Discriminant Feature Learning with Self-Attention for Image-based Person Re-Identification

Yang Li[1], Xiaoyan Jiang[1,⋆⋆⋆], and Jenq-Neng Hwang[2]

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, No.333 of Longteng Road, Shanghai, China
[2] Department of Electrical and Computer Engineering, University of Washington, Box 352500, Seattle WA 98195, USA

Abstract. Image-based person re-identification (re-ID) across cameras is a crucial task, especially when cameras' fields of views are non-overlapping. Feature extraction is challenging due to changing illumination conditions, complex background clutters, various camera viewing angles, and occlusions in this case. Moreover, the space mis-alignment of human corresponding regions caused by detectors is a big issue for feature matching across views. In this paper, we propose a strategy of merging attention models with resnet-50 network for robust feature learning. The efficient self-attention model is used directly on the feature map to solve the space mis-alignment and local features dependency problems. Furthermore, the loss function which jointly considers the cross-entropy loss and the triplet loss in training enables the network to capture both invariant features within the same individual and the distinctive features between different people. Extensive experiments show that our proposed mechanism outperforms the state-of-the-art approaches on large-scale datasets Market-1501 and DukeMTMC-reID.

Keywords: person re-identification · feature extraction · self-attention · cross-entropy loss · triplet loss.

1 Introduction

Person re-ID aims to retrieve specific pedestrians in images or video sequences obtained by multiple non-overlapping cameras. Given a query person's image, the task is to find the correct corresponding matches among a set of candidate images captured by different cameras in the gallery. Two crucial issues should be addressed: robust feature extraction and suitable distance metrics. It is a challenging topic because of the following difficulties:

Firstly, due to the changing of cameras' viewpoints, and the variability of the pedestrian's posture, the detected body parts of the same person in two images

[⋆] Corresponding Author: xiaoyan.jiang@sues.edu.cn.

[⋆⋆] Yang Li and Xiaoyan Jiang contribute equally to this work.

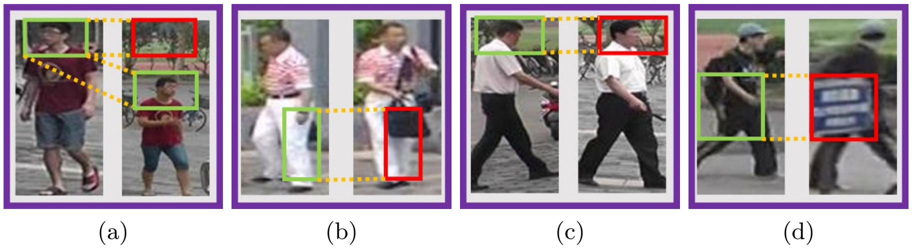

Fig. 1. Challenges of person re-ID: (a) misalignment due to inaccurate detection, (b) misalignment due to pose changes, (c) appearance in consistency, (d) occlusion.

might be mis-aligned. Sample detection bounding boxes of the same person are shown in Fig. 1(a). As shown in Fig. 1(b), the right leg region of the left image (green box) is spatially mis-aligned with the left leg region (red box) in the right image. Moreover, the black bag cannot be observed in the left image which causes changes in appearance characteristics. Secondly, different people may wear similarly or have similar gestures. As shown in Fig. 1(c), the general appearances of the two persons are close. In this case, feature extraction focusing on salient differences and fine-grained details between different categories is helpful for further feature matching. Lastly, some body parts of the person may be occluded by scene objects or other people in the image, which make the identifying more difficult. As shown in Fig. 1(d), the middle region of the pedestrian in the right image is occluded by a blue sign. Thus, the feature importance of this region should be weakened in the feature matching process.

