Multi-layer attention for person re-identification

Yuele Zhang, *Jie* Guo, *Zheng* Huang, *Weidong* Qiu*, and *Hexiaohui* Fan
School of Cyber Security, Shanghai Jiao Tong University, Shanghai, 200240, China

Abstract. Person re-identification has been a significant application in the field of video surveillance analysis, yet it remains a challenging work to recognize the person of interest across disjoint cameras of different viewpoints. The factors affecting the identification results include the variation in background, different illumination conditions and the changes of human body poses. Existing person re-identification methods mainly focus on the feature extraction of the whole frame and metric learning functions. However, most of those algorithms treat different areas without distinction. It is worth emphasizing that different local regions make different contributions to image representaion, which exactly conforms to the attention mechanism. In this paper, we introduce a novel attention network which explores spatial attention in a convolutional neural network. Our algorithm learns the visual attention in multi-layer feature maps. The proposed model not only pays attention to the spatial probabilities of local regions, but also takes the features in different levels into consideration. We evaluate this multi-layer spatial attention model on three benchmark person re-identification datasets: Market-1501, CUHK03, and DukeMTMC-reID. The experiment results validate the advances of our adopted network by comparing with state-of-the-art baselines.

1 Introduction

Recently, person re-identification (Re-ID) task, as an indispensable part of video behaviour analysis field, has received widespread attention. The task of person re-identification aims at recognizing the same people from multiple different cameras with non-overlapping views. Person re-identification problem has broad potential application prospects in many occasions, especially the security systems. It remains a challenging job since the same person changes a lot under different shooting conditions. As can be seen from the image sets in figure 1, the variation in pose, illumination and background has a great impact on the recognition results.

When dealing with person re-identification task, given an image captured by Camera A (probe image), it is compared with all images which come from Camera B (gallery images). The results are ranked according to the degree of similarity between the probe image and gallery ones. In order to achieve good performance, two steps are essentially important: (i) extract features that better describe the images; (ii) find a proper similarity measurement. Many attempts have been made to improve the above two steps. Some algorithms [1, 2, 3, 4] focus on the feature extraction schemes, including the representation of images in color and

* Corresponding author: wendypenny@sjtu.edu.cn

texture, and the fusion of different levels of features. Regarding to the similarity learning, several metric learning methods [5, 2, 6] have been proposed to learn a feature space in which the calculated distance of the feature vectors belong to the same person are smaller than those belong to different pedestrians.

Fig. 1. Sample image pairs from the Market-1501 Re-ID dataset. The images in (a) belong to the same pedestrian from different cameras. The images in (b) are different persons that look similar to the person in (a). It is obvious that the variation in human pose, the change of viewpoints and background, and the low photo quality make it hard to judge whether two similar images point to the same person.

With the success of deep learning models, many methods [7, 8, 9, 10, 11, 12, 13, 14] rely on deep neural networks to tackle the feature extraction and metric learning steps. The deep network architectures achieve more representative feature expressions than hand-crafted features. The most commonly employed convolutional neural network (CNN) models can be categorized into two types: (i) classification networks [8, 12, 13], which are originally used in the image classification task and object detection task; (ii) siamese networks [7, 10, 11, 14], which take a pair of images or triplet images as the network inputs. However, few algorithms take the human attention mechanism into consideration. With the assumption that when processing images, human visual system tends to selectively focus on the important regions instead of treating all regions equally, attention mechanism dynamically allows different levels of features to gain different weights. The adoption of attention model has been proved effective in various tasks such as image captioning [15, 16, 17], machine translation [18, 19], and question answering [20, 21].

In this paper, we adopt a novel algorithm which explores visual attention models in a CNN network. Our contributions are: (1) We propose a spatial attention-based convolutional neural network for person re-identification task. Existing models generally ignore the selection of attentive regions. The formulated attention of spatial regions conforms to the natural visual attention mechanism, resulting in a better expression of the whole image. (2) We explore the attention mechanism in multi-layer feature maps. The multiple layers of weighted attention features are fused. Since CNN networks are multi-layer, the proposed attention network takes full advantage of the CNN architecture. (3) The attention-based deep model completes the person re-identification procedure in an end-to-end manner. We validate the effectiveness of our framework by comparing the results with the state-of-the-art algorithms on three Re-ID benchmark datasets: Market-1501 [22], DukeMTMC-reID [23] and CUHK03 [9].

