Part-Aligned Bilinear Representations for Person Re-identification

Yumin Suh[1], Jingdong Wang[2], Siyu Tang[3,4], Tao Mei[5], and Kyoung Mu Lee[1]

[1] ASRI, Seoul National University, Seoul, Korea
[2] Microsoft Research Asia, Beijing, China
[3] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[4] University of Tübingen, Tübingen, Germany
[5] JD AI Research, Beijing, China
{n12345,kyoungmu}@snu.ac.kr, jingdw@microsoft.com,
stang@tuebingen.mpg.de, tmei@jd.com

Abstract. Comparing the appearance of corresponding body parts is essential for person re-identification. As body parts are frequently misaligned between the detected human boxes, an image representation that can handle this misalignment is required. In this paper, we propose a network that learns a part-aligned representation for person re-identification. Our model consists of a two-stream network, which generates appearance and body part feature maps respectively, and a bilinear-pooling layer that fuses two feature maps to an image descriptor. We show that it results in a compact descriptor, where the image matching similarity is equivalent to an aggregation of the local appearance similarities of the corresponding body parts. Since the image similarity does not depend on the relative positions of parts, our approach significantly reduces the part misalignment problem. Training the network does not require any part annotation on the person re-identification dataset. Instead, we simply initialize the part sub-stream using a pre-trained sub-network of an existing pose estimation network and train the whole network to minimize the re-identification loss. We validate the effectiveness of our approach by demonstrating its superiority over the state-of-the-art methods on the standard benchmark datasets including Market-1501, CUHK03, CUHK01 and DukeMTMC, and standard video dataset MARS.

Keywords: Person re-identification · Part alignment · Bilinear pooling

1 Introduction

The goal of person re-identification is to identify the same person across videos captured from different cameras. It is a fundamental visual recognition problem in video surveillance with various applications [56]. It is challenging because the camera views are usually disjoint, the temporal transition time between cameras varies considerably, and the lighting conditions/person poses differ across cameras in real-world scenarios.

Body part misalignment (i.e., the problem that body parts are spatially misaligned across person images) is one of the key challenges in person re-identification. Figure 1 shows some examples. This problem causes conventional strip/grid-based representations [25, 1, 72, 70, 10, 59] to be unreliable as they implicitly assume that every person

Fig. 1. (a, b) As a person appears in different poses/viewpoints in different cameras, and (c) human detections are imperfect, the corresponding body parts are usually not spatially aligned across the human detections, causing person re-identification to be challenging.

appears in a similar pose within a tightly surrounded bounding box. Thus, a body part-aligned representation, which can ease the representation comparison and avoid the need for complex comparison techniques, should be designed.

To resolve this problem, recent approaches have attempted to localize body parts explicitly and combine the representations over them [50, 79, 75, 23, 76]. For example, the body parts are represented by the pre-defined (or refined [50]) bounding boxes estimated from the state-of-the-art pose estimators [79, 50, 4, 75]. This scheme requires highly-accurate pose estimation. Unfortunately, state-of-the-art pose estimation solutions are still not perfect. Also, these schemes are bounding box-based and lack fine-grained part localization within the boxes. To mitigate the problems, we propose to encode human poses by feature maps rather than by bounding boxes. Recently, Zhao et al. [76] represented body parts through confidence maps, which are estimated using attention techniques. The method has a lack of guidance on body part locations during the training, thereby failing to attend to certain body regions consistently.

In this paper, we propose a part-aligned representation for person re-identification. Our approach learns to represent the human poses as part maps and combine them directly with the appearance maps to compute part-aligned representations. More precisely, our model consists of a two-stream network and an aggregation module. 1) Each stream separately generates appearance and body part maps. 2) The aggregation module first generates the part-aligned feature maps by computing the bilinear mapping of the appearance and part descriptors at each location, and then spatially averages the local part-aligned descriptors. The resulting image matching similarity is equivalent to an aggregation of the local appearance similarities of the corresponding body parts. Since it does not depend on the relative positions of parts, the misalignment problem is reduced.

Training the network does not require any body part annotations on the person re-identification dataset. Instead, we simply initialize the part map generation stream using the pre-trained weights, which are trained from a standard pose estimation dataset. Surprisingly, although our approach only optimizes the re-identification loss function, the resulting two-stream network successfully separates appearance and part information into each stream, thereby generating the appearance and part maps from each of them, respectively. In particular, the part maps adapt from the original form to further differentiate informative body parts for person re-identification. Through extensive experiments, we verify that our approach consistently improves the accuracy of the baseline and achieves competitive/superior performance over standard image datasets, Market-1501, CUHK03, CUHK01 and DukeMTMC, and one standard video dataset, MARS.

