Self-attention Learning for Person Re-identification

Minyue Jiang[1]
jiangmy0817@gmail.com
Yuan Yuan[1]
y.yuan1.ieee@gmail.com
Qi Wang*[12]
crabwq@gmail.com

[1] School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL)
Northwestern Polytechnical University
Xi'an, P. R. China

[2] Unmanned System Research Institute
Northwestern Polytechnical University
Xi'an, P. R. China

Abstract

Person re-identification is a critical yet challenging task in video surveillance. It aims to match the same person across cameras. Practically, people's appearances vary greatly across cameras. Most deep learning methods rely on single-level features of deep layer while ignoring low-level detailed features of shallow layers, since different layers have different sizes of feature maps and different layers' features cannot be concatenated without extra downsampling or upsampling. To remedy this problem, we propose a novel yet simple self-attention learning method for person re-identification. We design a convolutional neural network(CNN) to capture multi-level information from different layers while keeping spatial resolution of feature maps unchanged by using dilated convolution. Multi-level information consists of two parts: multi-level attention maps and multi-level feature maps. Multi-level attention maps are constrained and multi-level feature maps are concatenated easily. And we combine softmax loss with quadruplet loss, taking full advantages of labels and metric learning at the same time. Experimental results demonstrate the proposed method achieves excellent performance for person re-identification and self-attention constraint can also be used in many other tasks.

1 Introduction

Person re-identification has drawn increasing attention in both academia and industry due to its wide application prospect. Normally, in order to get a wide field of view, surveillance cameras are put in high position and have non-overlapping views. These camera settings increase the difficulty of person re-identification. People's appearances vary greatly due to variations in view angles, illumination condition, mutual occlusion, etc. The resolution of pedestrians' images is low. And people may have different poses across cameras. How to extract discriminative and robust features is still an open and fundamental problem for person re-identification when undergoing such challenging situation. In addition, the Euclidean distance metric may be insufficient to measure the similarity between pedestrians when intra-class has large variation and inter-class has high similarity across cameras. Hence we need to find a more suitable subspace to measure similarity where intra-class distance is smaller than inter-class distance.

Traditional methods solve the problem by exactly covering above the two aspects: feature extraction[15, 16, 23, 29] and metric learning[4, 9, 11, 16, 26, 31]. Color[16, 23] and texture[29] are the most commonly extracted features. These features are all hand-crafted features. But when undergoing such significant environmental changes, these pre-designed features may lack discrimination to distinguish similar people. The metric learning method can enhance discrimination by projecting the features into a subspace where intra-class distance is reduced and inter-class distance is enlarged, foaming a margin between intra-class and inter-class. There are still some drawbacks when combining the feature extraction with metric learning. Specifically, feature extraction method cannot adjust adaptively according to the requirements of metric learning. Two components are still independent. The lack of interaction between feature extraction and metric learning makes person re-identification systems separated.

Recently, deep learning person re-identification methods[1, 2, 3, 5, 12, 13, 18, 20, 21, 27, 28] have shown their strengths over traditional methods. Feature extraction[2, 5, 20, 21, 28] and metric learning[1, 3, 12, 13, 18, 27] can be combined as an entirety. Convolutional neural networks are considered as feature extractor and loss functions such as quadruplet loss[3] are considered as metric learning. Networks can adjust adaptively according to the requirements of the loss function through back-propagation. Nevertheless, most deep learning methods have two main drawbacks. On the one hand, these methods only use single-level features of deep layer. It is intrinsically hard for deep learning methods to use multi-level features from different layers. Max-pooling operation makes the sizes of feature maps from different levels inconsistent. Other than that, single-level feature may not provide enough detailed information on demand of re-identification. For example, small objects may only have activations on shallow layers such as small bags. On the other hand, single loss function is insufficient for person re-identification. Most deep neural networks are trained on softmax loss. Althouth the performance is well, classification models still have some drawbacks. The major drawback is that softmax loss lacks consideration of intrinsically large intra-class variation and high inter-class similarity across different views. The designed loss function needs to consider the intra-class and inter-class distance.

