Person Re-Identification by Localizing Discriminative Regions

Tanzila Rahman
rahmant4@cs.umanitoba.ca
Mrigank Rochan
mrochan@cs.umanitoba.ca
Yang Wang
ywang@cs.umanitoba.ca

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

Abstract

Person re-identification is a challenging task of matching a person's image across multiple images captured from different camera views. Recently, deep learning based approaches have been proposed that show promising performance on this task. However, most of these approaches use whole image features to compute the similarity between images. This is not very intuitive since not all the regions in an image contain information about the person identity. In this paper, we introduce an end-to-end Siamese convolutional neural network that firstly localizes discriminative salient image regions and then computes the similarity based on these image regions in conjunction with the whole image. We use Spatial Transformer Networks (STN) for localizing salient regions. Extensive experiments on CUHK01 and CUHK03 datasets show that our method achieves the state-of-the-art performance.

1 Introduction

Person re-identification is an important problem in many real-world applications, such as video surveillance. The goal of person re-identification is to identify a specific person in an input image (known as the probe image) from a set of gallery images captured by non-overlapping and different cameras. It is a very challenging problem due to the complex variations in viewpoints, poses, lighting, illuminations, blurring effects, and image resolutions. The intra-person variations can even be larger than inter-person variations in this task [27]. Backgrounds and occlusions also create challenges in person re-identification.

Sometimes small objects or regions convey important information about the person identity in an image. Humans can recognize person identity based on these salient regions. For example, in Fig.1, person (a) carries a backpack, person (b) wears a white jacket, person (c) holds an orange colored jacket in his hand and person (d) holds a file in her hand. These distinctive regions can be used to identify one person from others. Usually, if an object is salient in one camera view, it remains salient in another camera view too [28] even though there are variations in view points. In addition to salient objects, body parts as well as clothing can also be considered as informative region for identifying persons. Although salient regions in an image play a vital role in person re-identification for humans, most existing

Figure 1: Some examples of pedestrian images for person re-identification from CUHK01 train dataset. Each pair represents the same person from different camera views. The bounding box on each image shows the discriminative region localized by our proposed approach.

approaches in person re-identification do not capture this information. Most of the existing approaches [1, 11, 18, 21, 24] compute the similarity between two images based on whole image features.

In this paper, we propose a new person re-identification technique by explicitly localizing salient regions. In particular, we use Spatial Transformer Network (STN) [8] to localize the discriminative regions in the input images. Our multichannel CNN model then computes the similarity of the input images based on these discriminative regions in conjunction with whole image features.

The main contribution of this work is that it integrates attention-based STN in the person re-identification framework. This allows our model to focus on discriminative regions in the input images when computing their similarity. Moreover, we integrate global image features with the discriminative regions to produce final feature representation for person re-identification. To the best of our knowledge, this is the first CNN-based architecture that performs person re-identification by localizing discriminative image regions. Our model can be trained end-to-end and it does not require supervision or any prior knowledge about the discriminative regions. We demonstrate that our approach achieves state-of-the-art results on several benchmark datasets.

2 Related Work

Previous work on person re-identification can be classified into two broad groups: non-deep learning methods and deep learning methods.

Non-Deep Learning Methods: Most of the person re-identification methods consist of two components: (1) a method to extract features from the input images, and (2) a way of computing a similarity metric to decide whether the images belong to same person or not. Much of the previous research focuses on either improving feature extraction method [6, 10, 16, 27], or robust similarity metric learning [2, 7, 9, 15, 27], or their combination [11, 14, 17, 26]. Although these approaches are promising, their performance is limited due to the heavy reliance on handcrafted features. In contrast, our approach is based on deep learning which simultaneously learns the feature representation and a similarity metric to optimize the performance.

Deep Learning Methods: In recent years, deep neural networks have significantly improved the state-of-the-art in many computer vision tasks such as image classification and object detection. There are a few previous works that use deep learning for person re-identification problem in the literature. Our work is mostly related to the work by Yi et al. [25], Li et

