Deeply-Learned Part-Aligned Representations for Person Re-Identification

Liming Zhao[†] Xi Li[†*] Yueting Zhuang[†] Jingdong Wang[‡*]
[†]Zhejiang University [‡]Microsoft Research
{zhaoliming,xilizju,yzhuang}@zju.edu.cn jingdw@microsoft.com

Abstract

*In this paper, we address the problem of person re-identification, which refers to associating the persons captured from different cameras. We propose a simple yet effective human part-aligned representation for handling the body part misalignment problem. Our approach decomposes the human body into regions (parts) which are discriminative for person matching, accordingly computes the representations over the regions, and aggregates the similarities computed between the corresponding regions of a pair of probe and gallery images as the overall matching score. Our formulation, inspired by attention models, is a deep neural network modeling the three steps together, which is learnt through minimizing the triplet loss function without requiring body part labeling information. Unlike most existing deep learning algorithms that learn a global or spatial partition-based local representation, our approach performs human body partition, and thus is more robust to pose changes and various human spatial distributions in the person bounding box. Our approach shows state-of-the-art results over standard datasets, Market-*1501*, CUHK*03*, CUHK*01 *and VIPeR.* [1]

1. Introduction

Person re-identification is a problem of associating the persons captured from different cameras located at different physical sites. If the camera views are overlapped, the solution is trivial: the temporal information is reliable to solve the problem. In some real cases, the camera views are significantly disjoint and the temporal transition time between cameras varies greatly, making the temporal information not enough to solve the problem, and thus this problem becomes more challenging. Therefore, a lot of solutions exploiting various cues, such as appearance [12, 32, 23, 26], which is also the interest in this paper, have been developed.

Recently, deep neural networks have been becoming a dominate solution for the appearance representation. The straightforward way is to extract a global representation [33, 50, 6], using the deep network pretrained over ImageNet and optionally fine-tuned over the person re-identification dataset. Local representations are computed typically by partitioning the person bounding box into cells, e.g., dividing the images into horizontal stripes [56, 9, 44] or grids [23, 1], and extracting deep features over the cells. These solutions are based on the assumption that the human poses and the spatial distributions of the human body in the bounding box are similar. In real cases, for example, the bounding box is detected rather than manually labeled and thus the human may be at different positions, or the human poses are different, such an assumption does not hold. In other words, spatial partition is not well aligned with human body parts. Thus, person re-identification, even with subsequent complex matching techniques (e.g., [1, 23]) to eliminate the misalignment, is often not quite reliable. Figure 1 provides illustrative examples.

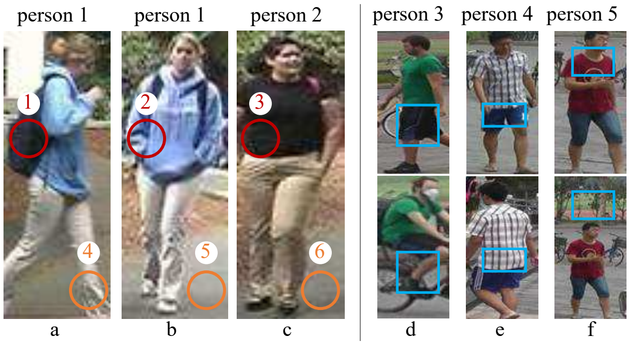

Figure 1. Illustrating the necessity of body part partition (best viewed in color). Using spatial partition without further processing, the regions (1) and (2), as well as (4) and (5), are not matched though they are from the same person; but the regions (1) and (3), as well as (5) and (6), which are from different persons, are matched. With body part decomposition, there is no such mismatch. More examples are shown in d, e, and f.

In this paper, we propose a part-aligned human representation, which addresses the above problem instead in the representation learning stage. The key idea is straightforward: detect the human body regions that are discrimina-

[*] Corresponding authors.

[1] This work was done when Liming Zhao was an intern at Microsoft Research.

tive for person matching, compute the representations over the parts, and then aggregate the similarities that are computed between the corresponding parts. Inspired by attention models [53], we present a deep neural network method, which jointly models body part extraction and representation computation, and learns model parameters through maximizing the re-identification quality in an end-to-end manner, without requiring the labeling information about human body parts. In contrast to spatial partition, our approach performs human body part partition, thus is more robust to human pose changes and various human spatial distributions in the bounding box. Empirical results demonstrate that our approach achieves competitive/superior performance over standard datasets: Market-1501, CUHK03, CUHK01 and VIPeR.

2. Related Work

There are two main issues in person re-identification: representation and matching. Various solutions, separately or jointly addressing the two issues, have been developed.

Separate solutions. Various hand-crafted representations have been developed, such as the ensemble of local features (ELF) [15], fisher vectors (LDFV) [29], local maximal occurrence representation (LOMO) [26], hierarchal Gaussian descriptor (GOG) [31], and so on. Most of the representations are designed with the goal of handling light variance, pose/view changes, and so on. Person attributes or salient patterns, such as female/male, wearing hat or not, have also been exploited to distinguish persons [40, 41, 61].

A lot of similarity/metric learning techniques [57, 58, 33, 27, 19] have been applied or designed to learn metrics, robust to light/view/pose changes, for person matching. The recent developments include soft and probabilistic patch matching for handling pose misalignment [4, 3, 36], similarity learning for dealing with probe and gallery images with different resolutions [24, 17], connection with transfer learning [34, 38], reranking inspired by the connection with image search [65, 13], partial person matching [66], human-in-the-loop learning [30, 46], and so on.

Deep learning-based solutions. The success of deep learning in image classification has been inspiring a lot of studies in person re-identification. The off-the-shelf CNN features, extracted from the model trained over ImageNet, without fine tuning, does not show the performance gain [33]. The promising direction is to learn the representation and the similarity jointly, except some works [51, 62] that do not learn the similarity but adopt the classification loss by regarding the images about one person as a category.