The rest of this paper is organized as follows. Section 2 briefly reviews the related work of Re-ID task. Section 3 describes the framework of our attention-based approach in details. Section 4 shows the experiment results on three public benchmark datasets and gives the corresponding analysis. Section 5 concludes our work.

2 Related Work

For an image-based person re-identification system, feature extraction and distance learning are typically considered as the two main components. Various algorithms have been proposed to solve the Re-ID problem. Some of them complete the above two steps separately, others utilize the deep learning-based method to achieve a jointly-learning procedure.

In order to present a better representation of given images, the basic principle is to find features unrelated to illumination, pose, background and viewpoints. Many existing algorithms have extracted more discriminative and viewpoint-invariant features. To get more information about the mean attributes of pixels, Matsukawa [3] exploit a hierarchical Gaussian distribution descriptor (GoG), modelling both means and covariances. In [24], An et al. conduct the matchings by projecting the original features into a reference subspace instead of matching directly. Experiments validate the effectiveness of the reference descriptors (RDs) generated with the correlations of the reference sets. To enhance the Re-ID performance, [1, 4] utilize different fusing schemes to combine the features obtained at different levels, including pixel-based low-level features (e.g. SIFT (scale-invariant feature transform) [25]), mid-level features (e.g. BoW (Bag of Words) [22]), and high-level features (e.g. features jointly learned in CNN [14]).

In addition to methods concerned with image representation, many approaches aim at finding a proper similarity measurement which makes the distance metric discrimitive. In [26], a null space is learned in which the in-class distance is minimized and between-class gap is maximized. Yu et al. [27] propose an unsupervised approach to deal with the limitation of lacking labelled sample images. The algorithm learns an asymmetric metric with projections on each viewpoint respectively. Considering the cross-view distortion of features, Chen et al. [28] formulate a Camera coRrelation Aware Feature augmenTation (CRAFT) framework, which automatically calculates the cross-view camera correlation. With learned features projected into an adaptive subspace, the CRAFT framework obtains view-specific features for Person Re-ID. [29, 30, 31] take the advantage of deep learning architecture and optimize the task of feature selection, similarity learning and ranking jointly.

Most of the Re-ID methods are proposed under the assumption that the probe and gallery images are well-aligned. However, due to the different viewpoints and the change of pedestrian poses, misalignment is a key issue to be considered. To solve this problem, visual attention schemes and saliency-based techniques are adopted. Some studies pay attention to saliency learning. Zhao et al. [32] apply pedestrian saliency distribution learning and estimate the scores based on the matching of constrained patches. The whole learning and matching procedure is unified into a RankSVM framework. In [33], a weighted integration scheme is adopted, combing human salience information with SDALF (Symmetry-Driven Accumulation of Local Features) [34]. With the rotation invariant attributes, the experimental performance is improved. Attention mechanism can also be seen as a concern for salient regions. With the recent success of exploiting attention modules in other fields, a few attention-based deep learning networks have been adopted to tackle the problem of misalignment in Re-ID. Gradient-based attention mechanism is exploited in [35]. In [36], both the CNN-extracted features and color histograms are fed into the recurrent attention architecture to perform a coarse-to-fine selection process. Liu et al. [37] formulate a Comparative Attention Network (CAN) architecture which takes image triplets as the training input. The CAN algorithm compares different local regions repeatedly instead of taking just one glimpse of the whole image. The global features learned from CNNs are delivered to a LSTM-based attention module to obtain the visual attentive features. The model in [38] also utilizes LSTM units to generate the spatial attetion encoding. In [39], a Harmonious Attention CNN (HA-CNN) model is proposed. HA-CNN adopts multi-branch scheme to process both local level and global level attention.

For each branch, soft attention and hard attention are combined to fully explore the complementary information.

Different from the attention-related algorithms mentioned above, our proposed attention network not only explores spatial attention, but also fuses the attention weighted feature maps learned at different layers for a richer representation. Consequently, the adopted method outperforms the state-of-the-art approaches.