al. [12], Ahmed et al. [1], and Subramaniam et al. [21]. Yi et al. [25] propose a Siamese convolutional network for re-identification. Their network takes a pair of images as its input to which three stages of convolution are performed followed by a fully-connected operation that outputs a vector for each input image. Lastly, cosine similarity function is used to compare the two output vectors. Li et al. [12] use a two-input network architecture that firstly performs a set of convolutions to the inputs and then multiplies the convolution feature maps at different horizontal offsets. This is followed by a max-out grouping which filters out the highest response from horizontal strips to which another convolution and max pooling operation is done. Finally, the output is used to compute the similarity. Ahmed et al. [1] introduce a deep architecture that contains two new layers: cross-input neighborhood layer and patch summary layer. Cross-input neighborhood layer is used to learn the relationship between feature maps of two input images. Patch summary layer is responsible for summarizing the neighborhood maps by analysing the differences in each 5x5 block, which are then used to measure the similarity of two input images. Our model is motivated by recent work in [21] which extends the work of [1]. The work in [21] uses a fused network that performs inexact matching through a novel layer called Normalized X-Corr whose output assists the subsequent layers in making decision on whether the two input images are similar or not. The main difference between these previous approaches and ours is that, instead of using only whole image feature maps to compare the two input person images similarity, we firstly localize discriminative regions in the images and then forward their feature maps in addition to the global images to subsequent layers for similarity computation. Our work is driven by the intuition that the input images contain a lot of background pixels which are irrelevant for person re-identification.

Our work is also related to the recent work on localizing and ranking visual attributes given a pairwise image comparison [20]. This work uses STN to localize the image regions that are relevant for the visual attribute. Similar to [20], we also incorporate STN to localize discriminative regions in images that are relevant for person re-identification.

3 Our Approach

We formulate person re-identification as a binary classification problem given two input images. Our proposed model learns a function f that maps an image pair (I_1, I_2) to a score that indicates how likely these two images correspond to the same person. During training, our network takes an image pair (I_1, I_2) and a binary label L indicating whether the images are similar or not. During testing, the input is an image pair (I_{test1}, I_{test2}) and the network uses the learned function f with parameters w to predict the similarity score $f(I_{test1}, I_{test2})$ between the image pair.

Figure 2 shows the overall architecture of our model. Our model is based on the Siamese network [4]. It has two Siamese networks whereas each network contains two branches with shared parameters. There are two main components in the network: (a) *Spatial Transformer Network* (SN) and (b) *Fused Network* (FN). The STN is used to learn to localize the discriminative region in an image and generate a feature representation based on this region. The FN is used to combine the features of discriminative regions in both input images and output a similarity score.

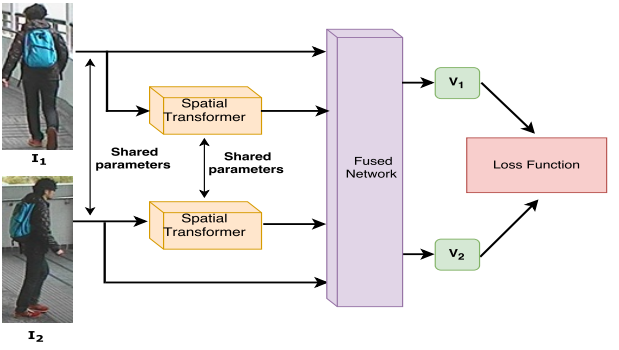

Figure 2: Overall architecture of our network. It takes two person images (I_1, I_2) as its input. Each image is forwarded to two Siamese-CNN architecture whereas one contains a Spatial Transformer Network (STN) with a Fused Network (FN) and another contains only fused network. The model finally produces two outputs/scores (v_1, v_2) indicating similarity strength of two input person images which is later fed to a loss function to update the parameters of the network.

3.1 Spatial Transformer Network

Previous work on person re-identification typically compares the similarity of two images based on features extracted from the entire image. We believe this is not optimal, since an image usually contains a lot of pixels (e.g. background pixels) that are irrelevant for person re-identification. Humans usually differentiate between a pair of images by focusing on certain distinct regions/parts of the person in the image (see Fig. 1). In our work, we develop a model that has the same capability. In our model architecture, we incorporate STN for localizing discriminative regions that are relevant for person re-identification. STN is a fully-differentiable module that can learn spatial transformations, such as scaling, rotation and translation without any additional supervision.

We incorporate STN in our network so that it can focus on discriminative regions which would be used for subsequent parts of the network. The output of STN will simplify the task of Fused network (FN) as it can be optimized efficiently over the localized discriminative regions for a given pair of images.