The network typically consists of two subnetworks: one for feature extraction and the other for matching. The feature extraction subnetwork could be simply (i) a shallow network [23] with one or two convolutional and max-pooling layers for feature extraction, or (ii) a deep network, e.g., VGGNet and its variants [39, 49] and GoogLeNet [42, 59], which are pretrained over ImageNet and fine-tuned for person re-identification. The feature representation can be (i) a global feature, e.g., the output of the fully-connected layer [6, 52], which does not explicitly model the spatial information, or (ii) a combination (e.g., concatenation [56, 9] or contextual fusion [44]) of the features over regions, e.g., horizontal stripes [56, 9, 44], or grid cells [23, 1], which are favorable for the later matching process to handle body part misalignment. Besides, the cross-dataset information [51] is also exploited to learn an effective representation.

The matching subnetwork can simply be a loss layer that penalizes the misalignment between learnt similarities and ground-truth similarities, e.g., pairwise loss [56, 44, 23, 1, 37], triplet loss and its variants [11, 9, 41, 45]. Besides using the off-the-shelf similarity function [56, 44, 9], e.g., cosine similarity or Euclidean distance, for comparing the feature representation, specific matching schemes are designed to eliminate the influence from body part misalignment. For instance, a matching subnetwork conducts convolution and max pooling operations, over the differences [1] or the concatenation [23, 59] of the representations over grid cells of a pair of person images, to handle the misalignment problem. The approach with so called single-image and cross-image representations [45] essentially combines the off-the-shelf distance and the matching network handling the misalignment. Instead of only matching the images over the final representation, the matching map in the intermediate features is used to guide the feature extraction in the later layers through a gated CNN [43].

Our approach. In this paper, we focus on the feature extraction part and introduce a human body part-aligned representation. Our approach is related to but different from the previous part-aligned approaches (e.g., part/pose detection [10, 54, 2, 63]), which need to train a part/pose segmentation or detection model from the labeled part mask/box or pose ground-truth and subsequently extract representations, where the processes are conducted separately. In contrast, our approach does not require those labeling information, but only uses the similarity information (a pair of person images are about the same person or different persons), to learn the part model for person matching. The learnt parts are different from the conventional human body parts, e.g., Pascal-Person-Parts [7], and are specifically for person matching, implying that our approach potentially performs better, which is verified by empirical comparisons with the algorithms based on the state-of-the-art part segmentation approach (deeplab [5]) and pose estimator (convolutional pose machine [47]).

Our human body part estimation scheme is inspired by the attention model that is successfully applied to many applications such as image captioning [53]. Compared to the work [28] that is based on attention models and LSTM, our approach is simple and easily implemented, and empirical results show that our approach performs better.

3. Our Approach

Person re-identification aims to find the images that are about the same identity with the probe image from a set of gallery images. It is often regarded as a ranking problem: given a probe image, the gallery images about the same identity are thought closer to the probe image than the gallery images about different identities.

The training data is typically given as follows. Given a set of images $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N\}$, we form the training set as a set of triplets, $\mathcal{T} = \{(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k)\}$, where $(\mathbf{I}_i, \mathbf{I}_j)$ is a positive pair of images that are about the same person and $(\mathbf{I}_i, \mathbf{I}_k)$ is a negative pair of images that are about different persons.

Our approach formulates the ranking problem using the triplet loss function,

$$\begin{aligned} &\ell_{\text{triplet}}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \\ =\ &[d(h(\mathbf{I}_i), h(\mathbf{I}_j)) - d(h(\mathbf{I}_i), h(\mathbf{I}_k)) + m]_+. \end{aligned} \tag{1}$$

Here $(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \in \mathcal{T}$. m is the margin by which the distance between a negative pair of images is greater than that between a positive pair of images. In our implementation, m is set to 0.2 similar to [35]. $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ is a Euclidean distance. $[z]_+ = \max(z, 0)$ is the hinge loss. $h(\mathbf{I})$ is a feature extraction network that extracts the representation of the image $\mathbf{I}$ and will be discussed in detail later. The whole loss function is as follows,

$$\mathcal{L}(h) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \in \mathcal{T}} \ell_{\text{triplet}}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k), \tag{2}$$

where $|\mathcal{T}|$ is the number of triplets in $\mathcal{T}$.

3.1. Part-Aligned Representation

The part-aligned representation extractor, is a deep neural network, consisting of a fully convolutional neural network (FCN) whose output is an image feature map, followed by a part net which detects part maps and outputs the part features extracted over the parts. Rather than partitioning the image box spatially to grid cells or horizontal stripes, our approach aims to partition the human body to aligned parts.

The part net, as illustrated in Figure 2, contains several branches. Each branch receives the image feature map from the FCN as the input, detects a discriminative region (part[2]), and extracts the feature over the detected region as the output. As we will see, the detected region usually lies in the human body region, which is as expected because these regions are informative for person matching. Thus, we call the net as a part net. Let a 3-dimensional tensor $\mathbf{T}$ represent the image feature maps computed from the FCN and thus $t(x, y, c)$ represent the cth response over the location (x, y). The part map detector estimates a 2-dimensional map $\mathbf{M}_k$, where $m_k(x, y)$ indicates the degree that the location (x, y) lies in the kth region, from the image feature map $\mathbf{T}$:

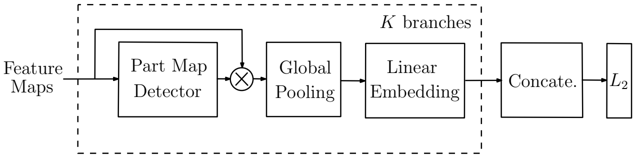

Figure 2. Illustrating the part net. It consists of K branches. Each branch takes the image feature map as the input and estimates a part map, which is used for weighting the image feature map followed by an average pooling operator. The part features from the K branches are concatenated as the final human representation.

$$\mathbf{M}_k = N_{\text{MapDetector}_\text{k}}(\mathbf{T}), \tag{3}$$

where$N_{\text{MapDetector}_\text{k}}(\cdot)$ is a region map detector implemented as a convolutional network.

