Person Re-Identification Post-print Based on Enhanced Feature Fusion Network

Authors: Deng Tao, Yang Juan, Wang Ronggui, Xue Lixia

Date: 2020-09-28T00:00:00+00:00

Abstract

Person re-identification primarily aims to determine whether pedestrian images captured by different cameras belong to the same individual. In real-world scenarios, factors such as human pose variations, camera viewpoint changes, and background interference cause the same pedestrian to exhibit substantial discrepancies across different cameras, rendering this a challenging task. In recent years, deep learning-based methods have achieved remarkable success in addressing the person re-identification problem. However, most existing methods consider only local or global features of pedestrians in isolation, thereby neglecting the holistic relationships within the pedestrian, i.e., the connection between global and local features. Therefore, this paper proposes an Enhanced Feature Convergent Network (EFCN). In the global branch, an attention network suitable for global feature extraction is proposed as an embedded module and integrated into the backbone network to extract global pedestrian features; in the local branch, a Gated Recurrent Unit Change Network (GRU-CN) is proposed to obtain representative local features, after which a feature fusion method combines global and local features into the final pedestrian representation, with the network being trained using loss functions. Through extensive comparative experiments, the proposed network model achieves favorable results on standard Re-ID datasets. The proposed enhanced feature fusion network can extract highly discriminative pedestrian features; the model is applicable to person re-identification in large-scale non-overlapping multi-camera scenarios, demonstrating high recognition capability and accuracy, and is capable of extracting robust features against background variations in pedestrian images.

Full Text

Enhanced Feature Convergent Network for Person Re-Identification

Deng Tao†, Yang Juan, Wang Ronggui, Xue Lixia
(Dept. of Computer & Information, Hefei University of Technology, Hefei 230601, China)

Abstract

Person re-identification (Re-ID) aims to determine whether pedestrian images captured by different cameras belong to the same individual. In real-world scenarios, this task is extremely challenging due to significant variations caused by pose changes, camera viewpoint shifts, and background interference, which can make the same person appear dramatically different across cameras. In recent years, deep learning-based methods have achieved remarkable success in addressing person Re-ID. However, most existing approaches treat local and global features separately, neglecting the crucial relationship between them—that is, the connection between global and local pedestrian features. To address this limitation, we propose an Enhanced Feature Convergent Network (EFCN). In the global branch, we introduce an attention network specifically designed for global feature extraction as an embedded module within the backbone network to capture comprehensive pedestrian representations. In the local branch, we propose a Gated Recurrent Unit Change Network (GRU-CN) to obtain representative local features, which are then fused with global features through a feature fusion method to generate the final pedestrian descriptor. The network is trained using a composite loss function. Extensive comparative experiments demonstrate that our proposed network achieves competitive results on standard Re-ID datasets. The enhanced feature fusion network can extract highly discriminative pedestrian features, making it applicable to large-scale non-overlapping multi-camera Re-ID scenarios with high recognition capability and accuracy, particularly in extracting robust features for pedestrian images with varying backgrounds.

Keywords: person re-identification; global features; local features; feature fusion

0 Introduction

Person re-identification (Re-ID) is a sub-problem of pedestrian retrieval that involves determining whether pedestrian images captured by different non-overlapping cameras depict the same individual. Re-ID technology enables automatic tracking and retrieval of suspects in video surveillance networks, thereby enhancing system performance and improving case processing efficiency. Given its critical role in video surveillance and public safety, an increasing number of researchers have devoted attention to this problem. The core

challenges in person Re-ID primarily involve feature representation [1] and distance metric learning. Due to variations in pedestrian pose, camera angles, and image quality within surveillance systems, the same person can appear vastly different across cameras, posing significant challenges. These challenges manifest in three main aspects:

First, captured pedestrian images are often misaligned across cameras. As shown in Figure 1: see original paper, for the same person, the red bounding box on the left indicates the head region, while the blue bounding box on the right captures background. Clearly, the feature maps extracted from these two regions by convolutional neural networks exhibit substantial differences and cannot be directly compared. To address misalignment, previous work [2] proposed a multi-feature subspace and kernel learning method for effective identity recognition, while [3] utilized keypoints to generate regions of interest for learning local features and achieving alignment. However, such approaches require training a model to practical performance levels at considerable cost. In contrast, our method employs an attention-embedded network to extract global features and uses horizontal slicing to convert global features into three local features, enabling indirect alignment and yielding significant improvements.