As outlined in [8], there are three main components in STN (see top of Fig. 3): i) Localization network, which takes the input image and produces the transformation parameters θ; ii) Grid generator, which generates a sampling grid using the transformation parameters. The sampling grid is a set of points where the input feature map should be sampled to produce the transformed output; and iii) Sampler, which uses a bilinear interpolation kernel to produce the output image. In this work, we use STN that has three transformation parameters $\theta = [s, t_x, t_y]$, where s, t_x and t_y represent isotropic scaling, horizontal and vertical translation respectively. This transformation parameters are constrained for attention [8], and the point transformation is

$$\begin{pmatrix} {x_i}^{in} \\ {y_i}^{in} \end{pmatrix} = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \begin{pmatrix} {x_i}^{out} \\ {y_i}^{out} \\ 1 \end{pmatrix} \tag{1}$$

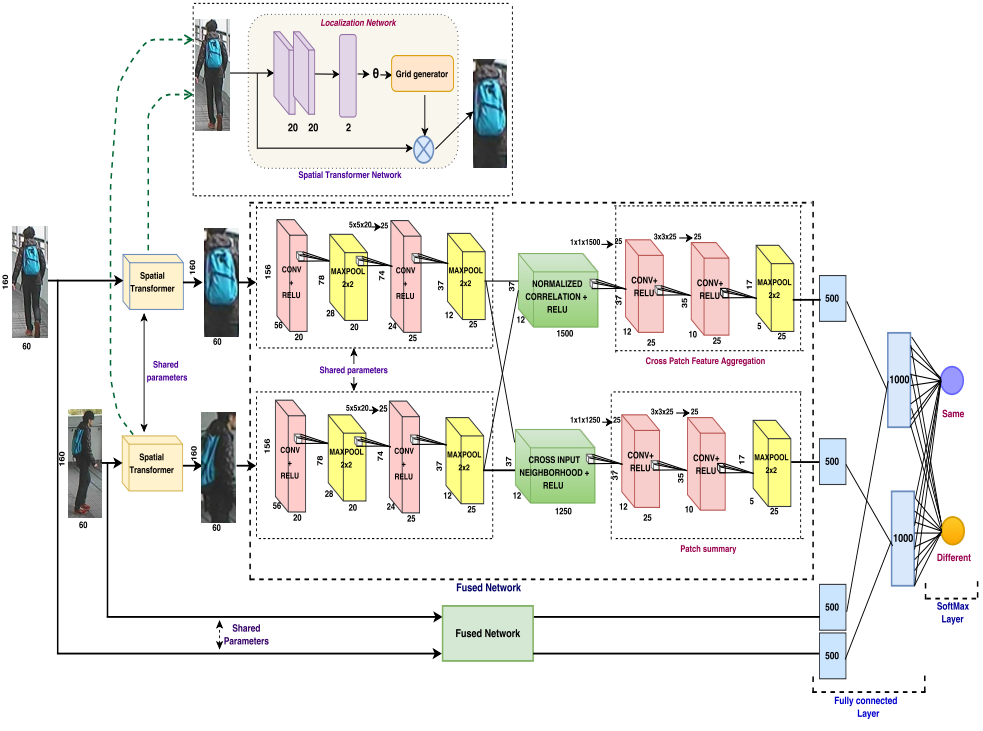

Figure 3: Detailed architecture of our proposed network. The network takes a pair of image as input. Each image goes through the spatial transformer network (STN), which localizes the discriminative image region. The output of STN is fed to Fused network which generates two linear layers with 500 output values as features of the discriminative region. At the same time, the input images go through another fused network which also produces two linear layers of 500 output values as global image features. The features from the localized regions and the global images are concatenated and finally used to compute the similarity score of the two input images.

Here, ${x_i}^{in}$ and ${y_i}^{in}$ represent coordinates of the input image, whereas ${x_i}^{out}$ and ${y_i}^{out}$ represent output image coordinates at the i-th index. The localization network within the STN can take any form of convolutional network or fully-connected network, but finally it should include a regression layer that generates transformation parameters θ [8]. In this paper, we follow their localization layer architecture which uses STN for digit localization in images. A convolutional layer with 20 filters of size 5×5 and two fully-connected layers are added towards the end. The first fully-connected layer takes 6120 values as its input and produces 20 output values, whereas the second one takes 20 input values and produces 2 transformation parameters (t_x, t_y) as output. Here the network is not learning scaling parameter (s) as we fix it to 0.5. These parameters are used to generate the transformed output image patch through the sampling mechanism. Figure 4 shows the STN's localization behavior during training.