The part feature map $\mathbf{T}_k$ for the kth region is computed through a weighting scheme,

$$t_k(x, y, c) = t(x, y, c) \times m_k(x, y), \tag{4}$$

followed by an average pooling operator, $\bar{\mathbf{f}}_k = \text{AvePooling}(\mathbf{T}_k)$, where $\bar{f}_k(c) = \text{Average}_{x,y}[t_k(x, y, c)]$. Then a linear dimension-reduction layer, implemented as a fully-connected layer, is performed to reduce $\bar{\mathbf{f}}_k$ to a d-dimensional feature vector $\mathbf{f}_k = \mathbf{W}_{FC_k}\bar{\mathbf{f}}_k$. Finally, we concatenate all the part features,

$$\mathbf{f} = [\mathbf{f}_1^\top\ \mathbf{f}_2^\top\ \dots\ \mathbf{f}_K^\top]^\top, \tag{5}$$

and perform an L_2 normalization, yielding the human representation $h(\mathbf{I})$.

3.2. Optimization

We learn the network parameters, denoted by θ, by minimizing the summation of triplet loss functions over triplets formulated in Equation 2. The gradient is computed as

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \in \mathcal{T}} \frac{\partial \ell_{\text{triplet}}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k)}{\partial \theta}. \tag{6}$$

We have[3]

$$\begin{aligned} &\frac{\partial \ell_{\text{triplet}}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k)}{\partial \theta} \\ =\ &\delta_{\ell_{\text{triplet}}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) > 0} \times 2[\frac{\partial h(\mathbf{I}_i)}{\partial \theta}(h(\mathbf{I}_k) - h(\mathbf{I}_j)) + \\ &\frac{\partial h(\mathbf{I}_j)}{\partial \theta}(h(\mathbf{I}_j) - h(\mathbf{I}_i)) + \frac{\partial h(\mathbf{I}_k)}{\partial \theta}(h(\mathbf{I}_i) - h(\mathbf{I}_k))]. \end{aligned}$$

[2] In this paper, we use the two terms, part and region, interchangeably for the same meaning.

[3] The gradient at the non-differentiable point is omitted like the common way to handle this case in deep learning.

Thus, we transform the gradient to the following form,

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{|\mathcal{T}|} \sum_{n=1}^{N} \frac{\partial h(\mathbf{I}_n)}{\partial \theta} \boldsymbol{\alpha}_n, \tag{7}$$

where $\boldsymbol{\alpha}_n$ is a weight vector depending on the current network parameters, and computed as follows,

$$\begin{aligned} \boldsymbol{\alpha}_n = 2[& \sum_{(\mathbf{I}_n,\mathbf{I}_j,\mathbf{I}_k)\in\mathcal{T}} \delta_{\ell_{\text{triplet}}(\mathbf{I}_n,\mathbf{I}_j,\mathbf{I}_k)>0}(h(\mathbf{I}_k) - h(\mathbf{I}_j)) + \\ & \sum_{(\mathbf{I}_i,\mathbf{I}_n,\mathbf{I}_k)\in\mathcal{T}} \delta_{\ell_{\text{triplet}}(\mathbf{I}_i,\mathbf{I}_n,\mathbf{I}_k)>0}(h(\mathbf{I}_n) - h(\mathbf{I}_i)) + \\ & \sum_{(\mathbf{I}_i,\mathbf{I}_j,\mathbf{I}_n)\in\mathcal{T}} \delta_{\ell_{\text{triplet}}(\mathbf{I}_i,\mathbf{I}_j,\mathbf{I}_n)>0}(h(\mathbf{I}_i) - h(\mathbf{I}_n))]. \end{aligned} \tag{8}$$

Equation 7 suggests that the gradient for the triplet loss is computed like that for the unary classification loss. Thus, in each iteration of SGD (stochastic gradient descent) we can draw a mini-batch of (M) samples rather than sample a subset of triplets: one pass of forward propagation to compute the representation $h(\mathbf{I}_n)$ of each sample, compute the weight $\boldsymbol{\alpha}_n$ over the mini-batch, compute the gradient $\frac{\partial h(\mathbf{I}_n)}{\theta}$, and finally aggregate the gradients over the mini-batch of samples. Directly drawing a set of triplets usually leads to that a larger number of (more than M) samples are contained and thus the computation is more expensive than our mini-batch sampling scheme.

3.3. Implementation details

Network architecture. We use a sub-network of the first version of GoogLeNet [42], from the image input to the output of *inception_4e*, followed by a 1×1 convolutional layer with the output of 512 channels, as the image feature map extraction network. Specifically, the person image box is resized to 160×80 as the input, and thus the size of the feature map of the feature map extraction network is 10×5 with 512 channels. For data preprocessing, we use the standard horizontal flips of the resized image. In the part net, the part estimator ($N_{\text{MapDetector}_k}$ in Equation 3) is simply a 1×1 convolutional layer followed by a nonlinear sigmoid layer. There are K part detectors, where K is determined by cross-validation and empirically studied in Section 4.3.

Network Training. We use the stochastic gradient descent algorithm to train the whole network based on Caffe [16]. The image feature map extraction part is initialized using the GoogLeNet model, pretrained over ImageNet. In each iteration, we sample a mini-batch of 400 images, e.g., there are on average 40 identities with each containing 10 images on Market-1501 and CUHK03. In total, there are about 1.4 million triplets in each iteration. From Equation 8, we see that only a subset of triplets, whose predicted similarity order is not consistent to the ground-truth order, i.e., $\ell_{\text{triplet}}(\mathbf{I}_n, \mathbf{I}_j, \mathbf{I}_k) > 0$, are counted for the weight (θ) update, and accordingly we use the number of counted triplets to replace $|\mathcal{T}|$ in Equation 7.