3 Network Architecture

3.1 Overview

We adopt the attention-based model for person Re-ID task. The overall framework is illustrated in figure 2. We formulate an end-to-end Convolutional Neural Network to learn attentive features. Through exploiting spatial attention on multiple layers, our proposed network is able to make the original CNN-based feature maps adaptive to the more discriminative local regions and attributes. The spatial attention aims at encoding "where" to pay attention to. The feature integration scheme results in a more robust feature representation.

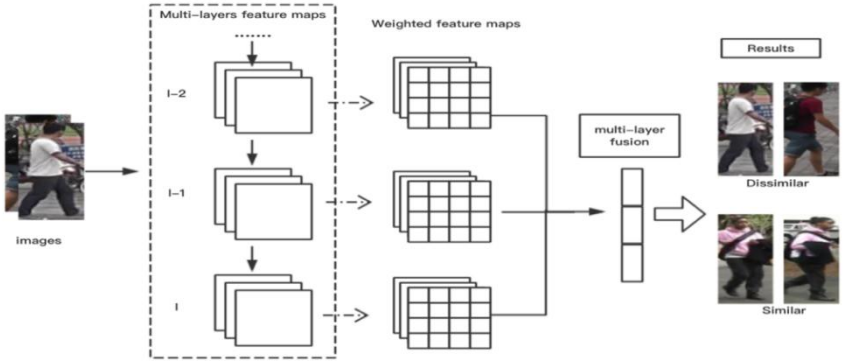

Fig. 2. The overview of the proposed multi-layer attention network architecture. For the feature map $\boldsymbol{V^l}$ on the l-th convolution layer, we learn a spatial attention weighted feature map $\boldsymbol{A^l}$ which is of the same size as $\boldsymbol{V^l}$. The attention modules are applied to multiple layers. The multi-layer attentive features are finally fused.

Given $\boldsymbol{n}$ distinct pedestrians, $\boldsymbol{m}$ training images captured from cameras of different viewpoints, the task is to learn a model that better describes all the pedestrians in spite of the variation of poses and changes. Our proposed model learns the attention-related features on the basis of a CNN base network. This basic CNN structure can be replaced by any CNN model, e.g. VGG-19, ResNet-101. The key components are the attention learning module and multi-layer fusion module. In additional to the final layer, one or more mid-layers of the CNN base network are exploited to construct the multiple attention weighted features.

For attention selection mechanism, we formulate an attention module which learns spatial attention. At the $\boldsymbol{l}$-th layer of the base CNN network, the feature map is denoted as $\boldsymbol{V^l}$. The $\boldsymbol{l}$ -th layer attention weights $\boldsymbol{\gamma^l}$ are learned from $\boldsymbol{V^l}$. The functions are formulated as:

$$V^l = CNN(V^{l-1}) \tag{1}$$

$$\gamma^l = \phi(V^l) \tag{2}$$

$$A^l = \eta(V^l, \gamma^l) \tag{3}$$

where $\boldsymbol{\phi}$ refers to the attention learning function, which will be elaborated in the next attention sections. η is a linear weighting calculation which combines both original feature representation $\boldsymbol{V^l}$ and learned weights $\boldsymbol{\gamma^l}$. $\boldsymbol{A^l}$ is the weighted map, considered as the output of an attention module.

Note that the obtained output map $A^l \in R^{h\times w\times c}$ is of the same size as the CNN original input map $V^l \in R^{h\times w\times c}$, where $\boldsymbol{h}$, $\boldsymbol{w}$ and $\boldsymbol{c}$ denote the pixels of the feature map in height dimension, width dimension and channel dimension. The spatial attention maps are generated from the representation map $\boldsymbol{V^l}$.

3.2 Spatial Attention

Spatial attention can be intuitively explained and conforms to the visual processing mechanism naturally. Instead of treating all the local regions equally, the spatial attention focuses on the selected regions and helps to enhance the representation. Inspired by the region enhancement characteristic, we adopt the spatial attention module in our formulated network to tackle the misalignment problem for person Re-ID.