In this paper, we propose a novel yet simple convolutional neural network framework which uses multi-level features to capture detailed information by self-attention learning. Firstly, we use dilated convolution to substitute traditional convolution in deeper layers to keep spatial resolution unchanged. With the help of the unchanged spatial resolution, multi-level information can be extracted easily. Then, multi-level attention maps are extracted and multi-level feature maps are concatenated. The network learns the attention by itself, so it is called self-attention learning. The whole network is trained on softmax loss and quadruplet loss, combining strengths of two loss functions and taking the intra-class and inter-class distance into consideration. Details are explained in Section 3. In general, the main contributions of this paper are summarized as follows:

- multi-level information from different layers are extracted. Low-level detailed information and high-level semantic information learn from each other by network itself.
- Complementary advantages of softmax loss and quadruplet loss are combined. The generalization capability can be improved.
- Experimental results show that self-attention learning method achieves excellent performances on person re-identification datasets(Market-1501 and DukeMTMC-reID)

and self-attention constraint can also be applied to many other tasks.

2 Related Works

Existing person re-identification methods mainly focus on either feature extraction, or metric learning, but traditional methods[4, 9, 11, 15, 16, 16, 23, 26, 29, 31] and deep learning method[1, 2, 3, 5, 12, 13, 18, 20, 21, 27, 28] conduct these two aspects in different ways.

For traditional methods, feature extraction[15, 16, 23, 29] and metric learning[4, 9, 11, 16, 19, 22, 26, 31] are main components. Yang [23]*et al*. combine salient color names with color histograms to represent color distribution in appearance feature extraction. In [29], Zhao *et al*. extracts dense scale invariant feature transform(SIFT) features to alleviate the misalignment across different views. In terms of metric learning, the purpose is to reduce intra-class distance and enlarge inter-class distance, foaming a margin between intra-class and inter-class. In [26], Zhang *et al*. learn a null space, projecting same persons' features into one single point. Zheng *et al*. [31] proposed probabilistic relative distance to maximize the probability of positive pairs having smaller distance than negative pairs. In [16], local maximal occurrence(LOMO) feature extraction method is fixed when Cross-view Quadratic Discriminant Analysis(XQDA) metric is learned. Compared with end-to-end deep learning method, feature extraction and metric learning are independent.

Deep learning methods have surpassed traditional hand-crafted designed methods. By designing a convolutional neural network[2, 5, 20, 21, 28] and a loss function[1, 3, 12, 13, 18, 27], feature extraction and metric learning are combined into an entirety. In [12], filter pairing neural network(FPNN) is proposed to jointly handle complicated environmental changes. Camera style adaptation can also be considered in [24]. In [21], a gated siamese convolutional neural network is utilized to emphasize discriminative local patterns by comparing the mid-level features across pairs of images. In terms of loss function, triplet loss is refined into triplet-hard loss[1]. Quadruplet loss[3] is designed to lead a larger inter-class variation and smaller intra-class variation by adding another constraint comparing with triplet loss. Self-attention learning is a deep learning method. Compared with other state-of-the-art deep learning methods, our method takes full advantages of multi-level features and loss functions. Multi-level features are extracted easily by using dilated convolution. And the network is learning by itself through self-attention constraint. The final loss combines complementary strengths of softmax loss and quadruplet loss.

3 Self-attention Learning

Self-attention Learning method mainly contains two components: network architecture and loss functions respectively. Figure 1 shows the overall architecture of our network.

3.1 Network Architecture

Our network architecture is designed for extracting multi-level information which contains two parts: multi-level attention maps and multi-level feature maps.

The network is based on ResNet50[8]. More specifically, after the stage3 of ResNet50[8], convolution is replaced by dilated convolution. Another block stage6 is added. The architecture of stage6 is the same as stage5 other than the number of channels is half of stage5.