To overcome the challenges discussed above, there have been numerous research efforts on person re-ID, which can be broadly categorized into the following techniques: representation learning, metric learning, local feature-based, generative adversarial networks, and unsupervised learning-based. More specifically, representation learning is a supervised learning relying on feature extraction methods [7]. Unlike representation learning method, metric learning aims to learn the similarity between two images through the network. Commonly used loss methods for metric learning are contrastive loss [15], triplet hard loss with batch hard mining [4], and margin sample mining loss [20]. The acquisition of person re-ID dataset is another challenge. Therefore, after the emergence of the generative adversarial networks (GAN) [28, 30, 19, 9], the model is used to expand our datasets as much as possible to improve the learning ability of the network. In contrast, the method of supervised learning relies on the demand for a large number of labeled data. Hence, unsupervised learning methods [8, 18] depending on unlabeled datasets are proposed to cross-view identity-specific information.

We propose a self-attention model extraction framework to obtain discriminant feature specifically for person re-ID. Distinct convolution kernels are combined with non-linear functions to get multi-scale spatial features, which express

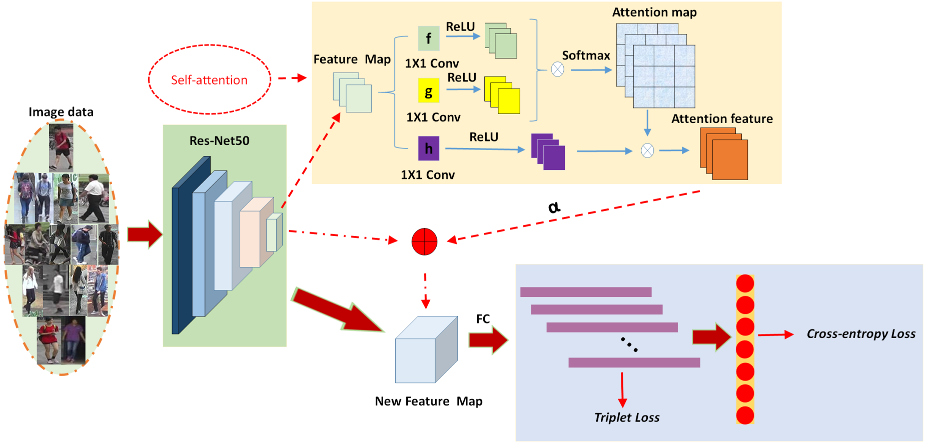

Fig. 2. The proposed network architecture. It consists of three parts: the resnet-50 network, the self-attention model, and the loss function. The self-attention model includes multi-scale transformation spaces and nonlinear combination of feature maps. The loss function considers both intra-person similarity and inter-person distinction.

images more accurately. Moreover, different loss items are fused in the loss function to generate inter-person similarity and enlarge intra-person distinction. Our approach achieves the state-of-the-art performance on standard datasets evaluated by standard metrics. The main contributions are summarized as following two-folds:

- A multi-scale attention model is jointly with resnet-50 to solve the spatial mis-alignment and local feature dependency problems of person re-ID. The nonlinear combination of features from multiple scales in the self-attention model merges the global and local features of the image effectively.
- The cross-entropy loss function usually used in multi-classification task is combined with the triplet loss function for person re-ID in the supervised training process of the network. The trained model extracts the similar characteristics of the same individuals better and significantly enlarges feature distinctions between different persons.

2 Proposed Technical Approaches

We propose a deep learning framework to deal with image-based person re-ID as shown in Fig. 2, where images are sequentially fed to the pre-trained resnet-50 network (Sec.2.1) for feature extraction to obtain corresponding feature maps. The feature map is then input to the self-attention network (Sec.2.2) to generate new discriminative spatial visual features. To better represent intra-person features and distinguish inter-person differences, the loss function fuses cross-entropy and triplet loss (Sec.2.3) to give reasonable feedback to the network finally. The details of the proposed framework are described as the following.

2.1 Baseline network resnet-50

We use the resnet-50 convolutional neural network(CNN) [3] as our basic network to extract spatial visual features based on sample 2D images. According to the conventional fine-tuning strategy [26], we use the pre-trained model on ImageNet [11], of which we only reserve the convolution layer by removing the fully connected layer. The simplest residual network block layer is used to reduce the amount of calculation and highlight the effect of the attention model.