Second, as illustrated in Figure 1: see original paper, many real-world camera images suffer from blur and low quality, increasing Re-ID difficulty. To mitigate this, [4] employed attention mechanisms to focus on local regions of interest, while [5] used attention to focus on local features from top to bottom for similarity comparison, but neglected global feature influence. Additionally, [6] proposed a multi-directional saliency weight learning method that produced better feature representations but suffered from low computational efficiency due to requiring paired input images. Our approach addresses these issues by using attention mechanisms for global feature extraction and horizontal slicing for local features. The proposed GRU-CN network emphasizes important local features, better resolving image blur while reducing background interference.

Third, when distinguishing very similar pedestrian images, subtle details become crucial. While [7] extracted both global and local features to capture details, it ignored correlations between them. [8] proposed a multi-level similarity metric that computed similarity scores at different levels, but incurred high computational costs. Therefore, we propose GRU-CN in the local branch to extract more discriminative local features and design a feature fusion method that tightly integrates global and local features to obtain more representative pedestrian descriptors. This approach better captures pedestrian details and significantly improves recognition accuracy for images with subtle differences.

1 Methodology

For a given query image I_p, the goal of person Re-ID is to find other images of the same identity within a gallery set $G = \{I_i\}_{i=1}^{N}$, where N is the total number of pedestrian images. One solution involves training the network to

learn discriminative feature vectors f for pedestrian identification. This section introduces our proposed enhanced feature network, whose architecture is shown in [Figure 2: see original paper]. The model comprises three main components: a global feature branch, a local feature branch, and a feature fusion branch. The following sections detail each component.

1.1 Global Branch Network

Recent research has increasingly focused on using deep learning to extract discriminative features. Our objective is to learn pedestrian feature maps for identity recognition. For the global branch, we utilize ResNet50 as the backbone network. However, due to the challenges in person Re-ID, using only the basic ResNet50 to learn global features yields insufficiently representative features while introducing interference factors. Therefore, we propose an attention-embedded network called Spatial and Channel Attention Network (SC-Net), which combines with ResNet50 to extract more representative global pedestrian features. We slightly modify ResNet50 by removing the downsampling operation in the fourth layer to obtain larger feature maps of size $2048 \times 24 \times 8$.

The SC-Net aims to enhance feature expressiveness through attention mechanisms: focusing on important features while suppressing unnecessary ones. The embedded attention network SC-Net consists of spatial and channel attention mechanisms. Since convolution operations mix cross-channel and spatial information for feature extraction, this module emphasizes meaningful features across both channel and spatial dimensions. Given a pedestrian image of size $3 \times 384 \times 128$, the image passes through ResNet50 to obtain corresponding feature vectors.

Assume the feature vector $F \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels and H, W represent height and width. The feature vector F is then input to the SC-Net attention-embedded network to obtain feature vector $F' \in \mathbb{R}^{C \times H \times W}$. The specific structure is shown in [Figure 3: see original paper].

First, the feature vector F is reshaped to $F_1 \in \mathbb{R}^{C \times N}$, where $N = H \times W$. The transpose is defined as $F_1^T \in \mathbb{R}^{N \times C}$. To extract channel attention feature maps, we directly compute the channel feature map $M_c \in \mathbb{R}^{C \times C}$ by multiplying F_1 and F_1^T, followed by a softmax layer:

$$M_c(i,j) = \frac{e^{M_{i,j}}}{\sum_{i=1}^{C} e^{M_{i,j}}}$$

where $M_c(i,j)$ measures the impact of the i-th channel on the j-th channel. The resulting feature vector after channel attention is $F_2 \in \mathbb{R}^{C \times N}$, reshaped as:

$$F_2(j,i) = \sum_{i=1}^{C} M_c(i,j) \cdot F_1(i,j)$$

where α is a weight learned from zero. We then perform matrix multiplication between F_2 and F_1^T, where $F_2 \in \mathbb{R}^{C\times N}$ and $F_1^T \in \mathbb{R}^{N\times C}$. The matrix multiplication yields values on a single channel with size $1 \times C$. Finally, we sum these channel values to obtain $F' \in \mathbb{R}^{C\times H\times W}$. This helps improve feature discriminability.