2 Related Work

The early solutions of person re-identification mainly relied on hand-crafted features [36, 27, 18, 39], metric learning techniques [71, 73, 42, 28, 22, 26, 20], and probabilistic patch matching algorithms [6, 5, 48] to handle resolution/light/view/pose changes. Recently, attributes [51, 52, 77], transfer learning [43, 49], re-ranking [81, 15], partial person matching [83], and human-in-the-loop learning [38, 61], have also been studied. More can be found in the survey [82]. In the following, we review recent spatial-partition-based and part-aligned representations, matching techniques, and some works using bilinear pooling.

Regular spatial-partition based representations. The approaches in this stream of research represent an image as a combination of local descriptors, where each local descriptor represents a spatial partition such as grid cell [25, 1, 72] and horizontal stripe [70, 10, 59]. They work well under a strict assumption that the location of each body part is consistent across images. This assumption is often violated under realistic conditions, thereby causing the methods to fail. An extreme case is that no spatial partition is used and a global representation is computed over the whole image [42, 65, 78, 64, 7, 66].

Body part-aligned representations. Body part and pose detection results have been exploited for person re-identification to handle the body part misalignment problem [12, 69, 3, 13, 63, 11]. Recently, these ideas have been re-studied using deep learning techniques. Most approaches [79, 50, 75] represent an image as a combination of body part descriptors, where a dozen of pre-defined body parts are detected using the off-the-shelf pose estimator (possibly an additional RoI refinement step). They usually crop bounding boxes around the detected body parts and compute the representations over the cropped boxes. In contrast, we propose part-map-based representations, which is different from the previously used box-based representations [79, 50, 75].

Tang et al [56] also introduced part maps for person re-identification to solve the multi-people tracking problem. They used part maps to augment appearances as another feature, rather than to generate part-aligned representations, which is different from our method. Some works [34, 76] proposed the use of attention maps, which are expected to attend to informative body parts. They often fail to produce reliable attentions as the attention maps are estimated from the appearance maps; guidance from body part locations is lacking, resulting in a limited performance.

Matching. The simple similarity functions [70, 59, 10], e.g., cosine similarity or Euclidean distance, have been adapted, for part-aligned representations, such as our approach, or under an assumption that the representations are body part/pose aligned. Various schemes [60, 1, 25, 72] were designed to eliminate the influence from body part misalignment for spatial partition-based representations. For instance, a matching sub-network was proposed to conduct convolution and max-pooling operations, over the differences [1] or the concatenation [25, 72] of grid-based representation of a pair of person images. Varior et al. [58] proposed the use of matching maps in the intermediate features to guide feature extraction in the later layers through a gated CNN.

Bilinear pooling. Bilinear pooling is a scheme to aggregate two different types of feature maps by using the outer product at each location and spatial pooling them to obtain a global descriptor. This strategy has been widely adopted in fine-grained recog-

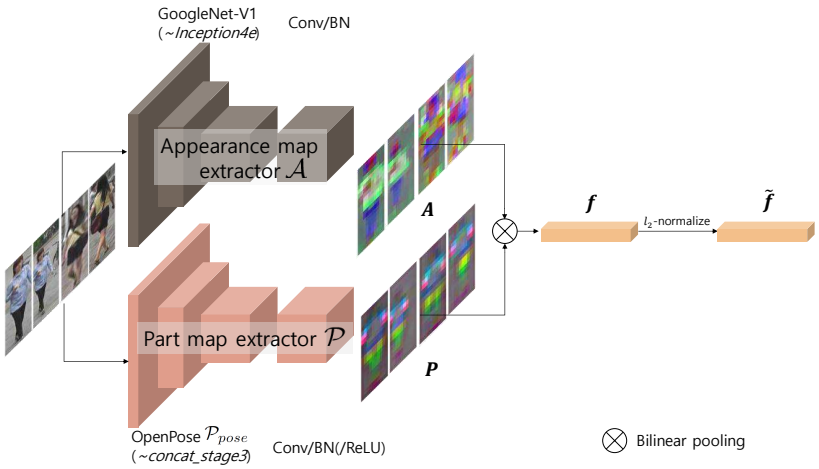


Fig. 2. Overview of the proposed model. The model consists of a two-stream network and an aggregator (bilinear pooling). For a given image $\mathbf{I}$, the appearance and part map extractors, $\mathcal{A}$ and $\mathcal{P}$, generate the appearance and part maps, $\mathbf{A}$ and $\mathbf{P}$, respectively. The aggregator performs bilinear pooling over $\mathbf{A}$ and $\mathbf{P}$ and generates a feature vector $\mathbf{f}$. Finally, the feature vector is l_2-normalized, resulting in a final part-aligned representation $\tilde{\mathbf{f}}$. Conv and BN denote the convolution and batch normalization layers, respectively.