The resnet-50 network contains an independent convolution layer (conv1) at the forefront, which is followed by four residual blocks (from res2c to res5c). BN refers to batch normalization [5] and the ReLU [17] is used as the activation function. Conv1 and res2c to res5c are used for feature extraction. Consequently, each image I_n is represented by a feature map $\{f_{n,l}\}_{l=1,\cdots,L}$ composed by 8×4 grids, where $L = 32$ is the number of batch size. Each feature map is a $D = 2048$ dimensional vector.

2.2 Self-attention model

The generalized attention mechanism is mostly used in the process of sequence decoder and encoder. As shown in [16], scaled dot-product attention computes attention functions on a set of queries simultaneously and pack them together into the matrix Q. The Keys and Values are also packed together into matrices K and V, respectively. The attention function is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt[2]{d_k}})V, \tag{1}$$

where $Q \in \mathbb{R}^{n \times d_k}$,$K \in \mathbb{R}^{m \times d_k}$,$V \in \mathbb{R}^{m \times d_v}$, and d_k is a scale factor.

Refer to the above attention mechanism, a novel self-attention model is proposed as shown in Fig. 2. We compute the visual space feature of the image $\{f_{n,l}\}_{l=1,\cdots,L}$ by resnet-50, which is simplified to $\mathbf{x} \in \mathbb{R}^{D \times L}$ while keeping the size as the original feature map size (h, w). Then we sent the feature $\mathbf{x}$ into two different feature transformation spaces $\mathbf{f}$ and $\mathbf{g}$ to calculate the attention score, and another different transformation space $\mathbf{h}$ to generate a new feature map. Thus, we have three different feature transformation spaces $\mathbf{k} \in (\mathbf{f}, \mathbf{g}, \mathbf{h})$. We indicate $W_f \in \mathbb{R}^{\bar{D} \times D}$, $W_g \in \mathbb{R}^{\bar{D} \times D}$, $W_h \in \mathbb{R}^{\bar{D} \times D}$ to be parameter matrices trained by the network. Dot-product with attention weight is defined by the following formula:

$$f(x) = W_f x, g(x) = W_g x, h(x) = W_h x. \tag{2}$$

Here, we use 1×1 convolution to perform linear combination of feature maps in different scales and different spaces. Since the size of the original feature map is large, the dot-product requires a large memory size. We convert them to a lower-dimensional feature space $\bar{D} = D/8$ by filtering the original feature map using a small number of kernels.

After linear space conversion, the ReLU activation function is applied to perform nonlinear processing to obtain the response $e_{i,j}$, which is corresponding to the original feature map. The characteristic response transformation formula is:

$$e_{i,j} = (max(f(x_i), 0))^T (max(g(x_i), 0)). \tag{3}$$

Based on the obtained response $e_{i,j}$, our attention score $s_{j,i}$ is calculated as follows:

$$s_{j,i} = \frac{exp(e_{i,j})}{\sum_{i=1}^{D} exp(e_{i,j})}, \tag{4}$$

where $s_{j,i}$ indicates the extent to which the model attends to the i^{th} location on the feature map when synthesizing the j^{th} dimension. By definition, each receptive field is a probability mass function since $\sum_{i=1}^{L} \sum_{j=1}^{D} s_{i,j} = 1$.

We denote the output of the attention model layer as $O = (o_1, o_2, \cdots, o_j, \cdots, o_N) \in \mathbb{R}^{D \times L}$, where o_j is defined as

$$o_j = \sum_{i=1}^{L} \sum_{j=1}^{D} s_{i,j} h(x). \tag{5}$$

In addition, we further multiply the output of the attention layer by a scale parameter α, of which the result is added back to the input feature map. Therefore, the final output y_i is given by $y = \alpha o_i + x_i$, where α is initialized as 0. Such parameter setting makes the network to give priority to the local neighborhood and then gradually learn to assign high weights to non-local evidence.