Given the original generated feature map $V^l \in R^{h\times w\times c}$, the spatial attention module aims at calculating a probability map that indicates the attention probabilities of all local regions. The spatial attention weighted feature can be formulated as:

$$f^S = \sum_i^h \sum_j^w f_{i,j}^S = \sum_i^h \sum_j^w \alpha_{i,j} f_{i,j}^V \ (0 \le \alpha_{i,j} \le 1) \tag{4}$$

where $f_{i,j}^V$ represents the feature at the location of (***i***, ***j***) in feature map $\boldsymbol{V^l}$, and $\alpha_{i,j}$ represents the corresponding attention probability at the same location. At (***i***, ***j***), the learned spatial attended feature is denoted as $f_{i,j}^S$. In the equation, $\boldsymbol{f^S}$ refers to the final spatial attention representation.

3.3 Multi-layer Fusion

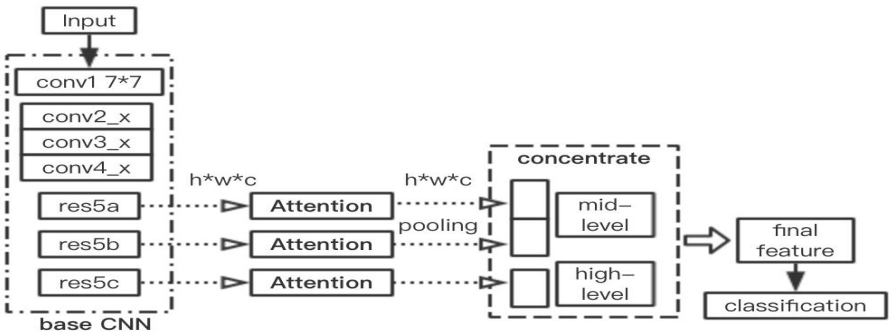

Fig. 3. The adopted CNN network structure. ResNet-50 is adopted as the base CNN network for Re-ID task. For the last three conv-layers ***res5a***, ***res5b*** and ***res5c***, the attention module is applied to obtain attention-based feature maps. After pooling methods, the mid-level features and high-level features are concentrated for final classification.

We choose ResNet-50 [40] as our CNN base network for person Re-ID task. Figure 3 shows the network structure of our proposed method. For ResNet-50 network, ***res5c*** represents the last convolutional layer, which contains the image representation in high-level. To take full advantage of the mid-level semantic information, we also select layers ***res5a*** and ***res5b*** as a supplement for better description. The feature maps of the above

layers are denoted as V^{res5a}, V^{res5b} and V^{res5c}. Instead of fusing the multiple original feature vectors, we concentrate the weighted feature maps after the attention modules, denoted as A^{5a}, A^{5b} and A^{5c}.

By adopting the appropriate pooling scheme, the feature vectors are constructed from multiple attention-weighted layers. For the maps A^{5a}, A^{5b}, A^{5c} that is corresponding to the selected layers, we leverage the global average pooling layer to generate the vectors f^{5a}, f^{5b} and f^{5c}. The concentration of mid-level feature vectors f^{5a} and f^{5b} are connected by a fully connected layer, which helps to reduce the feature dimension. After that, the generated mid-level vector f^{mid} is fused with f^{5c} to obtain the final feature vector for classification.

4 Experiments

In this section, we evaluate the performance of our proposed attention-based multi-layer integrated method. We validate the effectiveness of the multi-layer fusion strategy. The assessment is conducted on three benchmark datasets comparing with the state-of-the-art baseline algorithms.

4.1 Dataset

We perform the evaluation experiments on three public Re-ID benchmark datasets, i.e. Market-1501 [22], DukeMTMC-reID [23] and CUHK03 [9].

4.1.1 Market-1501 Dataset

The Market-1501 Dataset [22] contains 1,501 pedestrians captured from 6 non-overlapping cameras, with 5 of high resolution and 1 of low resolution. There are altogether 32,668 annotated bounding boxes, using DPM (Deformable Part Model) as the detector for pedestrian. In the experiment, we set the training and test sets as provided, including 750 persons in training set and 751 in testing set.