nition [30, 14, 21] and showed promising performance. For person re-identification, Ustinova et al. [57] adopted a bilinear pooling to aggregate two different appearance maps; this method does not generate part-aligned representations and leads to poor performance. Our approach uses a bilinear pooling to aggregate appearance and part maps to compute part-aligned representations.

3 Our Approach

The proposed model consists of a two-stream network and an aggregation module. It receives an image $\mathbf{I}$ as an input and outputs a part-aligned feature representation $\tilde{\mathbf{f}}$ as illustrated in Figure 2. The two-stream network contains two separate sub-networks, the appearance map extractor $\mathcal{A}$ and the part map extractor $\mathcal{P}$, which extract the appearance map $\mathbf{A}$ and part map $\mathbf{P}$, respectively. The two types of maps are aggregated through bilinear pooling to generate the part-aligned feature $\mathbf{f}$, which is subsequently normalized to generate the final feature vector $\tilde{\mathbf{f}}$.

3.1 Two-Stream Network

Appearance map extractor. We feed an input image $\mathbf{I}$ into the appearance map extractor $\mathcal{A}$, thereby outputting the appearance map $\mathbf{A}$:

$$\mathbf{A} = \mathcal{A}(\mathbf{I}). \tag{1}$$

$\mathbf{A} \in \mathbb{R}^{h \times w \times c_A}$ is a feature map of size $h \times w$, where each location is described by c_A-dimensional local appearance descriptor. We use the sub-network of GoogLeNet [55] to form and initialize $\mathcal{A}$.

Part map extractor. The part map extractor $\mathcal{P}$ receives an input image $\mathbf{I}$ and outputs the part map $\mathbf{P}$:

$$\mathbf{P} = \mathcal{P}(\mathbf{I}). \tag{2}$$

$\mathbf{P} \in \mathbb{R}^{h \times w \times c_P}$ is a feature map of size $h \times w$, where each location is described by a c_P-dimensional local part descriptor. Considering the rapid progress in pose estimation, we use the sub-network of the pose estimation network, OpenPose [4], to form and initialize $\mathcal{P}$. We denote the sub-network of the OpenPose as $\mathcal{P}_{pose}$.

3.2 Bilinear Pooling

Let $\mathbf{a}_{xy}$ be the appearance descriptor at the position (x, y) from the appearance map $\mathbf{A}$, and $\mathbf{p}_{xy}$ be the part descriptor at the position (x, y) from the part map $\mathbf{P}$. We perform bilinear pooling over $\mathbf{A}$ and $\mathbf{P}$ to compute the part-aligned representation $\mathbf{f}$. There are two steps, bilinear transformation and spatial global pooling, which are mathematically given as follows:

$$\mathbf{f} = \text{pooling}_{xy}\{\mathbf{f}_{xy}\} = \frac{1}{S}\sum_{xy}\mathbf{f}_{xy}, \qquad \mathbf{f}_{xy} = \text{vec}(\mathbf{a}_{xy} \otimes \mathbf{p}_{xy}), \tag{3}$$

where S is the spatial size. The pooling operation we use here is average-pooling. $\text{vec}(.)$ transforms a matrix to a vector, and $\otimes$ represents the outer product of two vectors, with the output being a matrix. The part-aligned feature $\mathbf{f}$ is then normalized to generate the final feature vector $\tilde{\mathbf{f}}$ as follows:

$$\tilde{\mathbf{f}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2}. \tag{4}$$

Considering the normalization, we denote the normalized part-aligned representation as $\tilde{\mathbf{f}}_{xy} = \text{vec}(\tilde{\mathbf{a}}_{xy} \otimes \tilde{\mathbf{p}}_{xy})$, where $\tilde{\mathbf{a}}_{xy} = \frac{\mathbf{a}_{xy}}{\sqrt{\|\mathbf{f}\|_2}}$ and $\tilde{\mathbf{p}}_{xy} = \frac{\mathbf{p}_{xy}}{\sqrt{\|\mathbf{f}\|_2}}$. Therefore, $\tilde{\mathbf{f}} = \frac{1}{S}\sum_{xy}\tilde{\mathbf{f}}_{xy}$.