2.3 Loss function

We combine the triplet loss function with hard mining [4] and the softmax cross-entropy loss function with label smoothing regularization in the training process of the CNN network. The triplet loss forces the network to learn robust features representing key traits of the same person in different views, while keeping the differences between different pedestrians.

The triplet loss function is originally proposed in [4], which is named as Batch Hard triplet loss function. For each batch, we randomly select the number of P different individuals in the dataset. For each person, K tracks are randomly selected. A single track is composed by images captured from different viewing angles or extracted from different sequences. Since the proposed person re-ID approach needs only images not videos, each track here can contains only one image. We set $T = 1$ and each batch contains $P \times K$ images in total.

For each sample a in the batch, the hardest positive and the hardest negative samples within the batch are selected to form the triplet for computing the loss

$L_{triplet}$:

$$L_{triplet} = \overbrace{\sum_{i=1}^{p}\sum_{a=1}^{k}}^{\text{all anchors}} [m + \overbrace{\underset{p=1\cdots K}{max} D(f_a^i, f_p^i)}^{\text{hardest positive}} - \underbrace{\underset{\substack{j=1\cdots P\\ n=1\cdots K\\ j\neq i}}{min} D(f_a^i, f_n^i)]_+}_{\text{hardest negative}}. \tag{6}$$

The original softmax cross-entropy loss function is given by:

$$L_{softmax} = -\frac{1}{P \times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} log q_{i,a}, \tag{7}$$

where $p_{i,a}$ is the ground truth identity and $q_{i,a}$ is the prediction of the sample $\{i, a\}$. The label-smoothing regularization is used to regularize the model resulting in

$$L'_{softmax} = -\frac{1}{P \times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} log((1-\varepsilon)q_{i,a} + \frac{\varepsilon}{N}), \tag{8}$$

which is a mixture of the original ground-truth distribution $q_{i,a}$ and the uniform distribution $u(x) = \frac{1}{N}$ with weights 1-ε and ε, respectively. We set $\varepsilon = 0.1$ and N is the number of classes.

The total loss L is the combination of the above mentioned two loss items:

$$L = L_{triplet} + L'_{softmax}. \tag{9}$$

3 Experiment

Our approach is evaluated on two public standard datasets: Market-1501 [25] and Duke MTMC-reID [10], which are collected by multiple cameras and have relatively large-scale sample images. The experimental evaluation demonstrates that our proposed framework obtains better performance than other state-of-the-art methods.

3.1 Dataset

Market-1501. The Market-1501 dataset was recorded by six different cameras. It consists of 32,668 images from 1501 people, which is modeled and deformed using the Deformable Part Model (DPM). The dataset is divided into three parts: 751 people with totally 12,936 images as the training set, 750 people with totally 19,732 images as the gallery, and 3,368 images selected from the same 750 people as the query. All images are with the size of 128 × 64.

DukeMTMC-reID. The DukeMTMC-reID dataset is a subset of the newly released multi-target multi-camera pedestrian tracking dataset. The dataset is divided into three parts: 702 people with totally 16,522 images as the training set, 17,661 images of 1110 different people were used as the gallery. In addition, a total number of 2,228 images of 702 people from the initial selection of the gallery were selected as the query. The size of the image is 128×64, which is the same as that of the Market-1501 dataset.

3.2 Evaluation criteria and parameter configuration

Experimental results are evaluated by the accuracy of rank-1, 5, 10, 20 and the mean average precision (mAP) metrics. It is a remarkable fact that all experiments use a single query.

The resnet-50 network pre-trained on ImageNet is used to initialize the convolutional layer of the proposed network. The size of all images is set to 256×128. The model is trained for 800 epochs. The starting learning rate is 0.0003 and is reduced by the optimization algorithm Adam with a decay rate of 0.0005 for each 100 epochs. The batch size is 32. In the process of testing, the last level of the feature map using the attention model is expanded into a 2048-dimensional vector. Similarly, features of images in the galley are extracted by the same network. The similarity of two features is calculated by the Euclidean distance and is normalized to [0, 1].