Next, using the channel attention-extracted feature vector F', we apply both average pooling and max pooling to obtain two feature vectors. These are integrated into an effective feature descriptor. Along the channel direction, this effectively highlights important information regions. A convolutional layer then acts on the feature descriptor to produce a spatial attention feature map $M_s \in \mathbb{R}^{H\times W}$, with the computation process given by:

$$M_s(F') = \sigma(f^{7\times 7}([\text{AvgPool}(F'); \text{MaxPool}(F')]))$$

where σ is the sigmoid function and $f^{7\times 7}$ denotes a convolution operation with a 7×7 kernel. F'_{avg} is the feature vector of size $1 \times H \times W$ obtained by average pooling, and F'_{max} is the feature vector of size $1 \times H \times W$ obtained by max pooling.

In summary, for feature vector F, the embedded attention network SC-Net combines channel and spatial attention as follows:

$$F'' = M_s(F') \otimes F'$$

where $\otimes$ denotes element-wise multiplication. F'' is the optimized feature vector obtained through the embedded SC-Net. Since SC-Net is embedded after the first three residual blocks of ResNet50, F'' represents the output feature after embedding network optimization. After ResNet50' s fourth layer, we obtain feature maps of $2048 \times 24 \times 8$. Average pooling extracts a 2048-dimensional feature vector, which then passes through a 1×1 convolutional layer, batch normalization, and ReLU layer to obtain a 512-dimensional feature vector. Finally, the global branch network is trained using triplet loss and softmax loss functions.

1.2 Local Branch Network

Many methods focus primarily on global features while ignoring pedestrian details, increasing Re-ID difficulty. Consequently, researchers increasingly consider local features. Our local branch differs from other methods by utilizing feature vectors extracted from the global branch as its foundation. Assuming feature vector F is obtained from ResNet50' s third layer, we use a bottleneck [9] to map F to T with dimensions $2048 \times 24 \times 8$. Simple partitioning divides T into three identical feature vectors t_1, t_2, t_3 of size $2048 \times 8 \times 8$. The proposed Gated Recurrent Unit Change Network (GRU-CN) then transforms these into three local features.

The GRU-CN architecture is an improvement upon Spatial Transformer Networks (STN) [10]. STN has proven effective at focusing on the most important image regions and can automatically locate multiple significant areas when only one foreground object exists. However, experiments reveal that STN' s localization network uses only simple convolutional operations, which cannot adequately handle partitioned local features, resulting in weak correlation between different local patches. Research shows that Gated Recurrent Units (GRU) [11] inherit LSTM characteristics, enabling stronger spatial dependencies between local features while improving computational efficiency. Therefore, we designed a new localization network by incorporating GRU followed by two fully connected layers. Our GRU-CN network obtains more important local feature information while maintaining connectivity between local features. Using GRU-CN on local features can identify important pedestrian regions, automatically align features, and produce better feature maps to improve Re-ID performance. The GRU-CN network structure is shown in [Figure 4: see original paper].

The localization network input is a feature map $U \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels and H, W represent height and width. The output is a 6-dimensional affine transformation parameter θ that can be described as:

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix}$$

Affine transformations allow scaling, rotation, and shearing of the input. The localization network predicts transformation parameters. In our network, the localization network combines GRU with two fully connected layers, so $\theta = FC(GRU(U))$, where $GRU(\cdot)$ is the gated recurrent unit and $FC(\cdot)$ represents two fully connected layers. Using the 6-dimensional parameter θ, we compute the affine transformation for the image to obtain the transformed feature map V with unchanged dimensions:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x_s \\ y_s \\ 1 \end{pmatrix}$$

where (x_t, y_t) are target coordinates in the output image, (x_s, y_s) are source coordinates in the input image, and $\theta_1, \theta_2, \theta_4, \theta_5$ are scaling and rotation parameters while θ_3, θ_6 are translation parameters. The final step requires computing the pixel value at each position (x_t^i, y_t^i) in the output feature map V:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \cdot \max(0, 1 - |x_t^i - m|) \cdot \max(0, 1 - |y_t^i - n|)$$

where V_i^c denotes the pixel value at channel c and position (x_t^i, y_t^i) in the output feature map, and U_{nm}^c denotes the pixel value at channel c and position

(x_s, y_s) in the input feature map. Finally, loss functions enable backpropagation. Through GRU-CN transformation, we obtain more robust pedestrian local information features. In the local branch, we similarly use a 1×1 convolutional layer, batch normalization, and ReLU layer to obtain 512-dimensional feature vectors, trained with triplet loss and softmax loss functions.