4.1.2 DukeMTMC-reID Dataset

The DukeMTMC-reID Dataset [23] involves 1,404 identities and 36,411 bounding box images. All the dataset images are cropped from video frames recorded by 8 different high-resolution cameras. The whole set is partitioned into two parts, with 702 individuals forming the training set, and 702 identities forming the testing set.

4.1.3 CUHK03 Dataset

The CUHK03 Dataset [9] consists of 13,164 images from 1,360 individuals captured by 6 cameras. Part of the images are manually cropped and the rest are detected by DPM. There are 767 individuals in training set and 700 identities in the test set.

4.2 Experimental Setup

In our formulated network, we exploit the widely-adopted CNN model: ResNet-50 [40] as the basic CNN network. The method is implemented using the deep learning framework Pytorch. All the pedestrian images in the datasets are resized to 160×64. The human body ratio is thus kept to remain undistorted. For optimization step, we choose Adam as the optimizer. The initial learning rate and the decay factor is set to 0.0005 and 0.95 respectively. We set the batch size as 32, and the maximum training epoch as 100. For performance evaluation, we use the standard CMC (Cumulative Matching Characteristic),

which represents the relationship between the correct identification rate and rank numbers, and mAP (Mean Average Precision), which measures the precision of classification. Our whole attention-based framework is completed in an end-to-end procedure.

4.3 Evaluation of Multi-layer Attention

In this section, we investigate the improvement effect of our multi-layer attention fusion mechanism. We conduct the experiments by reducing the number of layers followed by the attention module. For our adopted ResNet-50 network, 1-st layer, 2-nd layer, 3-rd layer refer to ***res5a***, ***res5b*** and ***res5c*** respectively. Table 1 shows the experiment results. In the table, ***1-layer*** refers to the method of applying the attention module to ***res5c*** layer only. ***2-layer*** represents the fusion of learned attention features on ***res5b*** and ***res5c***. And ***3-layer*** denote our adopted approach.

Table 1. Evaluation on the multi-layer fusion scheme (Market-1501 Dataset).

	Rank1	Rank5	Rank10	Rank20	mAP
1-layer	83.1	91.6	93.9	95.3	66.3
2-layer	86.6	93.3	95.1	96.4	69.5
3-layer	**87.1**	**94.8**	**96.6**	**97.8**	**70.1**

As can be observed from table 1, compared to modulating the attention network directly to the last convolutional layer, the 2-layer fusion method gains 3.5% for Rank-1, 2.3% for Rank-5, 2.8% for Rank-10, 0.9% for Rank-20 and 3.2% for mAP on Market-1501. The accuracy of 3-layer fusion method has a minor increase of 0.5% for Rank-1 accuracy. The experiments indicate that the multi-layer attention mechanism has a positive effect for better image representation.

4.4 Comparison with State-of-the-arts

4.4.1 Evaluation on Market-1501

We compare the results of our method with recent state-of-the-arts on Market-1501 dataset, as shown in table 2. The algorithms consist of feature extraction methods, i.e. CRAFT [28], GLAD [41], Zhao et al. [42], metric learning methods, i.e. SCSP [43], DNS [26], and deep network-based methods, i.e. CAN [37], PIE+Kissme [44], PDC [45]. As can be observed from table 2, our method outperforms the 2-nd best algorithm PDC by 3.0% in Rank-1 accuracy and 6.7% in mAP. Among all the compared approaches, CAN and PDC use attention mechanism for person Re-ID task in their algorithms as well. Statistics indicate that our attention-based deep network outperforms other approaches using attention information. It is worth noting that PIE also uses ResNet-50 as the base network architecture, with part-aligned representation based on the pose estimator. Our proposed method surpasses PIE by 8.4% for Rank-1 accuracy. These statistics show the advantage of our multi-layer attention model over the state-of-the-arts baseline.

Table 2. Evaluation on the Market-1501 Dataset.