Part-aligned interpretation. We can decompose $\mathbf{a} \otimes \mathbf{p}$[6] into c_P components:

$$\text{vec}(\mathbf{a} \otimes \mathbf{p}) = [(p_1\mathbf{a})^\top \ (p_2\mathbf{a})^\top \ \dots (p_{c_P}\mathbf{a})^\top]^\top, \tag{5}$$

where each sub-vector $p_i\mathbf{a}$ corresponds to a i-th part channel. For example, if $p_{knee} = 1$ on knee and 0 otherwise, then $p_{knee}\mathbf{a}$ becomes $\mathbf{a}$ only on the knee and $\mathbf{0}$ otherwise. Thus, we call $\text{vec}(\mathbf{a} \otimes \mathbf{p})$ as part-aligned representation. In general, each channel c does not necessarily correspond to a certain body part. However, the part-aligned representation remains valid as $\mathbf{p}$ encodes the body part information. Section 4 describes this interpretation in detail.

[6] We drop the subscript xy for presentation clarification

3.3 Loss

To train the network, we utilize the widely-used triplet loss function. Let $\mathbf{I}_q$, $\mathbf{I}_p$ and $\mathbf{I}_n$ denote the query, positive and negative images, respectively. Then, $(\mathbf{I}_q, \mathbf{I}_p)$ is a pair of images of the same person, and $(\mathbf{I}_q, \mathbf{I}_n)$ is that of different persons. Let $\tilde{\mathbf{f}}_q$, $\tilde{\mathbf{f}}_p$, and $\tilde{\mathbf{f}}_n$ indicate their representations. The triplet loss function is formulated as

$$\ell_{\text{triplet}}(\tilde{\mathbf{f}}_q, \tilde{\mathbf{f}}_p, \tilde{\mathbf{f}}_n) = \max(m + \text{sim}(\tilde{\mathbf{f}}_q, \tilde{\mathbf{f}}_n) - \text{sim}(\tilde{\mathbf{f}}_q, \tilde{\mathbf{f}}_p), 0), \tag{6}$$

where m denotes a margin and $\text{sim}(\mathbf{x}, \mathbf{y}) =< \mathbf{x}, \mathbf{y} >$. The margin is empirically set as $m = 0.2$. The overall loss function is written as follows.

$$\mathcal{L} = \frac{1}{|\mathcal{T}|} \sum\nolimits_{(\mathbf{I}_q, \mathbf{I}_p, \mathbf{I}_n) \in \mathcal{T}} \ell_{\text{triplet}}(\tilde{\mathbf{f}}_q, \tilde{\mathbf{f}}_p, \tilde{\mathbf{f}}_n), \tag{7}$$

where $\mathcal{T}$ is the set of all triplets, $\{(\mathbf{I}_q, \mathbf{I}_p, \mathbf{I}_n)\}$.

4 Analysis

Part-aware image similarity. We show that under the proposed part-aligned representation in Eqs.(3) and (4), the similarity between two images is equivalent to the aggregation of local appearance similarities between the corresponding body parts. The similarity between two images can be represented as the sum of local similarities between every pair of locations as follows.

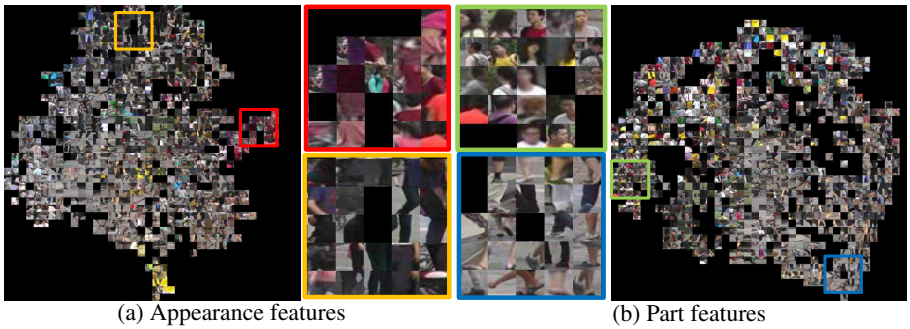
$$\begin{aligned} \text{sim}_I(\mathbf{I}, \mathbf{I}') =< \tilde{\mathbf{f}}, \tilde{\mathbf{f}}' > &= \frac{1}{S^2} < \sum_{xy} \tilde{\mathbf{f}}_{xy}, \sum_{x'y'} \tilde{\mathbf{f}}'_{x'y'} > \\ &= \frac{1}{S^2} \sum_{xy} \sum_{x'y'} < \tilde{\mathbf{f}}_{xy}, \tilde{\mathbf{f}}'_{x'y'} > \\ &= \frac{1}{S^2} \sum_{xy} \sum_{x'y'} \text{sim}(\tilde{\mathbf{f}}_{xy}, \tilde{\mathbf{f}}'_{x'y'}), \end{aligned} \tag{8}$$