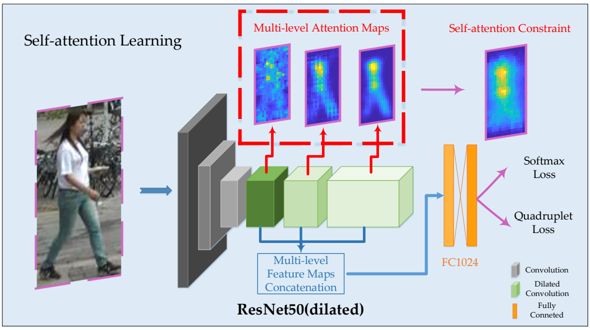

Figure 1: The whole network Architecture. Details are explained in 3.1. Different colors mean different operations in CNN. Grey, green and orange represents conventional convolution, dilated convolution and fully connected layer, respectively. The final network is trained on self-attention constraint, softmax loss and quadruplet loss. Best viewed in color.

Then the feature maps of the last convolution in stage4, stage5, stage6 are concatenated into a multi-level features. Multi-level attention maps are also extracted from these layers. The final compact feature vector is 1024-dim through global pooling and a fully connected layer.

Current CNNs are hard to use multi-level features because of the commonly used max-pooling operations. These operations make sizes of these feature maps from different layers inconsistent. Different layers cannot fuse without extra downsampling or upsampling. In order to keep the spatial resolution unchanged and extract multi-level features, we use dilated convolution. The dilated convolution is the key component in our network design. By using this operation, max-pooling operations are removed. And the spatial resolution is maintained from stage3 to stage6. As we can see from Figure 2, the dilation on the kernel can change the receptive field. In order to keep receptive field the same as ResNet50[8], the dilation is double after every original max-pooling operation. However, dilation also bring gridding artifacts. In order to alleviate gridding artifacts, we add another block stage6 at the end of stage5. The adding kernels remove artifacts with appropriate frequency.

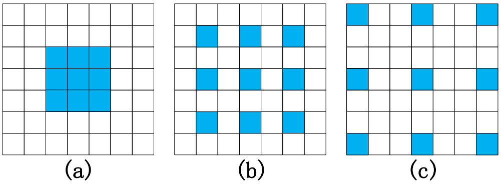

Figure 2: Dilated Convolution. Blue points represent the 3x3 convolutional kernel, dilation means the distance between points on the kernel. (a) conventional 3x3 convolutional, dilation=0. (b) dilated 3x3 convolution, dilation=1. (c) dilated 3x3 convolution, dilation=2.

With the help of dilated convolution, multi-level information is encoded easily. The information of shallow layers can be considered more local and detailed than deep layers and the information of deep layers can be considered more general and semantic than shallow layers. Multi-level information contains two parts: multi-level attention maps and multi-level feature maps. On the one hand, feature maps of different levels can be concatenated easily because of the unchanged spatial resolution. On the other hand, multi-level attention maps are extracted easily from the last convolution's output of stage4, stage5 and stage6.

3.2 Self-attention Constraint

Attention maps are considered as one-channel spatial maps that essentially try to encode the significance of spatial areas. Concretely, for each spatial point i on the spatial attention map[25], we first calculate the square-mean of all channels:

$$a_i = \frac{1}{C} \sum_{j=1}^{C} v_{ij}^2, \tag{1}$$

where C represents the channel number, v_{ij} is the activation value of spatial point i on the channel j, a_i represents unnormalized value of spatial point i on the attention map. Then we use l_2 normalization to normalize attention maps and get the final attention map att_{s4}, att_{s5}, and att_{s6} from stage4, stage5 and stage6 respectively.

These attention maps indicate the spatial significance across the feture maps. The network pays more attention to spatial points which have bigger values on these maps since they have higher activation through the network.