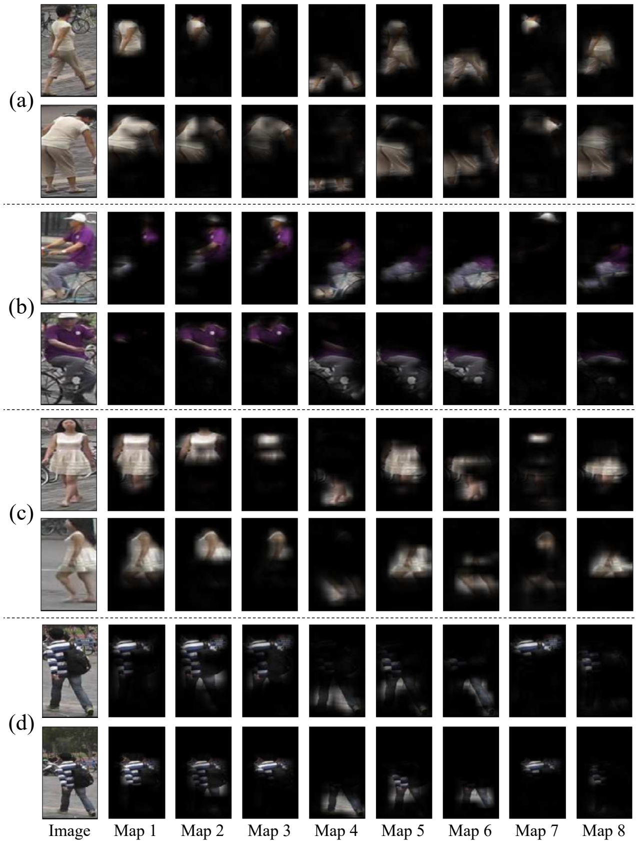

Figure 3. Examples of the part maps learnt by the part map estimator for test images (best viewed in color).

We adopt the initial learning rate, 0.01, and divide it by 5 every $20K$ iterations. The weight decay is 0.0002 and the momentum for gradient update is 0.9. Each model is trained for $50K$ iterations within around 12 hours on a K40 GPU. For testing, it takes on average 0.005 second on one GPU to extract the part-aligned representation.

3.4. Discussions

Body part partition and spatial partition. Spatial partition, e.g., grid or stride-based, may not be well aligned with human body parts, due to pose changes or various human spatial distributions in the human image box. Thus, matching techniques, e.g., through complex networks [1, 23, 59], have been developed to eliminate the misalignment problem. In contrast, our approach addresses this problem in the representation stage, with a simple Euclidean distance for person matching, which potentially makes existing fast similarity search algorithms easily applied, and thus the online search stage more efficient.

Figure 3 shows the examples about the parts our approach learns for the test images. It can be seen that the parts are generally well aligned for the pair of images about the same person: the parts almost describe the same human body regions, except that one or two parts in the pair of images describe different regions, e.g., the first part in Figure 3 (b). In particular, the alignment is also good for the examples of Figure 3 (c, d), where the person in the second image is spatially distributed very differently from the person in the first image: one is on the right in Figure 3 (c), and one is small and on the bottom in Figure 3 (d).

In addition, we empirically compare our approach with two spatial partition based methods: dividing the image box into 5 horizontal stripes or 5 × 5 girds to form region maps. We use the region maps to replace the part mask in our approach and then learn the spatial partition-based representation. The results shown in Table 1 demonstrate that the human body part partition method is more effective.

Table 1. The performance (%) of our approach and spatial partition based methods (stripe and grid) over Market-1501 and CUHK03.

Dataset	Method	rank-1	rank-5	rank-10	rank-20
Market-1501	ours	**81.0**	**92.0**	**94.7**	**96.4**
	stripe	74.1	89.0	92.3	95.1
	grid	73.4	88.2	91.8	94.4
CUHK03	ours	**85.4**	**97.6**	**99.4**	**99.9**
	stripe	81.4	97.1	99.3	99.7
	grid	78.2	96.7	99.2	99.8

Learnt body parts. We have several observations about the learnt parts. The head region is not included. This is because the face is not frontal and with low resolution and accordingly not reliable for differentiating different persons. The skin regions are often also not included except the arms located nearby the top body in Figure 3 (c) as the skin does not provide discriminant information, e.g., the leg skins in Figure 3 (c) are not included while the legs with trousers in Figure 3 (b) are included in Map4-6 .

From Figure 3, we can see that the first three maps, Map1 - Map3, are about the top clothing. There might be some redundancy. In the examples of Figure 3 (c,d), the first two masks are very close. In contrast, in the examples of Figure 3 (b), the masks are different, and are different regions of the top, though all are about the top clothing. In this sense, the first three masks act like a mixture model to describe the top clothing as the top parts are various due to pose and view variation. Similarly, Map4 and Map6 are both about the bottom.

Separate part segmentation. We conduct an experiment with separate part segmentation. We use the state-of-the-art part segmentation model [5] learnt from the PASCAL-Person-Part dataset [7] (6 part classes), to compute the mask for both training and test images. We modify our network by replacing the masks from the part net with the masks from the part segmentation model. In the training stage, we learn the modified network (the mask fixed) using the same setting with our approach.

Table 2. The performance of our approach, and separate part segmentation over Market-1501 and CUHK-03.