3.3 Ablation studies

The experimental results of three different methods are shown in Table 1. The first method is resnet-50-based without using the attention mechanism.The second uses spatial-attention and the third one uses our self-attention. The spatial-attention method directly performs a one-dimensional time-series convolution on each feature map to obtain the corresponding weight value, and each function value is transferred into a vector, which is substituted into the softmax function to obtain the weight value of each feature map. Finally, we fuse multiple feature maps with the last feature map for training based on the weight value. According to the experimental results, the attention mechanism plays a certain role in the two datasets. The Rank-1 accuracy is improved by 3%-5%.

Comparison with the state-of-the-art. As shown in Table 2 and Table 3, we compare our results with the state-of-the-art methods on the dataset Market-1501 and DukeMTMC-reID, respectively. In the comparison, the metric learning

Table 1. Ablation study using the single query.

Database	resnet-50	Spatial-attention	Ours	Evaluation				
				Rank-1	Rank-5	Rank-10	Rank-20	mAP
Market-1501	√	×	×	86.8%	94.0%	96.8%	98.0%	75.6%
	√	√	×	88.6%	93.9%	97.1%	98.0%	76.2%
	√	×	√	**90.2%**	96.7%	98.1%	99.0%	**82.7%**
DukeMTMC-reID	√	×	×	76.8%	88.0%	90.2%	91.6%	74.2%
	√	√	×	78.2%	89.6%	92.4%	92.3%	75.8%
	√	×	√	**81.0%**	92.4%	94.2%	95.9%	**78.0%**

based methods include Gated-Sia [15], Basel.+LSRO [28], DML [22], JLML [6], Basel.+OIM [21], and Verif.-Identif.+LSRO [28]. Deep learning based methods include Inception-V3 [29], PDC [13], and Deep Transfer [2].

Table 2. Comparison of state-of-the-art methods on the Market-1501 dataset.

Database	Market-1501	
Method	Rank-1	mAP
Gated-Sia[15]	65.88	39.55
Spindle[23]	76.90	-
Basel.+LSRO[28]	78.06	56.23
PIE[24]	79.33	55.95
Verif.-Identif.[27]	79.51	59.87
DLPAR[25]	81.00	63.40
Deep Transfer[2]	83.70	65.50
Verif.-Identif.+LSRO[28]	83.97	66.07
PDC[13]	84.14	63.41
DML[22]	87.70	68.80
JLML[6]	85.10	65.50
PN-GAN[9]	89.43	72.58
Ours(Self-attention)	**90.20%**	**82.70%**

Table 3. Comparison of state-of-the-art methods on the DukeMTMC-reID dataset.

Database	DukeMTMC-reID	
Method	Rank-1	mAP
Basel.+LSRO[28]	67.70	47.10
Basel.+OIM[21]	68.10	-
AttIDNet[7]	70.69	51.88
ACRN[12]	72.60	52.00
SVDNet[14]	76.70	56.80
Chen et. al[1]	79.20	60.60
Inception-V3[29]	80.48	63.27
Ours(Self-attention)	**81.00%**	**78.00%**

Table 4. Evaluation on the dataset Market-1501 using different loss function.

Database	Loss		Evaluation				
	Cross-Entropy	Triplet	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Market-1501	√		85.3%	94.8%	97.0%	98.3%	76.4%
		√	86.2%	95.1%	97.3%	98.8%	76.8%
	√	√	**90.2%**	96.7%	98.1%	99.0%	**82.7%**