Different stages pay different attention to the images [25]. Shallow layers(stage4) focus on detailed texture part while deep layers(stage5, stage6) focus on more semantic part. Concatenating feature maps of different levels directly may cause an average of these discriminative regions. To remedy this problem, we introduce self-attention constraint to guide different levels in focusing on the same discriminative regions. By self-attention constraint, this self-attention learning is achieved. Low-level detailed attention and high-level semantic attention are learning from each other. Formally, the self-attention constraint is:

$$\mathcal{L}_{sac} = \|att_{s6} - att_{s5}\|_2^2 + \|att_{s6} - att_{s4}\|_2^2. \tag{2}$$

We minimize the discrepancy among these attention maps. These attention maps are constrained to be as same as possible, making different layers focus on the same discriminative regions. Besides, the gradient can back propagate directly to shallow layers, making the network converge faster.

3.3 Loss Functions

Our loss function combines softmax loss and quadruplet loss, making up for the shortcomings of both sides. Softmax loss is widely used in classification tasks. Formally, softmax loss is:

$$\mathcal{L}_{cls} = \sum_{i=1}^{n} -p_i \log(q_i), \tag{3}$$

where n is the number of persons in training set, p_i and q_i are the target probability and the predicted probability, respectively. This loss is suitable for the case that inter-class distance is much larger than intra-class distance such as classification task on ImageNet[17] dataset. But softmax loss does not consider the intra-class distance and inter-class distance, which is not suitable for person re-identification. In fact, same person's appearances vary greatly and different people may be similar across views. Person re-identification have large intra-class variation and high inter-class similarity intrinsically. This loss need the help of loss which considers intra-class and inter-class distance.

Triplet loss is a commonly used loss which considers intra-class and inter-class distance. For different images i, j, k, their extracted feature vectors from the above network are $\{x_i, x_j, x_k\}$, corresponding person identities are $\{y_i, y_j, y_k\}$, $y_i = y_j$ and $y_i \neq y_k$, Triplet loss for x_i is:

$$\underset{x_i}{\mathcal{L}_{trip}} = \max(0, d(x_i, x_j) - d(x_i, x_k) + \varepsilon_1), \tag{4}$$

where d represents l_2 distance, ε_1 represent the margin between the intra-class and inter-class. By making a margin between intra-class and inter-class, the network can distinguish similar people. In [1], triplets are chosen based on the local anchor x_i. x_j has the largest intra-class distance with x_i, and x_k has the smallest inter-class distance with x_i. However, this loss only focuses on the local distance around x_i, the network can not get a wider range of neighbour samples' similarity distribution.

To alleviate this problem, we use quadruplet loss which can also remedy defects of softmax loss. The quadruplet loss used in this paper adds another term comparing with triplet loss:

$$\begin{aligned} \underset{x_i}{\mathcal{L}_{quad}} = &\max(0, d(x_i, x_j) - d(x_i, x_k) + \varepsilon_1) + \\ &\max(0, d(x_i, x_j) - d(x_k, x_h) + \varepsilon_2), \end{aligned} \tag{5}$$

where $\{x_i, x_j, x_k\}$ are the same in triplet loss, h represents the adding image, corresponding feature vector and identity are x_h and y_h, respectively, ε_1 and ε_2 mean different margins between intra-class and inter-class. It is important to notice that $y_i \neq y_k \neq y_h$. Compared with the most used triplet loss, quadruplet loss adds a margin ε_2 to constrain another distance between $\{x_i, x_j\}$ and $\{x_k, x_h\}$. With the help of the adding margin ε_2, the network can get a wider range of neighbour samples' similarity from different probe images. Quadruplet loss can have larger inter-class distance than triplet loss. More comparisons can be found in Section 4.3.