1.3 Feature Fusion Branch

To improve Re-ID accuracy, fusing extracted local and global features to generate more robust feature representations is essential. Research shows that simple concatenation or addition operations may introduce feature noise interference. Literature [12] demonstrates that using vector outer products makes extracted feature maps more expressive. Therefore, we employ Feature Descriptor Fusion (FDF) to combine global and local features.

To validate FDF effectiveness, we conducted comparative experiments detailed in Section 4.4. Assuming the global branch extracts pedestrian features $gF \in \mathbb{R}^{C\times H\times W}$ and the local branch extracts features $lF_i \in \mathbb{R}^{C\times H\times W}$ (where $i = 1, 2, 3$), we concatenate three local features to obtain local feature vector $lF \in \mathbb{R}^{C\times H\times W}$ with the same size as global feature gF. Let $gF(x, y)$ denote the feature descriptor at point (x, y) in global feature gF, and $lF(x, y)$ denote the descriptor at point (x, y) in local feature lF. FDF primarily uses outer products to combine global and local features, producing fused feature vector M:

$$M(x, y) = gF(x, y) \cdot lF(x, y)^T$$

where $\cdot$ denotes the outer product and T represents matrix transposition. S is the spatial size, i.e., $S = H \times W$. Finally, we normalize to obtain the fused feature vector M:

$$M = \frac{M}{\|M\|_2}$$

FDF fuses local and global features into a 2048-dimensional feature vector. Softmax loss optimizes the feature fusion stage learning parameters. In our network, dropout [13] is only used in the GRU-CN network. Finally, feature vectors from all three branches are combined to form the pedestrian image feature vector.

1.4 Loss Function

Both global and local branches share the same loss function. The total loss is the sum of improved triplet loss and classification loss:

$$L_{total} = L_{cls} + L_{triplet}$$

where L_{cls} is the classification loss:

$$L_{cls} = -\sum_{i=1}^{T} y_i \ln(f_s(x_i))$$

with y_i as the predicted label, T as pedestrian categories, and f_s as the classification function. $L_{triplet}$ is the improved triplet loss [9]:

$$L_{triplet} = \sum_{i=1}^{M}\sum_{a=1}^{N} \max(0, 1 + \max_{p=1\ldots N} D(x_a^i, x_p^i) - \min_{n=1\ldots N, n\neq i} D(x_a^i, x_n^i))$$

where M represents pedestrian categories, N denotes images per person, and D is Euclidean distance. For each sample x_a^i, x_p^i is the hardest positive sample (maximum distance within same identity) and x_n^i is the hardest negative sample (minimum distance across different identities). The loss sums over all triplets in a batch. For the feature fusion branch, the fused feature vector is fed into a softmax classification loss function as shown in Equation (13).

2 Experimental Results

We evaluate our model on three person Re-ID datasets: Market1501, CUHK03, and DukeMTMC-reID. For simplicity, all experiments use single-query mode. Our network is implemented using PyTorch, trained on an NVIDIA GeForce GTX 1080Ti GPU with an Intel i7 CPU and 32GB RAM. We adopt the Adam [14] optimizer, an extension of stochastic gradient descent. Training uses batches of 16 images, with test batches of 16 images. During preprocessing, images are resized to 384×128 and augmented via random horizontal flipping and normalization. The initial learning rate is 1×10^{-3}, decaying to 1×10^{-4} after 100 epochs and 1×10^{-5} after 300 epochs. Training runs for 400 epochs total, requiring 8-10 hours to converge.

Market1501 Dataset: Table 1 presents results compared with state-of-the-art methods including metric learning (LOMO+XQDA [15], BoW+KISSME [16]), attribute learning (APR [17]), deep learning (GLOB-TO-LOCAL [7], PCB [18], PCB-RPP [18]), attention mechanisms (MSCAN [19], HA-CNN [20]), and recent methods (DMA-CN [21], Pose [22]). Our method significantly outperforms metric learning approaches. Compared with deep learning methods requiring no prior knowledge, our Rank-1 accuracy reaches 94.4%. Versus PCB-RPP, which also uses attention, our mAP improves by 2.2%. Our baseline ResNet-50 achieves 71.59% mAP and 88.84% Rank-1. Our method surpasses MSCAN and HA-CNN by 14.1% and 3.2% in Rank-1, respectively. Combining with re-ranking (RK) further improves Rank-1 to 95.2% and mAP to 93.1%.