Dataset	Method	rank-1	rank-5	rank-10	rank-20
Market-1501	ours (6 parts)	**80.4**	**91.5**	**94.3**	**96.4**
	part seg. (6 parts)	61.2	80.3	86.9	91.0
CUHK03	ours (6 parts)	**85.1**	**97.6**	**98.2**	**99.4**
	part seg. (6 parts)	70.7	90.4	94.8	97.6

The results are shown in Table 2 and the performance is poor compared with our method. This is reasonable because the parts in our approach are learnt directly for person re-identification while the parts learnt from the PASCAL-Person-Part dataset might not be very good because it does not take consideration into the person re-identification problem. We also think that if the human part segmentation of the person re-identification training images is available, exploiting the segmentation as an extra supervision, e.g., the learnt part corresponds to a human part, or a sub-region of the human part, is helpful for learning the part net.

4. Experiments

4.1. Datasets

Market-1501. This dataset [64] is one of the largest benchmark datasets for person re-identification. There are six cameras: 5 high-resolution cameras, and one low-resolution camera. There are $32,668$ DPM-detected pedestrian image boxes of $1,501$ identities: 750 identifies are used for training and the remaining 751 for testing. There are $3,368$ query images and the large gallery (database) include $19,732$ images with $2,793$ distractors.

CUHK03**.** This dataset [23] consists of $13,164$ images of $1,360$ persons, captured by six cameras. Each identity only appears in two disjoint camera views, and there are on average 4.8 images in each view. We use the provided training/test splits [23] on the labeled data set. For each test identity, two images are randomly sampled as the probe and gallery images, respectively, and the average performance over 20 trials is reported as the final result.

CUHK01**.** This dataset [22] contains 971 identities captured from two camera views in the same campus with CUHK03. Each person has two images, each from one camera view. Following the setup [1], we report the results of two different settings: 100 identifies for testing, and 486 identities for testing.

VIPeR. This dataset [14] contains two views of 632 persons. Each pair of images about one person are captured by different cameras with large viewpoint changes and various illumination conditions. The 632 person images are divided into two halves, 316 for training and 316 for testing.

4.2. Evaluation Metrics

We adopt the widely-used evaluation protocol [23, 1]. In the matching process, we calculate the similarities between each query and all the gallery images, and then return the

Table 3. The *validation* performance with different numbers (K) of parts over CUHK03. The model is trained over a random half of the training data, and the performance is reported over the remaining half (as the validation set). The best results are in bold.

#parts	rank-1	rank-5	rank-10	rank-20
1	77.7	95.6	98.4	**99.7**
2	80.4	96.7	98.4	99.4
4	82.0	96.7	**98.8**	**99.7**
8	**83.8**	96.9	98.3	**99.7**
12	83.6	**97.3**	**98.8**	99.6

Table 4. The performances of our approach and human segmentation over Market-1501 and CUHK03.

Dataset	Method	rank-1	rank-5	rank-10	mAP
Market-1501	ours	**81.0**	**92.0**	**94.7**	**63.4**
	human seg.	74.2	90.0	93.8	58.9
CUHK03	ours	**85.4**	**97.6**	**99.4**	**90.9**
	human seg.	82.7	95.9	97.9	88.6

ranked list according to the similarities. All the experiments are under the single query setting. The performances are evaluated by the cumulated matching characteristics (CMC) curves, which is an estimate of the expectation of finding the correct match in the top n matches. We also report the mean average precision (mAP) score [64] over Market-1501.

4.3. Empirical Analysis

The number of parts. We empirically study how the number of parts affects the performance. We conduct an experiment over CUHK03: randomly partition the training dataset into two parts, one for model learning and the remaining for validation. The performances for various numbers of parts, $K = 1, 2, 4, 8, 12$, are given in Table 3. It can be seen that (i) more parts for the rank-1 score lead to better scores till 8 parts and then the scores become stable, and (ii) the scores of different number of parts at positions 5, 10, and 20 are close except the score of 1 part at position 5. Thus, in our experiments, we choose $K = 8$ in the part net for all the four datasets. It is possible that in other datasets the optimal K obtained through validation is different.

Human segmentation and body part segmentation. The benefit from the body part segmentation lies in two points: (i) remove the background and (ii) part alignment. We compare our approach and the approach with human segmentation that is implemented as our approach with 1 part and is able to remove the background. The comparison shown from Table 4 over Market-1501 and CUHK03 shows that body part segmentation performs superiorly in general. The results imply that body part segmentation is beneficial.

Comparison with non-human/part-segmentation. We compare the performances of two baseline networks without segmentation, which are modified from our network: (i) replace the part net with a fully-connected layer outputting the feature vector with the same dimension (512-d) and (ii) replace the part net with an global average-pooling layer which also produces a 512-d feature vector.

Table 5. The performances of our approach, two baseline networks without segmentation, modified by replacing the part net in our network with a fully-connected (FC) layer and an average pooling (pooling) layer over Market-1501 and CUHK03.

Dataset	Method	rank-1	rank-5	rank-10	mAP
Market-1501	ours	**81.0**	**92.0**	**94.7**	**63.4**
	FC	75.9	89.3	92.9	54.3
	pooling	75.9	89.0	92.2	55.6
CUHK03	ours	**85.4**	**97.6**	**99.4**	**90.9**
	FC	80.3	95.5	98.6	87.3
	pooling	82.4	96.8	99.0	88.9

The fully-connected layer followed by the last convolutional layer in (i) has some capability to differentiate different spatial regions to some degree through the linear weights, which are however the same for all images, yielding limited ability of differentiation. The average-pooling method in (ii) ignores the spatial information, though it is robust to the translations. In contrast, our approach is also able to differentiate body regions and the differentiation is adaptive to each input image for translation/pose invariance.

The comparison over two datasets, Market-1501 and CUHK03, is given in Table 5. It can be seen that our approach outperforms these two baseline methods, which indicates that the part segmentation is capable of avoiding the mismatch due to part misalignment in spatial partition and improving the performance.