where $\text{sim}_I(,)$ measures the similarity between images. Here, the local similarity is computed by an inner product:

$$\begin{aligned} \text{sim}(\tilde{\mathbf{f}}_{xy}, \tilde{\mathbf{f}}'_{x'y'}) &=< \text{vec}(\tilde{\mathbf{a}}_{xy} \otimes \tilde{\mathbf{p}}_{xy}), \text{vec}(\tilde{\mathbf{a}}'_{x'y'} \otimes \tilde{\mathbf{p}}'_{x'y'}) > \\ &=< \tilde{\mathbf{a}}_{xy}, \tilde{\mathbf{a}}'_{x'y'} >< \tilde{\mathbf{p}}_{xy}, \tilde{\mathbf{p}}'_{x'y'} > \\ &= \text{sim}(\tilde{\mathbf{a}}_{xy}, \tilde{\mathbf{a}}'_{x'y'}) \, \text{sim}(\tilde{\mathbf{p}}_{xy}, \tilde{\mathbf{p}}'_{x'y'}). \end{aligned} \tag{9}$$

This local similarity can be interpreted as the appearance similarity weighted by the body part similarity or vice versa. Thus, from Eqs(8) and (9), the similarity between two images is computed as the average of local appearance similarities weighted by the body part similarities at the corresponding positions:

$$\text{sim}_I(\mathbf{I}, \mathbf{I}') = \frac{1}{S^2} \sum_{xyx'y'} \text{sim}(\tilde{\mathbf{a}}_{xy}, \tilde{\mathbf{a}}'_{x'y'}) \, \text{sim}(\tilde{\mathbf{p}}_{xy}, \tilde{\mathbf{p}}'_{x'y'}).$$



(a) Appearance features (b) Part features

Fig. 3. The t-SNE visualization of the normalized local appearance and part descriptors on the Market-1501 dataset. It illustrates that our two-stream network decomposes the appearance and part information into two streams successfully. (a) Appearance descriptors are clustered roughly by colors, independently from the body parts where they came from. (b) Part descriptors are clustered by body parts where they came from, regardless of the colors. (Best viewed on a monitor when zoomed in)

As a result, the image similarity does not depend on the relative positions of parts in images, and therefore the misalignment problem is reduced. To make the local part similarity to be always non-negative and therefore the sign of the local similarity depends only on the sign of the local appearance similarity, we can also restrict the part descriptors $\mathbf{p}_{xy}$ to be element-wise non-negative by adding a ReLU layer after the part map extractor $\mathcal{P}$ as shown in Figure 2. As this variant results in similar accuracy to the original one, we used the model without the ReLU layer for all the experiments. See supplementary material for more details.

Relationship to the baseline models. Consider a baseline approach that only uses the appearance maps and spatial global pooling for image representation. Then, the image similarity is computed as $\mathrm{sim}_I(\mathbf{I}, \mathbf{I}') = \frac{1}{S^2} \sum_{xyx'y'} \mathrm{sim}(\tilde{\mathbf{a}}_{xy}, \tilde{\mathbf{a}}'_{x'y'})$. Unlike our model, this approach cannot reflect part similarity. Consider another model based on the box-based representation, which represents an image as a concatenation of K body part descriptors, where k-th body part is represented as the average-pooled appearance feature within the corresponding bounding box. This model is equivalent to our model when $\mathbf{p}_{xy}$ is defined as $\mathbf{p}_{xy} = [\delta[(x, y) \in R_1], \cdots, \delta[(x, y) \in R_K]]$, where R_k is the region within the k-th part bounding box and $\delta[\cdot]$ is an indicator function, i.e., $\delta[x] = 1$ if x is true otherwise 0. Because our model contains these baselines as special cases and is trained to optimize the re-identification loss, it is guaranteed to perform better than them.