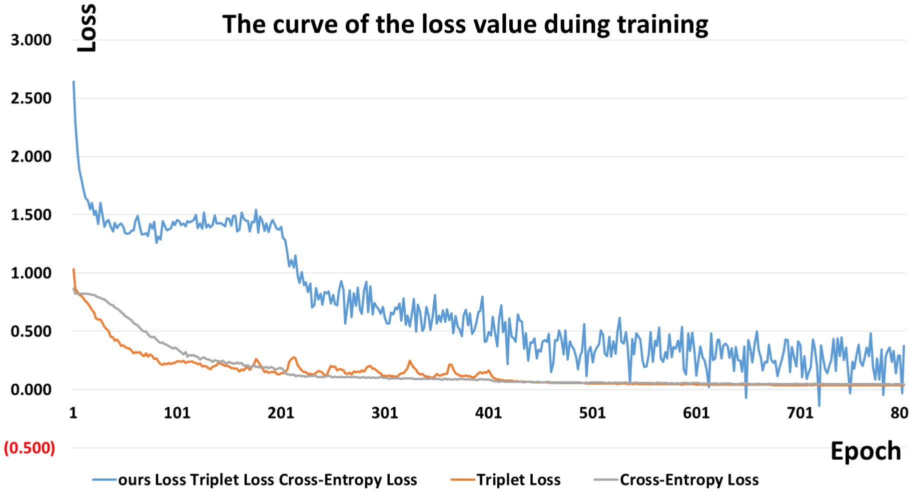

Fig. 3. Changes of loss values during training on the dataset Market-1501.

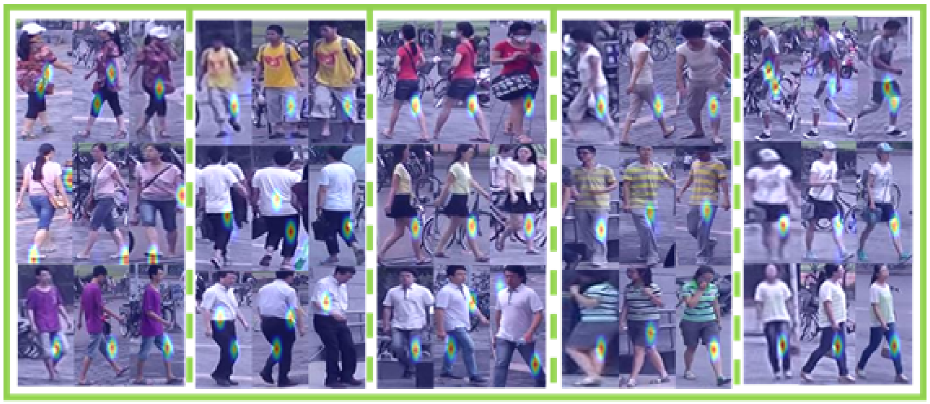

Fig. 4. Visualization of the self-attention model on the Market-1501. Our approach highlights distinctive image regions which are useful for person re-identification. The attention models primarily focus on foreground regions and generally correspond to specific body parts.

3.4 Analysis and visualization

We perform the relevant comparison on the dataset Market-1501, and visualize the results for analysis. In the following section, parameter configuration of the work influences the experimental results.

In the training process, our loss function is composed by the cross entropy and the triplet loss. In order to access the effect of each loss, we did experiments by using only one of the two loss functions for person re-ID as shown in Table 4.

As shown in Fig. 3, the curve of the loss value during training is smoothed using the logarithmic function. The cross entropy loss function mainly focuses on the common features within the class, without paying attention to the differences with other classes, which is beneficial for person re-ID task to learn the characteristics of the same person in different conditions. In contrast, the triplet loss function considers both intra-class features and inter-class properties, helping the network obtaining more invariant features in the training process. Two loss functions are fused in our approach so that they are mutually constrained and interactive with each other. Moreover, the network convergences in a better performance, although in a slower influence of the speed.

To analyze the attention model in the network, Fig. 4 shows the resulting effect of the attention model on different pedestrians. The low-level network extracts simple appearance features such as pedestrian contour, texture, color, and etc. The features extracted by high-level networks contain rich semantic information to make the final feature map more representative. Fig. 4 shows the effect of the partial kernel matrix on the corresponding original image which is different among various pedestrians. The reason is that the kernel matrix of the attention mechanism is obtained by multiplying the actual position of the corresponding pixel on the feature map. The final highlight is more concentrated on certain parts.