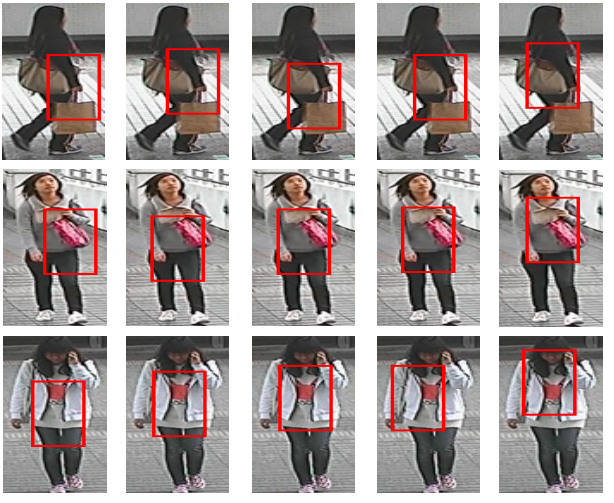

Figure 4: STN's localization behavior during training on CUHK01 dataset. Each row shows the localized image patch (in the red box) by STN for different training iterations. We find that STN converges to distinctive image regions after certain iterations.

3.2 Fused Network

The input images and the output of STN are fed to the Fused Network (FN) [21]. In our model, we use two fused networks separately. One of them takes a pair of image patches as its input whereas another one considers a pair of whole input images. Finally these two fused network outputs the similarity score indicating whether two image belong to the same person or not. The fused network is also a Siamese network where each branch contains two stage of convolutions (with shared parameters) and pooling layers. These convolution layers take input image of size $60 \times 160 \times 3$ and generate 25 feature maps of dimension 12×37 which is fed to the normalized correlation and the cross-input neighborhood layers. Given two feature maps, the normalized correlation layer [21] computes the correlation between every pair of 5×5 patch matrices. For given matrices X and Y, the Normalized Correlation is defined as [21]:

$$normCorr(X,Y) = \frac{\Sigma_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{(N-1).\sigma_X.\sigma_Y} \tag{2}$$

Here, μ_X and μ_Y are the mean values for two matrices X and Y respectively. Cross-input neighborhood layer [1] computes the difference between feature maps produced by the convolution layers of two branches of the Siamese network. The output of normalized correlation and the cross-input neighborhood are fed to separate cross patch feature aggregation layers which incorporate the contextual information and summarizes it. The feature aggregation layer is composed of two convolutions followed by max-pooling layer, and the output is 25 feature maps of size 5×17. The output feature maps are then fed to fully-connected layers of 500 hidden units. The fully-connected layers (one for patch image and another for global image) for each images are joined together with two softmax units. The output of the first softmax represents the likelihood that two images are same, and the other one

represents the likelihood that the images are different. To train our network, we use the standard cross-entropy loss and optimize the network parameters using the Stochastic Gradient Descent (SGD) algorithm.

Figure 3 shows the detailed architecture of our proposed network. Subramaniam et al. [21] also use FN for person re-identification. But the model in [21] computes the similarity of two input images only from the whole image features. In contrast, our proposed model first uses STN to localize the discriminative regions from the two input images, then the similarity score is computed based on these regions in addition to the whole images.

4 Experiments

4.1 Datasets

We conduct experiments on two benchmark datasets: CUHK01 [11] and CUHK03 [12] .

CUHK01 Dataset: This dataset consists of 3,884 images of 971 people [11]. For each person (or identity), there are 4 images captured from 2 different cameras. Following Subramaniam et al. [21], we conduct experiments in two different settings. In the first setting, we use 871 identities for training and the remaining 100 identities for testing. In the second setting, we use 485 identities for training and the remaining 486 identities for testing.

CUHK03 Dataset: This is one of the largest benchmark dataset for person re-identification. It consists of 13,164 images of 1,360 pedestrians captured by 6 different surveillance cameras [12]. Each person is observed by 2 disjoint camera views. The dataset contains two different types of pedestrian bounding boxes – one as a result of manually labeling (referred as Labeled dataset) and the other that is algorithmically generated (referred as Detected dataset). In this work, we conduct experiments on both types. Again, we follow the experiment protocol of Subramaniam et al. [21] by randomly picking 1,260 identities for training and the rest for testing.