CUHK03 Dataset: This challenging dataset contains many occlusions. Table 2 compares our results with low-level feature methods and deep learning approaches. Our method shows significant improvement over low-level methods,

achieving 61.9% Rank-1 and 65.3% mAP on CUHK03-detected, outperforming PCB+RPP by 2.5% and 5.2% respectively. On CUHK03-labeled, accuracy reaches 65.0% Rank-1 and 67.6% mAP. Re-ranking integration yields further improvements shown in Table 2.

DukeMTMC-reID Dataset: Table 3 compares our method with IDE [1], ARP, HA-CNN, PCB+RPP, DMA-CN, and Pose. Our method achieves 72.9% mAP and 86.8% Rank-1. Compared with GP-ReID [23], our Rank-1 improves by 1.6% despite similar mAP. Our approach surpasses several classic methods, and with RK achieves 86.8% mAP and 89.7% Rank-1.

3 Experimental Analysis

We conduct ablation studies on Market1501 to validate each component' s effectiveness.

3.1 Impact of Attention Mechanism

Table 4 presents four experiments validating SC-Net. Both channel-only and spatial-only attention improve results, as attention mechanisms focus on important information while reducing background interference. However, using them separately may lose some pedestrian information, reducing recognition rate. Our SC-Net combines spatial and channel attention, preserving spatial invariance while emphasizing channel information. This integrated approach focuses on important features without excessive information loss, demonstrating that SC-Net helps extract global features and improves accuracy.

3.2 Impact of Local Feature Transformation Network

Table 5 validates GRU-CN superiority through three comparison experiments. Both LSTM-STN and STN improve performance, but our GRU-CN increases mAP by 2.12% and Rank-1 by 0.52% over STN. Compared to LSTM-STN, GRU-CN shows slight improvement because GRU has fewer parameters, trains faster, and requires less data for generalization—advantageous for Re-ID datasets with limited samples. STN can spatially transform and align data without keypoint annotations, improving classification accuracy when spatial variations are large. GRU-CN maintains STN' s alignment capability while establishing connections between partitioned local features, better preserving pedestrian holistic information.

3.3 Impact of Partitioning

Our local branch uses horizontal partitioning into three local features. Table 6 validates partitioning effectiveness. Partitioning improves baseline recognition, but different partition numbers yield varying results. Our three-partition approach performs best. Two partitions may miss details, while four partitions

over-segment images, introducing background interference. Thus, three horizontal partitions produce optimal results.

3.4 Impact of Feature Fusion Technology

Table 7 demonstrates that combining global and local branches significantly improves results over baseline. To validate FDF accuracy, we compare with three fusion methods: concat, Fisher Vector (FV) [24], and bilinear [25]. Compared to simple concatenation, FDF improves mAP by 2.94% and Rank-1 by 3.13%. FDF also slightly outperforms mainstream methods FV and bilinear. FDF preserves original feature information without degradation, proving its effectiveness in extracting more discriminative pedestrian features.

Finally, [Figure 5: see original paper] visualizes retrieval results on three Re-ID datasets. The first column shows query images, with retrieved images ranked by similarity from left to right. Blue rectangles indicate correct matches, red rectangles show errors. Our model achieves high recognition rates on Market1501 and DukeMTMC-reID. On CUHK03, some errors occur due to limited gallery images or high similarity between persons, but remain within acceptable bounds. Our enhanced feature network effectively identifies pedestrian identities.

4 Conclusion

This paper proposes an Enhanced Feature Convergent Network (EFCN) for person re-identification. The model addresses challenges including pose variation, blurry images, and similar appearances. Using ResNet50 as the backbone, EFCN comprises three branches: global, local, and feature fusion. The global branch employs SC-Net as an embedded attention network after ResNet50' s first three layers to extract representative global features. The local branch uses GRU-CN to extract important local information. The fusion branch combines global and local features via FDF to obtain robust, representative pedestrian descriptors, trained with composite loss functions. Evaluated on three Re-ID datasets, our method demonstrates superior performance compared to mainstream approaches, confirming its effectiveness in extracting discriminative features for large-scale non-overlapping multi-camera Re-ID scenarios.