4 Conclusion

In this paper, an end-to-end framework for solving person re-ID is proposed. By fusing different spatial feature positions and combining two kinds of loss functions for network training, we obtain more discriminative and representative features. We utilize this attention mechanism to pay more attention to spatial position relationships between local features and global features, and use different nonlinear functions to carry out effective feature combination. In the mean time, to obtain the similarity and difference characteristics of pedestrians, we combine two high-efficiency loss functions to supervise training the network to improve the performance. Experiments show that our approach achieves better results than the state-of-the-art methods.

5 Acknowledgments

The work is supported by the following projects: National Natural Science Foundation of China, Nr.: 61702322, 6177051715, 61831018.

References

1. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2590–2600 (Oct 2017)

2. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person reidentification. IEEE Transactions on Image Processing **25**(12), 5576 – 5588 (Dec 2016)
3. He, K., Zhang, X., Ren, S., , Sun, J.: Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (Dec 2015)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv:1703.07737 pp. 1–10 (Mar 2017)
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning pp. 448–456 (Mar 2015)
6. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multiloss classification. arXiv:1705.04724 pp. 1–10 (Aug 2017)
7. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 1–10 (Mar 2017)
8. Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K., Zhong, Y.: Person reidentification by unsupervised video matching. Pattern Recognition **65**, 197–210 (May 2017)
9. Qian, X., Fu, Y., Wang, W., Xiang, T., Qiu, J., Wu, Y., Jiang, Y., Xue, X.: Pose-normalized image generation for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 1–10 (Apr 2018)
10. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. European Conference on Computer Vision(ECCV) pp. 17–35 (Nov 2016)
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., M. Bernstein, e.a.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision(IJCV) pp. 211–252 (Apr 2015)
12. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In Computer Vision and Pattern Recongnition Workshops(CVPRW) pp. 1435–1443 (May 2017)
13. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. IEEE International Conference on Computer Vision (ICCV) pp. 3980 – 3989 (Oct 2017)
14. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. IEEE International Conference on Computer Vision (ICCV) pp. 3820 – 3828 (Oct 2017)
15. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. European Conference on Computer Vision(ECCV) **9912**, 791–808 (Sep 2016)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv:1706.03762 pp. 1–10 (Dec 2017)
17. V.Nair, Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning(ICML) **27**, 807–814 (Jun 2010)
18. Wang, H., Zhu, X., Xiang, T., Gong, S.: Towards unsupervised open-set person re-identification. IEEE International Conference on Image Processing (ICIP) pp. 769 – 773 (Sep 2016)
19. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–10 (Jun 2018)

20. Xiao, Q., Luo, H., Zhang, C.: Margin sample mining loss: A deep learning based method for person re-identification. arXiv:1710.00478 pp. 1–10 (Oct 2017)
21. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3376–3385 (Jul 2017)
22. Zhang, Y., Xiang, T., Hospedales, T.M., , Lu, H.: Deep mutual learning. arXiv:1706.00384 pp. 1–10 (Jun 2017)
23. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 907 – 915 (Jul 2017)
24. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person reidentification. arXiv:1701.07732 pp. 1–10 (Jan 2017)
25. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. IEEE International Conference on Computer Vision(ICCV) pp. 1116–1124 (Dec 2015)
26. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv:1610.02984 pp. 1–10 (Oct 2016)
27. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person re-identification. arXiv:1611.05666 **14**, 1–10 (Nov 2016)
28. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. IEEE International Conference on Computer Vision (ICCV) pp. 3774 – 3782 (Oct 2017)
29. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 3652–3661 (Jul 2017)
30. Zhong, Z., Zheng, L., Zheng, Z.: Camera style adaptation for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 5157–5166 (Apr 2018)