4.2 Network Training Strategies

We treat person re-identification as a binary classification problem. So we train the network using pairs of similar (i.e. positive pair) and dissimilar (i.e. negative pair) images. There exists data imbalance in the datasets – there are more negative pairs than positive pairs. Following previous work [21], we perform data augmentation to deal with the data imbalance. For every training set image of size $W \times H$, we sample several image patches (2 image patches for CUHK03 and 5 image patches for CUHK01 Dataset) around the image center and then apply random 2D translation drawn from a uniform distribution within the range of $[-0.05W, 0.05W] \times [-0.05H, 0.05H]$. This data augmentation strategy alleviates the training data imbalance issue across the datasets.

We implement our network using Torch 7 [5]. We train our network with mini-batch of size 128. We use 0.9 as momentum and 0.05 as initial learning rate. Learning rate decay and weight decay are set to 1×10^{-4} and 5×10^{-4} respectively. We also fix the scaling value to 0.5 in the Spatial Transformer Network and learn translation parameters (t_x and t_y) only. Due to the data imbalance in most of the person re-identification dataset, after certain iteration the STN begins to consider whole image as patch. To mitigate this issue, we use fix scaling value to learn STN which gives better result along with the global image.

4.3 Evaluation Protocol

We present a comprehensive evaluation of our proposed method by comparing it with several state-of-the-art methods on CUHK01 and CUHK03 datasets. Following previous work, we rank the images present in the gallery image set based on the similarity with a probe image. Note that both the gallery images and the probe images are from test set. The intuition of this type of ranking is that the ground-truth matching gallary image should have the highest rank in the ideal case. In our experiments, we randomly select one image for each person/identity in the test set as a probe image and consider the remaining images as gallery images. For a probe image of a person, there is exactly one match in the gallery images. We perform 10 test trials on every probe image and report the averaged results in the tables along with several baselines. Note that the comparison with Subramaniam et al. [21], Ahmed et al. [1], and Li et al. [12] is of particular interest to us since they use similar deep learning architectures.

4.4 Results

CUHK01 Dataset: Table 1 and 2 summarize the experimental results on the CUHK01 dataset with 100 and 486 test identities. Our model outperforms the state-of-the-art method by nearly 5% in terms of the rank-1 accuracy. We believe that this performance gain is due to the discriminative regions learned by our network that is able to effectively distinguish between similar and dissimilar person images. Moreover, we train our network from scratch rather than pre-training it on a larger CUHK03 Labeled dataset, which is done by the state-of-the-art method in [21]. Note that the method in [21] is equivalent to our model without localizing the discriminative regions. Our model outperforms [21] by a large margin. This demonstrates the advantage of localizing discriminative regions in images for person re-identification.

Method	Rank-1	Rank-10	Rank-20
eSDC[26]	22.84	57.67	69.84
LDML[9]	26.45	72.04	84.69
KISSME[8]	29.40	72.43	86.07
Li et al.[12]	27.87	73.46	86.31
Ahmed et al.[1]	65.00	93.12	97.20
Wang et al.[23]	71.80	–	–
Subramaniam et al.[21]	81.23	97.39	98.60
Ours	**86.67**	**99.17**	**99.87**

Table 1: Performance of different methods at ranks 1, 10, and 20 on CUHK01 with 100 test IDs.

Method	Rank-1	Rank-10	Rank-20
Mid-Level Filters [27]	34.30	65.00	74.90
Mirror-KFMA [3]	40.40	75.30	84.10
Ahmed et al.[1]	47.50	80.00	87.44
Ensembles [19]	51.90	83.00	89.40
CPDL[10]	59.50	89.70	93.10
Subramaniam et al. [21]	65.04	89.76	94.49
Ours	**71.35**	**93.08**	**96.80**

Table 2: Performance of different methods at ranks 1, 10, and 20 on CUHK01 with 486 test IDs.

CUHK03 Dataset: Table 3 and 4 summarize the experimental results on the CUHK03 Labeled and Detected datasets, respectively. Our model outperforms the state-of-the-art [21] method by nearly 2% in terms of the rank-1 accuracy. Figure 5 shows some qualitative retrieval results on this dataset.

Method	Rank-1	Rank-10	Rank-20
eSDC [26]	8.76	38.28	53.44
LDML [9]	13.51	52.13	70.81
KISSME [9]	14.17	52.57	70.03
Li et al. [12]	20.65	68.74	83.06
LOMO+XQDA [12]	52.20	92.14	96.25
Ahmed et al. [1]	54.74	93.88	98.10
LOMO+MLAPG [13]	57.96	94.74	98.00
Ensembles [19]	62.10	92.30	97.20
Subramaniam et al. [21]	72.43	95.51	98.40
Ours	**77.80**	**98.49**	**99.52**

Table 3: Performance of different methods at ranks 1, 10, and 20 on the CUHK03 Labeled dataset.