The two-stream network yields a decomposed appearance and part maps. At the beginning of the training, the two streams of the network mainly represent the appearance and part maps because the appearance map extractor $\mathcal{A}$ and the part map extractor $\mathcal{P}$ are initialized using GoogleNet [54] pre-trained on ImageNet [46] and OpenPose [4] model pre-trained on COCO [29], respectively. During training, we do not set any constraints on the two streams, i.e., no annotations for the body parts, but only optimize the re-identification loss. Surprisingly, the trained two-stream network maintains to decom-

Fig. 4. Visualization of the appearance maps $\mathbf{A}$ and part maps $\mathbf{P}$ obtained from the proposed method. For a given input image (left), appearance (center) and part (right) maps encode the appearance and body parts, respectively. For both appearance and part maps, the same color implies that the descriptors are similar, whereas different colors indicate that the descriptors are different. The appearance maps share similar color patterns among the images from the same person, which means that the patterns of appearance descriptors are similar as well. In the part maps, the color differs depending on the location of the body parts where the descriptors came from. (Best viewed in color)

Fig. 5. Comparing the body part descriptors. For a given image (left), the conventional joint-based (center) and the proposed (right) descriptors are visualized. (Best viewed in color)

pose the appearance and part information into two streams: one stream corresponds to the appearance maps and the other corresponds to the body part maps.

We visualize the distribution of the learned local appearance and part descriptors using t-SNE [37] as shown in Figures 3 (a) and (b). Figure 3 (a) shows that the appearance descriptors are clustered depending on the appearance while being independent on the parts that they come from. For example, the red/yellow box shows that the red/black-colored patches are closely embedded, respectively. By contrast, Figure 3 (b) illustrates that the local part embedding maps the similar body parts into close regions regardless of color. For example, the green/blue box shows that the features from the head/lower leg are clustered, respectively. In addition, physically adjacent body parts, such as head–shoulder and shoulder–torso, are also closely embedded.

To understand how the learned appearance/part descriptors are used in person re-identification, we visualize the appearance maps $\mathbf{A}$ and the part maps $\mathbf{P}$ following the visualization used in SIFTFlow [32], as shown in Figure 4.For a given input image (left), the appearance (center) and part (right) maps encode the appearance and body parts, respectively. The figure shows how the appearance maps differentiate different persons

while being invariant for each person. By contrast, the part maps encode the body parts independently from their appearance. In particular, a certain body part is represented by a similar color across images, which confirms our observation in Figure 3 that the part features from physically adjacent regions are closely embedded.

Our approach learns the optimal part descriptor for person re-identification, rather than relying on the pre-defined body parts. Figure 5 qualitatively compares the conventional body part descriptor and the one learned by our approach. [7] In the previous works on human pose estimation [62, 4, 41], human poses are represented as a collection of pre-defined key body joint locations. It corresponds to a part descriptor which one-hot encodes the key body joints depending on the existence of a certain body joint at the location, e.g, $p_{knee} = 1$ on knee and 0 otherwise. Compared to the baseline, ours smoothly maps the body parts. In other words, the colors are continuous over the whole body in ours, which implies that the adjacent body parts are mapped closely. By contrast, the baseline not always maps adjacent body parts maps closely. For example, the upper leg between the hip and knee is more close to the background descriptors than to ankle or knee descriptors. This smooth mapping makes our method to work robustly against the pose estimation error because the descriptors do not change rapidly along the body parts and therefore are insensitive to the error in estimation. In addition, the part descriptors adopt to distinguish the informative parts more finely. For example, the mapped color varies sharply from elbow to shoulder and differentiates the detailed regions. Based on these properties, the learned part descriptors better support the person re-identification task and improve the accuracy.

5 Implementation Details

Network architecture. We use a sub-network of the first version of GoogLeNet [54] as the appearance map extractor $\mathcal{A}$, from the image input of size 160×80 to the output of *inception4e*, which is followed by a 1×1 convolution layer and a batch normalization layer to reduce the dimension to 512 (Figure 2). Moreover, we optionally adopt dilation filters in the layers from the *inception4a* to the final layer, resulting in 20×10 response maps. Figure 2 illustrates the architecture of the part map extractor $\mathcal{P}$. We use a sub-network of the OpenPose network [4], from the image input to the output of stage2 (i.e., *concat_stage3*) to extract 185 pose heat maps, which is followed by a 3×3 convolution layer and a batch normalization layer, thereby outputting 128 part maps. We adopt the compact bilinear pooling [14] to aggregate the two feature maps into a 512-dimensional vector $\mathbf{f}$.