Method	Rank1	Rank5	Rank10	Rank20	mAP
SCSP [43]	51.9	72.0	79.0	-	26.4
CAN [37]	60.3	-	-	-	35.9
DNS [26]	61.0	-	-	-	35.7
CRAFT [28]	68.7	-	-	-	42.3
GLAD [41]	61.4	76.8	81.6	85.9	34.0
PIE [44]	78.7	90.3	93.6	95.7	53.9
PIE+Kissme [44]	79.3	90.7	94.4	96.5	56.0
Zhao et al. [42]	81.0	92.0	94.7	-	63.4
PDC [45]	84.1	92.7	94.9	96.8	63.4
Ours	**87.1**	**94.8**	**96.6**	**97.8**	**70.1**

4.4.2 Evaluation on DukeMTMC-reID

Table 3 indicates the superiority of our multi-layer attention integration method compared to the state-of-the-arts on DukeMTMC-reID dataset. Compared to Market-1501 dataset, DukeMTMC-reID dataset has more complex scenes and more changes in background. Except for LOMO+XQDA [2], all the other compared approaches, i.e. SVDNet [46], APR [47] and PAN [48], formulate the framework on the basis of the deep learning network. On DukeMTMC-reID dataset, we achieve Rank-1 accuracy at 78.8%, and mAP at 60.0%, surpassing the 2-nd method SVDNet (ResNet-50) by 0.1% and 3.2%, respectively. This suggests that the attention mechanism plays a significant role in feature representation and our proposed network adapts to the complex scenes well.

Table 3. Evaluation on the DukeMTMC-reID Dataset.

Method	Rank1	Rank5	Rank10	mAP
LOMO+XQDA [2]	51.9	72.0	79.0	17.0
SVDNet (CaffeNet) [46]	60.3	-	-	45.8
APR [47]	61.0	-	-	51.9
PAN+Re-rank [48]	61.4	76.8	81.6	56.7
SVDNet (ResNet-50) [46]	78.7	90.3	93.6	56.8
Ours	**78.8**	**90.6**	**93.9**	**60.0**

4.4.3 Evaluation on CUHK03

Table 4 shows the CMC Rank accuracy and mAP of several recent proposed algorithms reported on CUHK03 dataset. We conduct the evaluation on both manually cropped bounding boxes and automatically detected images. Compared to the manually labelled images, the detected ones have more misalignment problems, presenting a more difficult task. In our evaluation, the compared counterparts include HACNN [39], LOMO+XQDA [2], LOMO+XQDA+re-rank [49], Ahmed et al. [7] and IDE+Re-rank [49]. HACNN proposes a harmonious network which combines soft attention in pixel and hard attention in region, exploiting attention mechanism in deep learning network. Our attention method achieves the performance of Rank-1 = 59.2%, mAP = 57.9% for manually labelled images, and Rank-1 = 56.3%, mAP = 54.8% for detected images. It shows the superiority of the adopted attention learning approach over other attention driven mechanism.

Table 4. Evaluation on the CUHK03 Dataset.

	Manual		**Detected**	
Method	Rank1	mAP	Rank1	mAP
HACNN [39]	44.4	41.0	41.7	38.6
LOMO+XQDA [2]	49.7	56.4	44.6	51.5
LOMO+XQDA+Re-rank [49]	50.0	56.8	45.9	52.6
Ahmed et al. [7]	54.7	-	45.0	-
IDE+Re-rank [49]	57.2	**63.2**	54.2	**60.5**
Ours	**59.2**	57.9	**56.3**	54.8

Despite the disparities among these three Re-ID baseline datasets, i.e. scenes, camera views, and image processing methods, we show that our proposed approach improves the performance for Re-ID task. The experiment results demonstrate the excellence of our attention-based framework.

5 Conclusion

In this paper, we present a novel attention-based method for person re-identification task. The whole framework is formulated in an end-to-end procedure. Our network benefits from both attention mechanism and multi-layer integration mechanism. For attention selection, we explore spatial attention on different local regions. The introduction of attention module takes full advantage of the complementary attention information. For multi-layer fusion module, the concentration of mid-level features and high-level features performs a better image representation. The detailed analysis of feature fusion scheme on multiple layers is provided. We conduct the performance evaluation on three baseline Re-ID datasets, i.e. Market-1501, DukeMTMC-reID and CUHK03. Experiments validate the effectiveness of our proposed model when compared to the state-of-the-art approaches.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant 2017YFB1002401.

