed algorithm with other advanced algorithms. To further improve the performance of the algorithm, Imagenet pretraining model is used in the training. Rank1 and mAP are 94.45% and 85.78% in market1501, 88.6% and 76.9% in DukeMTMC-ReID, 79.44% and 53.57% in MSMT17_V1, respectively. The results on the two datasets are higher than OSNet [7] based on multi-scale feature fusion and attention mechanism IANet [24] proposed in CVPR2019. The experimental results show that the improved strategy with attention mechanism and loss function can effectively improve the performance of ReID.

Table 2: The performance comparison of the proposed algorithm and other algorithms on market1501

Methods	Rank-1 (%)	mAP (%)
SVDNet	82.3	62.1
AACN	85.9	66.9
HAC	89.0	71.25
DPFL	88.6	72.6
DaRe	89.0	76.0
GP-reid	92.2	81.2
PCB	92.3	77.4
Alignedreid	92.6	82.3
PCB+RPP	93.8	81.6
IANet	94.4	83.1
OSNet	94.8	84.9
Ours	94.45	85.78

Table 3: The performance comparison of the proposed algorithm and other algorithms on DukeMTMC-ReID

Methods	Rank-1 (%)	mAP (%)
SVDNet	76.7	56.8
AACN	76.8	58.3
HAC	78.5	60.3
DPFL	79.2	60.0
DaRe	80.2	64.5
GP-reid	85.2	72.8
PCB	81.9	65.3
PCB+RPP	83.3	69.2
IANet	87.1	73.4
OSNet	88.6	73.5
Ours	88.6	76.9

Table 4: The performance comparison of the proposed algorithm and other algorithms on MSMT17_V1

Methods	Rank-1 (%)	mAP (%)
PCB	68.2	40.4
IANet	75.5	45.8
Auto-ReID	78.2	52.5
OSNet	78.7	52.9
Circle Loss	76.9	52.1
Ours	79.44	53.57

Fig. 6 shows the visualization of the proposed algorithm in the Market1501 and the left image is the probe. The first line image is the query result of our algorithm, Sort and display the query results corresponding to the probe according to the accuracy from the largest to the smallest. Here the red box indicates the correct prediction and the blue box indicates the prediction error. The second line shows the annotation images, that is, the correct results. By comparing the first and the second line, it can be found that although the algorithm query results can hit the target, the results are greatly affected by changes in environmental factors such as illumination, angle, and occlusion.

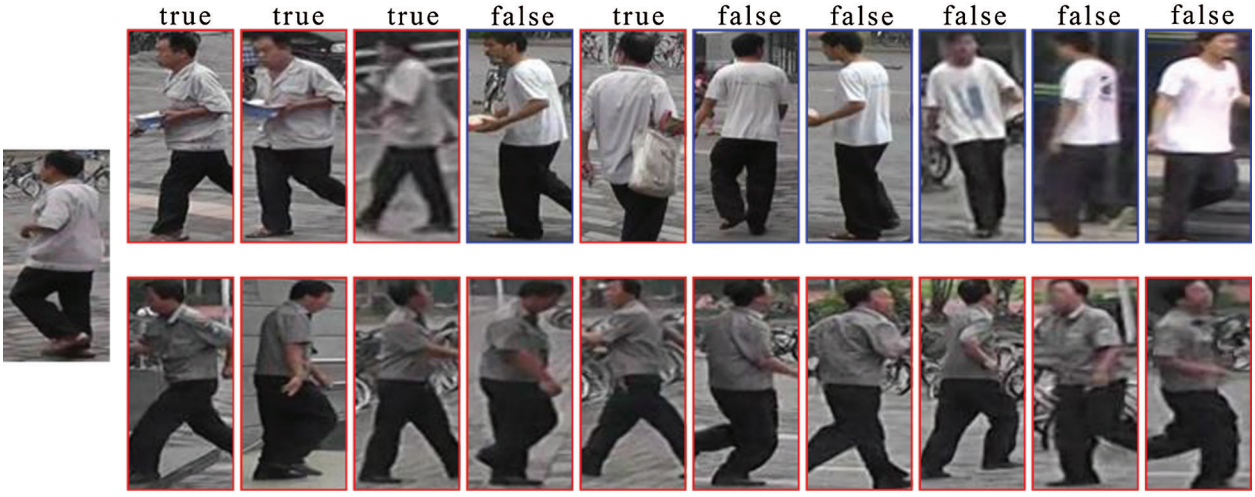

Figure 6: Visualization results of Market1501

5 Conclusion

In this paper, the person re-identification algorithm based on global features is improved, aiming at solving the problems of occlusion, attitude change and angle change in person re-identification. Two improvement strategies are proposed here. First, spatial attention and channel attention are added to ReID residual network to improve the feature extraction capability of the network. Second, joint loss functions is constructed based on cross entropy loss and triplet loss and circle Loss is adopted to design the classifier. By improving the loss function, the intra-class distance is reduced and the inter-class distance is increased. Through ablation experiments, the performance of the algorithm with different attention mechanisms and loss functions is compared. The effectiveness and necessity of each module of the algorithm in this work is analyzed. Finally, the proposed algorithm is trained and evaluated on three ReID