Image feature map extraction networks. We show that the part net can boost the performance for various feature map extraction FCNs. We report two extra results with using AlexNet [21] and VGGNet [39] as well as the result using GoogLeNet [42]. For AlexNet and VGGNet, we remove the fully connected layers and use all the remaining convolutional layers as the feature map extraction network, and the training settings are the same as provided in Section 3.3. The results are depicted in Figure 4. It can be seen that our approach consistently gets the performance gain for AlexNet, VGGNet and GoogLeNet. In particular, the gains with AlexNet and VGGNet are more significant: compared with the baseline method with FC, the gains are 6.8, 6.4, and 5.1 for AlexNet, VGGNet and GoogLeNet, respectively, and compared with the baseline method with pooling, the gains are 5.9, 4.4, and 3.0, respectively.

Comparison with other attention models. The part map detector is inspired by the spatial attention model. It is slightly different from the standard attention model: using sigmoid to replace softmax, which brings more than 2% gain for rank-1 scores. The comparative attention network (CAN) approach [28] is also based on the attention model and adopts LSTM to help learn part maps. It is not easy for us to have a good implementation for CAN. Thus, we report the results with AlexNet, which CAN is based on, as our base network. The comparison is given in Table 6. We can see that the overall performance of our approach is better except on the CUHK01 dataset for 100 test IDs.

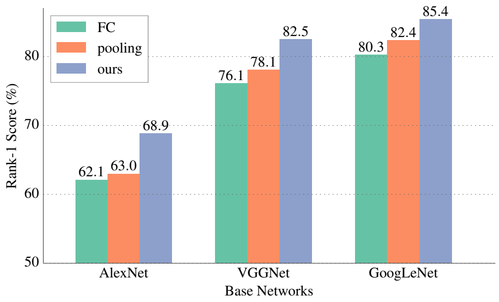

Figure 4. The performance of our approach and the two baseline networks (FC and Pooling) with different feature map extraction networks over CUHK03. Our approach consistently boosts the performance for all the three networks (best viewed in color).

Table 6. Compared with softmax over spatial responses and CAN [28]. All are based on AlexNet. Larger is better.

		rank-1	rank-5	rank-10	rank-20
CUHK03 (labeled)	CAN	65.65	91.28	**96.29**	98.17
	Ours	**68.90**	**91.40**	95.25	**98.3**
	Softmax	65.14	90.64	95.43	97.79
CUHK03 (detected)	CAN	63.05	82.94	88.17	93.29
	Ours	**65.64**	**89.50**	**93.93**	**96.71**
	Softmax	64.36	89.50	94.71	97.43
CUHK01-100	CAN	**81.04**	**96.89**	**99.67**	**100**
	Ours	79.25	94.00	96.37	98.75
	Softmax	74.64	91.27	94.55	97.27
Market	CAN	48.24	mAP = 24.43		
	Ours	**64.22**	mAP = **41.80**		
	Softmax	62.23	mAP = 41.01		

4.4. Comparison with State-of-the-Arts

Market-1501**.** We compare our method with recent state-of-the-arts, which are separated into four categories: feature extraction (F), metric learning (M), deeply learnt feature representation (DF), deep learning with matching subnetwork (DMN). The results in Table 7 are obtained under the single query setting.

The competitive algorithm, pose-invariant embedding (PIE) [63] extracts part-aligned representation, based on state-of-the-art pose estimator CPM [47] for part detection that is different from ours. PIE uses ResNet-50 which is more powerful than GoogLeNet our approach uses. We observe that our approach performs the best and outperforms PIE: 2.35 gain for rank-1 and 9.5 gain for mAP compared to PIE w/o using KISSME, and 1.67 for rank-1 and 7.4 gain for mAP compared to PIE w/ using KISSME.

CUHK03**.** There are two versions of person boxes: one is manually labeled and the other one is detected with a pedestrian detector. We report the results for both versions and all the previous results on CUHK03 are reported on the labeled version. The results are given in Table 8 for manually-labeled boxes and in Table 9 for detected boxes.

Our approach performs the best on both versions. On the one hand, the improvement over the detected boxes is more significant than that over the manually-labeled boxes. This is because the person body parts in the manually-labeled boxes are spatially distributed more similarly. On the other hand, the performance of our approach over the manually-labeled boxes are better than that over the detected-labeled boxes. This means that the person position in the box (manually-labeled boxes are often better) influences the part extraction quality, which suggests that it is a necessity to learn a more robust part extractor with more supervision information or over a larger dataset.

Table 7. Performance comparison of state-of-the-art methods on the recently released challenging dataset, Market-1501. The methods are separated into four categories: feature extraction (F), metric learning (M), deeply learnt feature representation (DF), deep learning with matching subnetwork (DMN).

Method		rank-1	rank-5	rank-10	mAP
F	LOMO [26] (CVPR15)	26.1	-	-	7.8
	BoW [64] (ICCV15)	35.8	52.4	60.3	14.8
M	KISSME [20] (CVPR12)	44.4	63.9	72.2	20.8
	WARCA [18] (ECCV16)	45.2	68.2	76.0	-
	TMA [30] (ECCV16)	47.9	-	-	22.3
	SCSP [3] (CVPR16)	51.9	72.0	79.0	26.4
	DNS [57] (CVPR16)	55.4	-	-	29.9
DMN	PersonNet [49] (ArXiv16)	37.2	-	-	18.6
	Gated S-CNN [43] (ECCV16)	65.9	-	-	39.6
DF	PIE [63]	78.65	90.26	93.59	53.87
	PIE [63] + KISSME [20] (Arxiv 2016)	79.33	90.76	94.41	55.95
	SSDAL [41] (ECCV16)	39.4	-	-	19.6
	Our Method	**81.0**	**92.0**	**94.7**	**63.4**

Table 8. Performance comparison on CUHK03 for manually labeled human boxes.