Compact bilinear pooling. The bilinear transformation over the 512-dimensional appearance vector and the 128-dimensional part vector results in an extremely high dimensional vector, which consumes large computational cost and memory. To resolve this issue, we use the tensor sketch approach [44] to compute a compact representation as in [14]. The key idea of the tensor sketch approach is that the original inner product, on which the Euclidean distance is based, between two high-dimensional vectors can be approximated as an inner product of the dimension-reduced vectors, which are random projections of the original vectors. Details can be found in [44].

[7] We used the visualization method proposed in SIFTFlow [32]

Network training. The appearance map extractor $\mathcal{A}$ and part map extractor $\mathcal{P}$ are fine-tuned from the network pre-trained on ImageNet [46] and COCO [29], respectively. The added layers are initialized following [17]. We use the stochastic gradient descent algorithm. The initial learning rate, weight decay, and the momentum are set to 0.01, 2×10^{-4}, and 0.9, respectively. The learning rate is decreased by a factor of 5 after every $20,000$ iterations. All the networks are trained for $75,000$ iterations.

We follow [76] to sample a mini-batch of samples at each iteration and use all the possible triplets within each mini-batch. The gradients are computed using the acceleration trick presented in [76]. In each iteration, we sample a mini-batch of 180 images, e.g., there are on average 18 identities with each containing 10 images. In total, there are approximately $10^2 \cdot (180 - 10) \cdot 18 \approx 3 \times 10^5$ triplets in each iteration.

6 Experiments

6.1 Datasets

Market-1501 [80]. This dataset is one of the largest benchmark datasets for person re-identification. Six cameras are used: five high-resolution cameras and one low-resolution camera. There are $32,668$ DPM-detected pedestrian image boxes of $1,501$ identities: 750 identities are utilized for training and the remaining 751 identities are used for testing. There are $3,368$ query images and $19,732$ gallery images with $2,793$ distractors.
CUHK03 [25]. This dataset consists of $13,164$ images of $1,360$ people captured by six cameras. Each identity appears in two disjoint camera views (i.e., 4.8 images in each view on average). We divided the train/test set following the previous work [25]. For each test identity, two images are randomly sampled as the probe and gallery images and the average accuracy over 20 trials is reported as the final result.
CUHK01 [24]. This dataset comprises 3884 images of 971 people captured in two disjoint camera views. Two images are captured for each person from each of the two cameras (i.e., a total of four images). Experiments are performed under two evaluation settings [1], using 100 and 486 test IDs. Following the previous works [1, 7, 10, 76], we fine-tuned the model from the one learned from the CUHK03 training set for the experiments with 486 test IDs.
DukeMTMC [45]. This dataset is originally proposed for video-based person tracking and re-identification. We use the fixed train/test split and evaluation setting following [31][8]. It includes $16,522$ training images of 702 identities, $2,228$ query images of 702 identities and $17,661$ galley images.
MARS [78]. This dataset is proposed for video-based person re-identification. It consists of 1261 different pedestrians captured by at least two cameras. There are $509,914$ bounding boxes and $8,298$ tracklets from 625 identities for training and $681,089$ bounding boxes and $12,180$ tracklets from 636 identities for testing.

6.2 Evaluation Metrics

We use both the cumulative matching characteristics (CMC) and mean average precision (mAP) to evaluate the accuracy. The CMC score measures the quality of identi-

[8] https://github.com/layumi/DukeMTMC-reID_evaluation

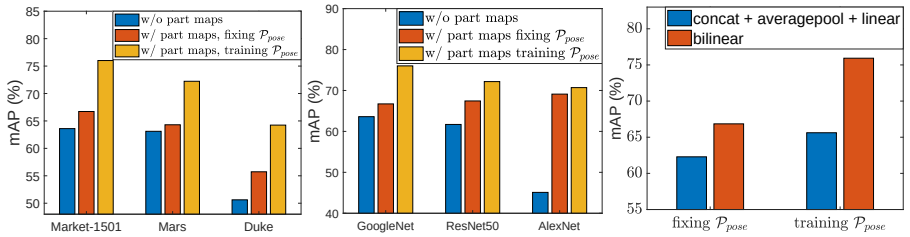


Fig. 6. (a) Comparison of different pooling methods on the appearance maps. (c) Comparing models, with and without part maps, on different datasets. (d) Comparing models, with and without part maps, on different architectures of the appearance map extractor. If not specified, the results are reported on Market-1501. (b) Comparison of different methods to aggregate the appearance and part maps.