datasets, Market1501, DukeMTMC-ReID and MSMT17, and he the advantages of the proposed algorithm is verified in comparison to other advanced algorithms.

Acknowledgement: This work is supported by Subject of the 13th Five-Year Plan for National Education Science (JYKYB2019012); Foundation of Basic Research of Engineering University of PAP (WJY201907); Foundation of military-civilian integration of Engineering University of PAP (WJM201905)

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Jiang, "A review of person re-identification," *Journal of New Media*, vol. 2, no. 2, pp. 45–60, 2020.

[2] H. Luo, W. Jiang, X. Fan and S. P. Zhang, "A survey on deep learning based person re-identification," *Acta Automatica Sinica*, vol. 45, no. 11, pp. 2032–2049, 2019.

[3] X. L. Tang, X. Sun, Z. Z. Wang, P. P. Yu, N. Cao *et al.,* "Research on the pedestrian re-identification method based on local features and gait energy images," *Computers Materials & Continua*, vol. 64, no. 2, pp. 1185–1198, 2020.

[4] E. Ahmed, M. Jones and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3908–3916, 2015.

[5] F. Q. Wang, W. G. Zuo, L. Lin, D. Zhang and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1288–1296, 2016.

[6] X. L. Qian, Y. W. Fu, Y. G. Jiang, T. Xiang and X. Y. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 5399–5408, 2017.

[7] K. Y. Zhou, Y. X. Yang, A. Cavallaro and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. of the IEEE International Conference on Computer Vision*, Seoul, Korea, pp. 3702–3712, 2019.

[8] Y. F. Sun, L. Zheng, Y. Yang, Q. Tian and S. J. Wang, "Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 480–496, 2018.

[9] S. S. Jiao, J. B. Wang, G. Y. Hu, Z. S. Pan, L. Du *et al.,* "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019.

[10] F. Yang, K. Yan, S. J. Lu, H. Z. Jia, X. D. Xie *et al.,* "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[11] B. H. Chen, W. H. Deng and J. N. Hu, "Mixed high-order attention network for person re-identification," in *Proc. of the IEEE International Conference on Computer Vision*, Seoul, Korea, pp. 371–381, 2019.

[12] S. S. Jiao, J. B. Wang, G. Y. Hu, Z. S. Pan, L. Du *et al.,* "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019.

[13] C. Xu, Z. Q. Su, Q. Jia, D. Z. Zhang, Y. H. Xie *et al.,* "Neural dialogue model with retrieval attention for personalized response generation," *Computers Materials & Continua*, vol. 61, pp. 113–122, 2019.

[14] J. Qiu, Y. Liu, Y. H. Chai, Y. Q. Si, S. Su *et al.,* "Dependency-based local attention approach to neural machine translation," *Computers Materials & Continua*, vol. 58, pp. 547–562, 2019.

[15] Y. Li and H. B. Sun, "An attention-based recognizer for Scene Text," *Journal on Artificial Intelligence*, vol. 2, no. 2, pp. 103–112, 2020.

[16] F. Wang, M. Q. Jiang, C. Qian, S. Yang, C. Li *et al.,* "Residual attention network for image classification," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3156–3164, 2017.

[17] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 7132–7141, 2018.

[18] Q. L. Wang, B. G. Wu, P. F. Zhu, P. H. Li, W. M. Zuo *et al.,* "ECA-net: efficient channel attention for deep convolutional neural networks," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 11534–11542, 2020.

[19] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 815–823, 2015.

[20] S. Y. Ding, L. Lin, G. R. Wang and H. Y. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[21] Y. Sun, C. M. Cheng, Y. H. Zhang, C. Zhang, L. Zheng *et al.,* "Circle loss: a unified perspective of pair similarity optimization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6398–6407, 2020.

[22] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.

[23] S. Woo, J. Park, J. Y. Lee and K. So, "Cbam: convolutional block attention module," in *Proc. of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.

[24] R. Hou, B. P. Ma, H. Chang, X. Q. Gu, S. G. Shan *et al.,* "Interaction-and-aggregation network for person re-identification," in *Proc of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 9317–9326, 2019.