Method		rank-1	rank-5	rank-10	rank-20
F	BoW [64] (ICCV15)	18.9	36.2	46.8	-
	LOMO [26] (CVPR15)	52.2	82.2	92.1	96.3
	GOG [31] (CVPR16)	67.3	91.0	96.0	-
M	KISSME [20] (CVPR12)	47.9	69.3	78.9	87.0
	SSSVM [58] (CVPR16)	57.0	84.8	92.5	96.4
	DNS [57] (CVPR16)	58.9	85.6	92.5	96.3
	Ensembles [33] (CVPR15)	62.1	89.1	94.3	97.8
	WARCA [18] (ECCV16)	78.4	94.6	97.5	99.1
DMN	DeepReID [23] (CVPR14)	20.7	51.3	68.7	83.1
	IDLA [1] (CVPR15)	54.7	86.4	93.9	98.1
	PersonNet [49] (ArXiv16)	64.8	89.4	94.9	98.2
	DCSL [59] (IJCAI16)	80.2	**97.7**	99.2	99.8
DF	Deep Metric [37] (ECCV16)	61.3	88.5	96.0	99.0
	Our Method	**85.4**	**97.6**	**99.4**	**99.9**

Compared with the competitive method DCSL [59] which is also based on the GoogLeNet, the overall performance of our approach, as shown in Table 8, is better on CUHK03 except that the rank-5 score of DCSL is slightly better by 0.1%. This is an evidence demonstrating the powerfulness of the part-aligned representation though DCSL adopts the strong matching subnetwork to improve the matching quality. Compared with the second best method, PIE, on the detected case as shown in Table 9, our approach achieves 4.5 gain at rank-1.

CUHK01**.** There are two evaluation settings [1]: 100 test IDs, and 486 test IDs. Since there are a small number (485) of training identities for the case of 486 test IDs, as done in [1, 6, 59], we fine-tune the model, which is learnt from the CUHK03 training set, over the 485 training identities: the rank-1 score from the model learnt from CUHK03 is 44.59% and it becomes 72.3% with the fine-tuned model.

The results are reported in Table 10 and Table 11, re-

Table 9. Performance comparison on CUHK03 for detected boxes.

	Method	rank-1	rank-5	rank-10	rank-20
F	LOMO [26] (CVPR15)	46.3	78.9	88.6	94.3
	GOG [31] (CVPR16)	65.5	88.4	93.7	-
M	LMNN [48] (NIPS05)	6.3	17.5	28.2	45.0
	KISSME [20] (CVPR12)	11.7	33.9	48.2	65.0
	SSSVM [58] (CVPR16)	51.2	81.5	89.9	95.0
	DNS [57] (CVPR16)	53.7	83.1	93.0	94.8
DMN	DeepReID [23] (CVPR14)	19.9	50.0	64.0	78.5
	IDLA [1] (CVPR15)	45.0	76.0	83.5	93.2
	SIR-CIR [45] (CVPR16)	52.2	85.0	92.0	97.0
DF	PIE [63] + KISSME [20] (Arxiv 2016)	67.10	92.20	96.60	98.10
	Deep Metric [37] (ECCV16)	52.1	84.0	92.0	96.8
	Our Method	**81.6**	**97.3**	**98.4**	**99.5**

Table 10. Performance comparison on CUHK01 for 100 test IDs.

	Method	rank-1	rank-5	rank-10	rank-20
DMN	DeepReID [23] (CVPR14)	27.9	58.2	73.5	86.3
	IDLA [1] (CVPR15)	65.0	88.7	93.1	97.2
	Deep Ranking ([6] (TIP16))	50.4	70.0	84.8	92.0
	PersonNet [49] (ArXiv16)	71.1	90.1	95.0	98.1
	SIR-CIR [45] (CVPR16)	71.8	91.6	96.0	98.0
	DCSL [59] (IJCAI16)	**89.6**	97.8	98.9	99.7
DF	Deep Metric [37] (ECCV16)	69.4	90.8	96.0	-
	Our Method	**88.5**	**98.4**	**99.6**	**99.9**

Table 11. Performance comparison on CUHK01 for 486 test IDs.

	Method	rank-1	rank-5	rank-10	rank-20
F	Semantic [38] (CVPR15)	31.5	52.5	65.8	77.6
	MirrorRep [8] (IJCAI15)	40.4	64.6	75.3	84.1
	LOMO [26] (CVPR15)	49.2	75.7	84.2	90.8
	GOG [31] (CVPR16)	57.8	79.1	86.2	92.1
M	LMNN [48] (NIPS05)	13.5	31.3	42.3	54.1
	SalMatch [60] (ICCV13)	28.5	45.9	55.7	68.0
	DNS [57] (CVPR16)	65.0	85.0	89.9	94.4
	WARCA [18] (ECCV16)	65.6	85.3	90.5	95.0
	SSSVM [58] (CVPR16)	66.0	89.1	92.8	96.5
DMN	IDLA [1] (CVPR15)	47.5	71.6	80.3	87.5
	Deep Ranking [6] (TIP16)	50.4	70.0	84.8	92.0
	DCSL [59] (IJCAI16)	**76.5**	**94.2**	**97.5**	-
DF	TCP-CNN [9] (CVPR16)	53.7	84.3	91.0	96.3
	Our Method	72.3	91.0	94.9	97.2
	Our Method + remove pool3	**75.0**	**93.5**	**95.7**	**97.7**

spectively. Our approach performs the best among the algorithms w/o using matching subnetwork. Compared to the competitive algorithm DCSL [59] that uses matching subnetwork, we can see that for 100 test IDs, our approach performs better in general except a slightly low rank-1 score and that for 486 test IDs our initial approach performs worse and with a simple trick, removing one pooling layer to double the feature map size, the performance is much closer. One notable point is that our approach is advantageous in scaling up to large datasets.