However, quadruplet loss only use weakly supervised information. The performances of networks heavily depends on the quality of quadruplet. At the beginning of training, the network itself cannot provide enough information about similarity for choosing effect quadruplet. How to choose appropriate quadruplet $\{x_i, x_j, x_k, x_h\}$ is another significant issue. The number of quadruplets increases in fourth-power with the growth size of dataset. And most quadruplets are trivial which cannot provide useful information to optimize the network. To remedy this problem, the network is trained on softmax loss firstly. Based on the first stage's classification training, the network itself can provide enough information to mine hard quadruplet. $\{x_i, x_j, x_k\}$ is chosen as same as the triplet loss. And x_h has the smallest inter-class distance with x_k. We call it quadruplet-hard loss, mining hard quadruplet from mini-batch. Hard quadruplet can provide more information than trivial quadruplet. Noticed that inter-class $\{x_i, x_k\}$ is chosen based on the anchor x_i, while $\{x_k, x_h\}$ is chosen based on the anchor x_k. Two anchors are different.

ε_1 and ε_2 are also very important for quadruplet loss, controling how hard the network learns. ε_1 represents a more **local** and strong margin based on the anchor x_i. ε_2 represents a more **general** and weak margin based on two anchors x_i and x_k, thus ε_1 is larger than ε_2. We perform grid search on ε_1 and ε_2. ε_1 and ε_2 are fixed 0.5 and 0.2 for all experiments.

Different from [3], our quadruplet-hard loss function calculates loss for each sample in the mini-batch, ε_1 and ε_2 are fixed in the training procedure, while [3] calculates loss for the entire mini-batch and ε_1, ε_2 change adaptively. [3] needs more complex computation which may not suit for large mini-batch. And softmax loss is not combined with [3].

Combining two loss functions can alleviate each other's drawbacks. Softmax loss takes full advantages of labels. It can provide effect information for choosing quadruplets. And quadruplet-hard loss considers intra-class and inter-class distance.

The final loss is formulated as:

$$\mathcal{L}_{final} = \mathcal{L}_{sac} + \mathcal{L}_{quad} + \mathcal{L}_{cls}, \tag{6}$$

where $\mathcal{L}_{sac}$ represents self-attention constraint, $\mathcal{L}_{quad}$ represents quadruplet-hard loss, $\mathcal{L}_{cls}$ represents softmax loss. The weight of each part is 1, every part plays an equal role in the final loss since adjusting these weights in a large range does not affect the performance. More implementation details can be found in Section 4.1.

4 Experiments

4.1 Implementation Details

All training images are resized to 256 × 128. The training procedure is two-stage. Firstly, the network is just trained on softmax loss and self-attention constraint. The learning rate is 0.1 and gradually decreases by multiplying 0.1 after each 50 epoches. The batch size is 128. Then we train the network on Eq 6. The batch-size is 72 of 18 identities with 4 images per identity. The learning rate strategy is the same as the first stage except that the beginning learning rate is 0.0001. We use mini-batch stochastic gradient descent to train the network.

4.2 Datasets and Protocols

Market-1501[30]: It contains 32668 images collected from 6 camera views of 1501 identities. Deformable Part Model(DPM) [6] is used as the person detector. Training set and testing set contain 750 and 751 identities, respectively.

DukeMTMC-reID[7]: It contains 16,522 training images of 702 identities, 2,228 query images and 17,661 gallery images of the other 702 identities.

Protocols. We adopt two evaluation protocols in two datasets. Rank-1 identification rate and mean Average Precisoin(mAP). mAP takes both precision and recall rate into consideration, which is complementary to Rank-1.

4.3 Ablation Experiments

Settings	Softmax	Quadruplet	DIL	SAC	Multi	mAP	Rank-1
1	√					68.5	83.7
2		√				60.6	77.2
3	√	√				73.5	87.3
3(T)	√	Triplet				71.6	84.3
4	√		√		√	72.7	85.6
5	√	√	√		√	78.4	90.5
6	√		√	√	√	76.7	88.2
7		√	√	√	√	66.0	80.3
8	√	√	√	√	√	**81.6**	**92.8**

Table 1: Ablation experiments on Market-1501 dataset using ResNet50[8]. DIL and SAC means dilated convolution and self-attention constraint, respectively. Multi means whether the feature is multi-level or not.

Dilated Convolution and multi-level features. We first study the role of dilated convolution and multi-level features. Dilated convolution not only maintains spatial resolution to obtain more details, but also helps the network extract multi-level features. As we can see from Table 1 setting 1(3) and 4(5), changing the way of convolution can improve mAP by a large margin.