References

1. Aske R Lejbølle, Nasrollahi K and Thomas B Moeslund 2018 Enhancing person re-identification by late fusion of low-, mid- and high-level features *Iet Biometrics* **7(2)** 125-135
2. Liao S, Hu Y, Zhu X and Li S Z 2015 Person re-identification by Local Maximal Occurrence representation and metric learning *IEEE Conference on Computer Vision and Pattern Recognition* **Vol.8** 2197-2206
3. Matsukawa T, Okabe T, Suzuki E and Sato Y 2016 Hierarchical Gaussian Descriptor for Person Re-identification *Computer Vision and Pattern Recognition* 1363-1372
4. Tan F, Liu K and Zhao X 2017 Person Re-Identification Based on Multi-Level and Multi-Feature Fusion *International Conference on Smart City and Systems Engineering* 184-187
5. Jose C and Fleuret F 2016 Scalable Metric Learning via Weighted Approximate Rank Component Analysis *European Conference on Computer Vision* 875-890
6. Yang Y, Liao S, Lei Z and Li S Z 2016 Large scale similarity learning using similar pairs for person verification *Thirtieth AAAI Conference on Artificial Intelligence* 3655-3661
7. Ahmed E, Jones M and Marks T K 2015 An improved deep learning architecture for person re-identification *Computer Vision and Pattern Recognition* 3908-3916
8. Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *International Conference on Neural Information* Processing *Systems* **Vol.60** 1097-1105
9. Li W, Zhao R, Xiao T, and Wang X 2014 DeepReID: Deep Filter Pairing Neural Network for Person Re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 152-159
10. Varior R R, Haloi M and Wang G 2016 Gated Siamese Convolutional Neural Network Architecture for Human Re-identification *European Conference on Computer Vision* 791-808

11. Varior R R, Shuai B, Lu J, Xu D and Wang G 2016 A Siamese Long Short-Term Memory Architecture for Human Re-identification *European Conference on Computer Vision* 135-153
12. Yu Q, Chang X, Song Y Z, Xiang T and Hospedales T M 2017 The devil is in the middle: exploiting mid-level representations for cross-domain instance matching *ArXiv e-prints* 1711.08106
13. Wu S, Chen Y C, Li X, Wu A C, You J J and Zheng W S 2016 An enhanced deep feature representation for person re-identification *Applications of Computer Vision* 1-8
14. Yi D, Lei Z, Liao S and Li S Z 2014 Deep Metric Learning for Person Re-identification *International Conference on Pattern Recognition* 34-39
15. Chen L, Zhang H, Xiao J, Nie L, Shao, J and Liu W 2017 Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning *Computer Vision and Pattern Recognition* 6298-6306
16. Xu K, Ba J, Kiros R, Cho K, Courville A and Salakhutdinov R 2015 Show, attend and tell: neural image caption generation with visual attention *Computer Science* 2048-2057
17. You Q, Jin H, Wang Z, Fang C and Luo J 2016 Image Captioning with Semantic Attention *Computer Vision and Pattern Recognition* 4651-4659
18. Choi H, Cho K and Bengio Y 2018 Fine-grained attention mechanism for neural machine translation *Neurocomputing* 284
19. Luong M T, Pham H and Manning C D 2015 Effective approaches to attention-based neural machine translation *Computer Science*
20. Chen K, Wang J, Chen L C, Gao H, Xu W and Nevatia R 2015 Abc-cnn: an attention based convolutional neural network for visual question answering *Computer Science*
21. Lioutas V, Passalis N and Tefas A 2018 Visual Question Answering using Explicit Visual Attention *IEEE International Symposium on Circuits and Systems* 1-5
22. Zheng L, Shen L, Tian L, Wang S, Wang J and Tian Q 2015 Scalable Person Re-identification: A Benchmark *IEEE International Conference on Computer Vision* 1116-1124
23. Ristani E, Solera F, Zou R, Cucchiara R and Tomasi C 2016 Performance Measures and a Data Set for Multi-target, Multi-camera Tracking *European Conference on Computer Vision* 17-35
24. An L, Kafai M, Yang S and Bhanu B 2016 Person reidentification with reference descriptor *IEEE Transactions on Circuits & Systems for Video Technology* **26(4)** 776-787
25. Zhao R, Ouyang W and Wang X 2014 Learning Mid-level Filters for Person Re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 144-151
26. Zhang L, Xiang T and Gong S 2016 Learning a Discriminative Null Space for Person Re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 1239-1248
27. Yu H X, Wu A and Zheng W S 2017 Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification *IEEE International Conference on Computer Vision* 994-1002
28. Chen Y C, Zhu X, Zheng W S and Lai J H 2018 Person re-identification by camera correlation aware feature augmentation *IEEE Transactions on Pattern Analysis & Machine Intelligence* **40(2)** 392-408.