Method	Rank-1	Rank-10	Rank-20
eSDC [26]	7.68	33.38	50.58
LDML [9]	10.92	47.01	65.00
KISSME [9]	11.70	48.08	64.86
Li et al. [12]	19.89	64.79	81.14
LOMO+XQDA [12]	46.25	88.55	94.25
Ahmed et al. [1]	44.96	83.47	93.15
LOMO+MLAPG [13]	51.15	92.05	96.90
Subramaniam et al. [21]	72.04	96.00	98.26
Ours	**74.48**	**96.16**	**98.28**

Table 4: Performance of different methods at ranks 1, 10, and 20 on the CUHK03 Detected dataset.

Figure 6 shows some typical failure cases of our approach.

5 Conclusion

In this paper, we have proposed an end-to-end deep neural network architecture that localizes discriminative image regions for person re-identification. Our proposed model achieves state-of-the-art results on CUHK01 and CUHK03 datasets. Currently, our model only localizes one discriminative region in each image. As future work, we plan to extend our model to localize multiple discriminative regions by using more than one STN in the model.

6 Acknowledgements

This work was supported by NSERC. We thank NVIDIA for donating some of the GPUs used in this work.

Figure 5: Qualitative retrieval results of our approach on CUHK03 dataset. The first column in each row represents a probe image. The remaining columns represent the retrieved results. The column highlighted in green is the ground-truth match.

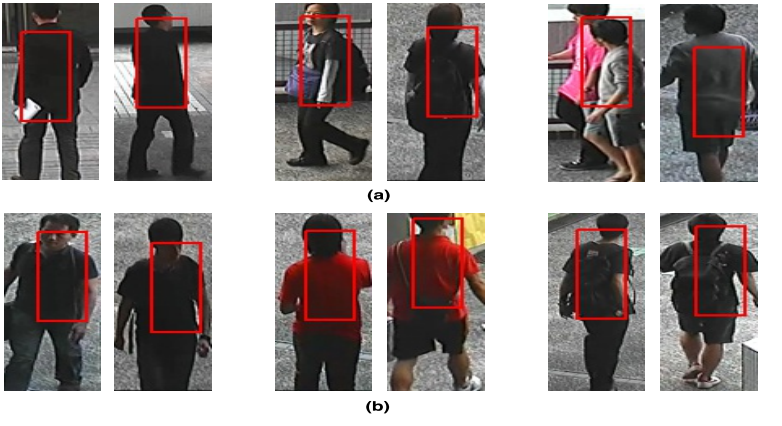

Figure 6: Some failure cases of our approach. (a) image pairs of the same person: our method incorrectly predict them as being dissimilar due to the lack of discriminative regions in these images; (b) image pairs of different persons: our method incorrectly recognize them as the same person, possibly because the localized discriminative regions in these image pairs have similar appearance.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[2] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1565–1573, 2015.

[3] Ying-Cong Chen, Wei-Shi Zheng, and Jianhuang Lai. Mirror representation for modeling view-specific transform in person re-identification. In *International Joint Conference on Artificial Intelligence*, pages 3402–3408. Citeseer, 2015.

[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–546, 2005.

[5] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, Neural Information Processing Systems Workshop*, 2011.

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.

[7] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, pages 498–505, 2009.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[9] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[10] Sheng Li, Ming Shao, and Yun Fu. Cross-view projective dictionary learning for person re-identification. In *International Joint Conference on Artificial Intelligence*, pages 2155–2161, 2015.

[11] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44. Springer, 2012.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[13] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

[14] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[15] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing*, pages 3567–3571, 2013.

[16] Niki Martinel, Christian Micheloni, and Gian Luca Foresti. Saliency weighted features for person re-identification. In *European Conference on Computer Vision*, pages 191–208. Springer, 2014.

[17] Niki Martinel, Christian Micheloni, and Gian Luca Foresti. Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, 2015.

[18] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2672, 2012.

[19] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.

[20] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016.

[21] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016.

[22] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[23] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1296, 2016.

[24] Lin Wu, Chunhua Shen, and Anton van den Hengel. Personnet: person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016.

[25] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, pages 34–39, 2014.

[26] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[27] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151, 2014.

[28] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2): 356–370, 2017.