References

[1] Ben Xianye, Xu Sen, Wang Jun. Review on pedestrian gait feature expression and recognition [J]. Pattern Recognition and Artificial Intelligence, 2012, 25 (1): 000071-81.

[2] Qi Meibin, Tan Shengshun, Wang Yunxia, et al. Multi-feature Subspace and Kernel Learning for Person Re-identification [J]. ACTA AUTOMATICA SINICA, 2016, 42 (2): 299-308.

[3] Zhao Haiyu, Tian Maoqing, Sun Shuyang, et al. Spindle net: person re-identification with human body region guided feature decomposition and fusion

[C]// Proc of IEEE the European Conference on Computer Vision (ECCV). 2017: 907–915.

[4] Zhao Rui, Wanli Ouyang, Wang Xiaogang. Person re-identification by salience matching [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 2528-2535.

[5] Liu Hao, Feng Jiashi, Qi Meibin, et al. End-to-End comparative attention networks for person re-identification [C]// Proc of IEEE the 26th International Conference on Neural Information Processing Systems. 2015: 3492–3506.

[6] Chen Ying, Huo Zhonghua. Person re-identification based on multi-directional saliency metric learning [J]. Journal of Image and Graphics, 2015, 20 (12): 1674-1683.

[7] Wei Longhui, Zhang Shiliang, Yao Hantao, et al. GLAD: Global-local-alignment descriptor for pedestrian retrieval [C]// Proc of the 25th ACM International Conference on Multimedia. 2017: 420–428.

[8] Guo Yiluan, Ngai-Man Cheung. Efficient and deep person re-identification using multi-level similarity [C]// Proc of IEEE on Computer Vision and Pattern Recognition (CVPR). 2018: 2335–2344.

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770–778.

[10] Jaderberg M, Simonyan K, and Zisserman A, et al. Spatial transformer networks [C]// Proc of the 28th International Conference on Neural Information Processing Systems. 2015: 2017–2025.

[11] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proc of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724–1734.

[12] Gao Yang, Oscar B, Zhang Ning, et al. Compact Bilinear Pooling [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 317–326.

[13] Srivastava N., Hinton G. Krizhevsky E, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research. 2014, 15 (1): 1929–1958.

[14] Zheng Liang, Yang Yi, Hauptmann A G. Person re-identification: past, present and future [J]. 2016, arXiv preprint: 1610. 02984. https://arxiv.org/abs/1610.02984.

[15] Liao Shengcai; Hu Yang; Zhu Xiangyu, et al. Person re-identification by local maximal occurrence representation and metric learning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 2197-2206.

[16] Zheng Liang, Shen Liyue, Lu Tian, et al. Scalable person re-identification: a benchmark [C]// Proc of IEEE International Conference on Computer Vision (ICCV). 2015: 1116-1124.

[17] Lin Yutian, Zheng Liang, Zheng Zhedong, et al. Improving person re-identification by attribute and identity learning [J]. pattern recognition. 2019: 151-161.

[18] Sun Yifan, Zheng Liang, Yang Yi, et al. Beyond part models: Person retrieval with refined part pooling [C]// Proc of the International Conference on European Conference on Computer Vision (ECCV). 2018: 481-496.

[19] Li Dangwei, Chen Xiaotang, Zhang Zhang, et al. Learning deep context-aware features over body and latent parts for person re-identification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 7398-7407.

[20] Li Wei, Zhu Xiatian, Gong Shaogang. Harmonious attention network for person re-identification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 2285-2294.

[21] Liu Chang, Qiu Weigen, Zhang Lichen. Person re-identification based on deformable mask alignment convolution model [J]. Computer Engineering and Applications, 2020 (03). http://dx.doi.org/10.3778.

[22] Pei Jiazhen, Xu Zengchun, Hu Ping. Person re-identification fusing viewpoint mechanism and pose estimation [J]. Computer Science, 2020 (02). http://dx.doi.org/10.11896/2019/500013.

[23] Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person re-Identification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3908-3916.

[24] Sanchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: theory and practice [J]. International Journal of Computer Vision. 2013, 105 (3): 222-245.

[25] Lin T, Roychowdhury A, Maji S, et al. Bilinear CNN Models for Fine-Grained Visual Recognition [C]// Proc of the international conference on computer vision (ICCV). 2015: 1449-1457.

Note: Figure translations are in progress. See original paper for figures.