VIPeR. The dataset is relatively small and the training images are not enough for training. We fine-tune the model learnt from CUHK03 following [43, 1]. The results are presented in Table 12. Our approach outperforms other deep learning-based approaches except PIE [63] with complicated schemes while performs poorer than the best-performed feature extraction approach GOG [31] and metric learning method SCSP [3]. In comparison with PIE [63], our approach performs better than PIE with data augmentation Mirror [8] and metric learning MFA [55] and lower than PIE with a more complicated fusion scheme, which our approach might benefit from. In general, the results suggest that like in other tasks, e.g., classification, training deep neural networks from a small data is still an open and challenging problem.

Table 12. Results on a relatively small dataset, VIPeR.

	Method	rank-1	rank-5	rank-10	rank-20
F	ELF [15] (ECCV 2008)	12.0	44.0	47.0	61.0
	BoW [64] (ICCV15)	21.7	42.0	50.0	60.9
	LOMO [26] (CVPR15)	40.0	68.1	80.5	91.1
	Semantic [38] (CVPR15)	41.6	71.9	86.2	95.1
	MirrorRep [8] (IJCAI15)	43.0	75.8	87.3	94.8
	GOG [31] (CVPR16)	**49.7**	**79.7**	**88.7**	**94.5**
M	LMNN [48] (NIPS05)	11.2	32.3	44.8	59.3
	KISSME [20] (CVPR12)	19.6	47.5	62.2	77.0
	LADF [25] (CVPR13)	30.0	64.7	79.0	91.3
	WARCA [18] (ECCV16)	37.5	70.8	82.0	92.0
	DNS [57] (CVPR16)	42.3	71.5	82.9	92.1
	SSSVM [58] (CVPR16)	42.7	-	84.3	91.9
	TMA [30] (ECCV16)	43.8	-	83.9	91.5
	SCSP [3] (CVPR16)	**53.5**	**82.6**	**91.5**	**96.7**
DMN	IDLA [1] (CVPR15)	34.8	63.6	75.6	84.5
	Gated S-CNN [43] (ECCV16)	37.8	66.9	77.4	-
	SSDAL [41] (ECCV16)	37.9	65.5	75.6	88.4
	SIR-CIR [45] (CVPR16)	35.8	67.4	83.5	-
	Deep Ranking [6] (TIP16)	38.4	69.2	81.3	90.4
	DCSL [59] (IJCAI16)	**44.6**	**73.4**	**82.6**	**91.9**
DF	PIE [63] + Mirror [8] + MFA [55] (Arxiv 2016)	43.3	69.4	80.4	90.0
	Fusion [63] + MFA [55] (Arxiv 2016)	**54.5**	**84.4**	**92.2**	**96.9**
	Deep Metric [37] (ECCV16)	40.9	67.5	79.8	-
	TCP-CNN [9] (CVPR16)	47.8	74.7	84.8	91.1
	Our Method	48.7	74.7	85.1	93.0

Summary. The overall performance of our approach is the best in the category of deeply-learnt feature representation (DF) and better than non-deep learning algorithms except in the small dataset VIPeR. In comparison to the category of deep learning with matching subnetwork (DMN), our approach in general is good, and performs worse than DCSL in CUHK01 with 486 test IDs. It is reasonable as matching network is more complicated than the simple Euclidean distance in our approach. One notable advantage is that our approach is efficient in online matching and cheap in storage, while DCSL stores large feature maps of gallery images for online similarity computation, resulting in larger storage cost and higher online computation cost.

5. Conclusions

In this paper, we present a novel part-aligned representation approach to handle the body misalignment problem. Our formulation follows the idea of attention models and is in a deep neural network form, which is learnt only from person similarities without the supervision information about the human parts. Our approach aims to partition the human body instead of the human image box into grids or strips, and thus is more robust to pose changes and different human spatial distributions in the human image box and thus the matching is more reliable. Our approach learns more useful body parts for person re-identification than separate body part detection. [4]

[4] **Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China under Grant U1509206 and Grant 61472353, in part by the National Basic Research Program of China under Grant Grant 2015CB352302, and partially funded by the MOE-Microsoft Key Laboratory of Visual Perception, Zhejiang University.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, pages 435–440, 2010.

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, June 2016.

[4] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, pages 1565–1573, 2015.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[6] S.-Z. Chen, C.-C. Guo, and J.-H. Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Trans. Image Processing*, 25(5):2353–2367, 2016.

[7] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014.

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

[9] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, June 2016.

[10] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, pages 1–11, 2011.

[11] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, June 2010.

[13] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV*, December 2015.

[14] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on PETS*, 2007.

[15] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014.

[17] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, June 2015.

[18] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, 2016.

[19] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised ℓ_1 graph learning. In *ECCV*, pages 178–195, 2016.

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

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[22] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

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

[24] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, December 2015.

[25] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610–3617, 2013.

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

[27] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[28] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *CoRR*, abs/1606.04404, 2016.

[29] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, pages 413–422, 2012.

[30] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, pages 858–877, 2016.

[31] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, June 2016.

[32] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[33] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, June 2015.

[34] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, June 2016.

[35] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[36] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015.

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

[38] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, June 2015.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, December 2015.

[41] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, pages 475–491, 2016.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

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

[44] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, pages 135–153, 2016.

[45] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, June 2016.

[46] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *ECCV*, pages 405–422, 2016.

[47] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016.

[48] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005.

[49] L. Wu, C. Shen, and A. van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. *CoRR*, abs/1601.07255, 2016.

[50] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, pages 1–8, 2016.

[51] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, June 2016.

[52] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *CoRR*, abs/1604.01850, 2016.

[53] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[54] Y. Xu, L. Lin, W. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, pages 3152–3159, 2013.

[55] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, 2007.

[56] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICLR*, 2014.

[57] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[58] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, June 2016.

[59] Y. Zhang, X. Li, L. Zhao, and Z. Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*, pages 3545–3551, 2016.

[60] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013.

[61] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, June 2014.

[62] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[63] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *CoRR*, abs/1701.07732, 2017.

[64] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[65] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, June 2015.

[66] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, December 2015.