Self-Attention Constraint. Different layers have different attention information. Figure 3 visualizes the attention maps extracted from different stages(stage4, stage5, stage6), (b-d) indicate the network trained without self-attention constraint, while (e-g) indicates the network trained with self-attention constraint. We can see from Figure 3 that by using self-attention constraint, deep layer and shallow layers can focus on same discriminative region. small representative objects are still activated on deep layer which is important for person re-identification. Without self-attention constraint, it is hard for shallow and deep layers to focus on same areas. Table 1 shows that this self-attention learning can improve performance 3.2% in mAP(setting5,8).

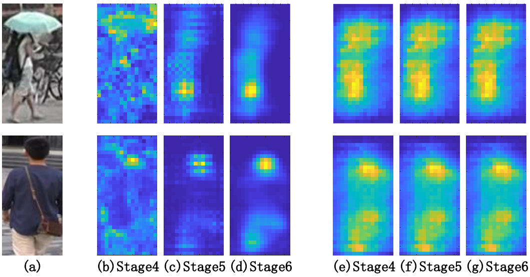

Figure 3: Self-Attention Constraint(SAC).(a) original images.(b-d) are attention maps from stage4, 5, 6 without SAC. (e-g) are attention maps with SAC from same stages. Best viewed in color.

Loss functions. Compared setting1(6) with setting2(7) in Table 1, we can see that the performance of quadruplet loss is inferior than softmax loss which means that without effective information to choose quadruplets, the network is struggled in trivial quadruplets. Comparing with triplet loss(setting3 and 3(T)), quadruplet loss has better performance (+1.9% in mAP and +3.0% in Rank-1). Better performance indicates that quadruplet loss can enlarge inter-class further than triplet loss. And with the help of the adding term, the network can achieve better generalization ability on the testing set.

Combining two loss functions(setting3,8), the performance is much higher than the network trained on softmax loss or quadruplet loss separately. It means that the network can project the feature vectors into a more discriminative space to calculate the similarities. The two loss functions are complementary to each other.

4.4 Comparison with State-of-the-art Methods

The results on **Market-1501** are shown in Table 2. All experiments are conducted in single query setting. Our method outperforms all the other methods by a large margin. More precisely, our method achieves 81.6% mAP and 92.8% Rank-1 performance. Even without re-ranking[32] algorithm, self-attention learning method is superior to most state-of-the-art methods. The performance is even higher than the second best model with re-ranking[32]. Furthermore, with the help of re-ranking[32], our method can get large improvements on

Method	mAP	Rank-1
re-rank[32]	63.6	77.1
MLC[13]	64.4	83.9
DML[27]	68.8	87.7
TRIP[11]	69.1	84.9
DPFL[5]	72.6	88.6
MLFN[2]	74.3	90.0
HAN[14]	75.7	**91.2**
Ours	**81.6**	**92.8**
MLC[13]+re-rank[32]	72.9	88.8
TRIP[11]+re-rank[32]	**81.1**	86.7
Ours+re-rank[32]	**91.2**	**93.3**

Table 2: Comparison with state-of-the-art methods on the Market-1501 dataset(Single query). $1^{st}/2^{nd}$ best in red/blue.

mAP and Rank-1 to 91.2% and 93.3% respectively, improving 10.1% in mAP and 2.1% in Rank-1.

The results on **dukeMTMC-ReID**[7] are shown in Table 3. The dataset is more challenging than Market-1501[30] dataset due to inconsistent sizes of images. Similarly, self-attention learning method achieves the best result, higher than the second best model 1.6% in mAP and 1.1% in Rank-1. The performance can improve 7.5% and 3.0% in mAP and Rank-1 by using re-ranking[32]. This indicates that by fusing multi-level features, our network can better handle various scales of pedestrians' images. From Table 2 and 3, we can see that our method achieves excellent performance on large person re-identification datasets.