29. Chen W, Chen X, Zhang J and Huang K 2017 A multi-task deep network for person re-identification *AAAI Conference on Artificial Intelligence*
30. Li W, Zhu X and Gong S 2017 Person re-identification by deep joint learning of multi-loss classification *International Joint Conference on Artificial Intelligence* 2194-2200.
31. Wang F, Zuo W, Lin L, Zhang D and Zhang L 2016 Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 1288-1296
32. Zhao R, Oyang W and Wang X 2014 Person re-identification by saliency learning *IEEE Transactions on Pattern Analysis & Machine Intelligence* **39(2)** 356-370
33. Yang M, Wan W, Hou L and Zhang Y 2016 Person re-identification using human salience based on multi-feature fusion *International Conference on Smart and Sustainable City and Big Data* 5
34. Bazzani L, Cristani M and Murino V 2014 Sdalf: modeling human appearance with symmetry-driven accumulation of local features *Person Re-Identification* **63(4)** 43-69
35. Rahimpour A, Liu L, Taalimi A, Song Y and Qi H 2017 Person re-identification using visual attention *IEEE SigPort* http://sigport.org/2046
36. Zhuang Z, Ai H, Shang C and Xiao L 2017 Person re-identification with coarse-to-fine visual attention *IEEE International Conference on Image Processing* 1097-1101
37. Liu, H., Feng, J., Qi, M., Jiang, J and Yan S 2017 End-to-end comparative attention networks for person re-identification *IEEE Transactions on Image Processing* **26(7)** 3492-3506
38. Wu L, Wang Y, Li X and Gao J 2018 Deep attention-based spatially recursive networks for fine-grained visual recognition *IEEE Transactions on Cybernetics* **PP(99)** 1-12
39. Li W, Zhu X and Gong S 2018 Harmonious attention network for person re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 2285-2294
40. Yang Y, Liao S, Lei Z and Li S Z 2016 Large scale similarity learning using similar pairs for person verification *Thirtieth AAAI Conference on Artificial Intelligence* 3655-3661
41. Wei L, Zhang S, Yao H, Gao W and Tian Q 2017 GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval *ACM Multimedia*
42. Zhao L, Li X, Zhuang Y and Wang J 2017 Deeply-Learned Part-Aligned Representations for Person Re-identification *IEEE International Conference on Computer Vision* 3239-3248
43. Chen D, Yuan Z, Chen B and Zheng N 2016 Similarity Learning with Spatial Constraints for Person Re-identification *IEEE Conference on Computer Vision and Pattern Recognition* 1268-1277
44. Zheng L, Huang Y, Lu H and Yang Y 2017 Pose invariant embedding for deep person re-identification *ArXiv e-prints* 1701.07732
45. Su C, Li J, Zhang S, Xing J, Gao W and Tian Q 2017 Pose-driven deep convolutional model for person re-identification *IEEE International Conference on Computer Vision* 3980-3989
46. Sun Y, Zheng L, Deng W and Wang S 2017 SVDNet for Pedestrian Retrieval *IEEE International Conference on Computer Vision* 3820-3828
47. Lin Y, Zheng L, Zheng Z, Wu Y and Yang Y 2017 Improving person re-identification by attribute and identity learning *ArXiv e-prints* 1703.07220

48. Zheng Z, Zheng L and Yang Y 2017 Pedestrian alignment network for large-scale person re-identification *ArXiv e-prints* 1707.00408
49. Zhong Z, Zheng L, Cao D and Li S 2017 Re-ranking Person Re-identification with k-Reciprocal Encoding *IEEE Conference on Computer Vision and Pattern Recognition* 3652-3661