Method	mAP	Rank-1
SVDNET[20]	56.8	76.7
DPFL[5]	60.6	79.2
MLFN[2]	62.8	**81.0**
HAN[14]	**63.8**	80.5
Ours	**65.4**	**82.1**
Ours+re-rank[32]	**72.9**	**85.1**

Table 3: Comparison with state-of-the-art methods on the DukeMTMC-ReID dataset. $1^{st}/2^{nd}$ best in red/blue.

4.5 Task Extension

CIFAR-100[10]: This dataset is used to demonstrate self-attention constraint can also be applied to many other tasks such as classification.

From Table 4, we can see that by self-attention constraint, the error rate decrease 0.27%. The performance is higher than ResNet101 which is much deeper than ResNet50. The improved performance demonstrates that self-attention learning can be applied to many other tasks.

Method	Error Rates(%)
ResNet50[8](Soft)	22.02
ResNet50(Soft+*DIL*+*SAC*)	**21.75**
ResNet101[8](Soft)	**21.95**

Table 4: Ablation experiments on CIFAR-100 dataset. Soft means softmax loss. The meanings of *DIL* and *SAC* are the same as those in Table 1.

5 Conclusion

We propose self-attention learning method for person re-identification which constrains different layers' attention maps focusing on the same discriminative regions. And quadruplet-hard loss combining with softmax loss can not only reduce intra-class distance and enlarge inter-class distance, but also fully use label information. Experimental results show that self-attention learning method achieves excellent performance. Ablation studys demonstrate self-attention constraint and combination of loss functions play crucial roles on enhancing performance. The concept of self-attention learning can also be applied to many other tasks.

6 Acknowledgement

This work was supported by the National Key R&D Program of China under Grant 2017YFB-1002202, National Natural Science Foundation of China under Grant 61773316, Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

References

[1] H. Alexander, L. Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[2] X. Chang, T. M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proc. CVPR*, 2018.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. CVPR*, 2017.

[4] Y. Chen, W. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *Proc. IJCAI*, pages 3402–3408, 2015.

[5] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Proc. ICCV*, pages 2590–2600, 2017.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Visual object detection with deformable part models. *Communications of the ACM*, 56(9):97–105, 2013.

[7] h. Zhedong, Z. Liang, and Y. Yi. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. ICCV*, pages 3774–3782, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[9] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, pages 2288–2295, 2012.

[10] A. Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

[11] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. CVPR*, pages 384–393, 2017.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. CVPR*, pages 152–159, 2014.

[13] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[14] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proc. CVPR*, 2018.

[15] X. Li, W. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *Proc. ICCV*, pages 3765–3773, 2015.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. CVPR*, pages 2197–2206, 2015.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein et al. Imagenet large scale visual recognition challenge. *arXiv:1409.0575*, 2014.

[18] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification a study against large variations. In *Proc. ECCV*, pages 732–748, 2016.

[19] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embeddingn. In *Proc. CVPR*, pages 4004–4012, 2016.

[20] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proc. CVPR*, pages 2590–2600, 2017.

[21] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proc. ECCV*, pages 791–808, 2016.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. ECCV*, pages 499–515, 2016.

[23] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *Proc. ECCV*, pages 536–551, 2014.

[24] Z. Zheng S. Li Y. Yang Z. Zhong, L. Zheng. Camera style adaptation for person re-identification. *arXiv:1711.10295*, 2017.

[25] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[26] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proc. CVPR*, pages 1239–1248, 2016.

[27] Y. Zhang, T. Xiang, T. Hospedales, and H. Lu. Dual mutual learning. In *Proc. CVPR*, 2018.

[28] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proc. CVPR*, pages 1077–1085, 2017.

[29] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proc. ICCV*, pages 2528–2535, 2013.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, pages 1116–1124, 2015.

[31] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. CVPR*, pages 649–656, 2011.

[32] Z. Zhong, L. Zheng, D. Cao, and S. Z. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. CVPR*, pages 3652–3661, 2017.