fying the correct match at each rank. For multiple ground truth matches, CMC cannot measure how well all the images are ranked. Therefore, we report the mAP scores for Market-1501, DukeMTMC, and MARS where more than one ground truth images are in the gallery.

6.3 Comparison with the Baselines

We compare the proposed method with the baselines in three aspects. In this section, when not specified, all the experiments are performed on the Market-1501 dataset, all the models do not use dilation, and $\mathcal{P}_{pose}$ is trained together with the other parameters.
Effect of part maps. We compare our method with a baseline that does not explicitly use body parts. As a baseline network, we use the appearance map extractor of Eq.(1), which is followed by a global spatial average pooling and a fully connected layer, thereby outputting the 512-dimensional image descriptor. Figures 6 (a) and (b) compare the proposed method with the baseline, while varying the training strategy: *fixing* and *training* $\mathcal{P}_{pose}$. *Fixing* $\mathcal{P}_{pose}$ initializes $\mathcal{P}_{pose}$ using the pre-trained weights [4, 29] and fixes the weight through the training. *Training* $\mathcal{P}_{pose}$ also initializes $\mathcal{P}_{pose}$ in the same way, but fine-tunes the network using the loss of Eq.(7) during training. Figure 6 (a) illustrates the accuracy comparison on three datasets, Market-1501, MARS, and Duke. It shows that using part maps consistently improves the accuracy on all the three datasets from the baseline. In addition, training $\mathcal{P}_{pose}$ largely improves the accuracy than fixing $\mathcal{P}_{pose}$. It implies that the part descriptors are adopted to better serve the person re-identification task. Figure 6 (b) shows the accuracy comparison while varying the appearance sub-network architecture. Similarly, the baseline accuracy is improved when part maps are introduced and further improved when $\mathcal{P}_{pose}$ is fine-tuned during training.
Effect of bilinear pooling. Figure 6 (c) compares the proposed method (*bilinear*) to the baseline with a different aggregator. For the given appearance and part maps, *concat+averagepool+linear* generates a feature vector by concatenating two feature maps, spatially average pooling, and feeding through a fully connected layer, resulting in a

512-dimensional vector. The result shows that bilinear pooling consistently achieves higher accuracy than the baseline, for both cases when $\mathcal{P}_{pose}$ is fixed/trained.

Comparison with previous pose-based methods. Finally, we compare our method with three previous works [79, 75, 50], which use human pose estimation, on Market-1501. For a fair comparison, we use the reduced CPM(R-CPM [∼3M param]) utilized in [50][9] as $\mathcal{P}_{pose}$. The complexity of the R-CPM is lower than the standard FCN (∼6M param) used in [75] and CPM (30M param) used in [79]. As the appearance network, [75] used GoogLeNet and [79] used ResNet50. [50] used 13 inception modules, whereas we use 7. Table 1 shows the comparison. In comparison with the method adopted by [79, 75, 50], the proposed method (Inception V1, R-CPM) achieves an increase of 4% and 9% for rank@1 accuracy and mAP, respectively. It shows that our method effectively uses the part information compared with the previous approaches.

6.4 Comparison with State-of-the-Art Methods

Market-1501. Table 1 shows the comparison over two query schemes, single query and multi-query. Single query takes one image from each person whereas multi-query takes multiple images. For the multi-query setting, one descriptor is obtained from multiple images by averaging the feature from each image. Our approach achieves the best accuracy in terms of both mAP and rank@K for both single and multi-query. We also provide the result after re-ranking [86], which further boosts accuracy. In addition, we conduct the experiment over an expanded dataset with additional $500K$ images [80]. Following the standard evaluation protocol [19], we report the results over four different gallery sets, $19,732$, $119,732$, $219,732$, and $519,732$, using two evaluation metrics (i.e., rank-1 accuracy and mAP). Table 2 reports the results. The proposed method outperforms all the other methods.

CUHK03. We report the results with two person boxes: manually labeled and detected. Table 3 presents the comparison with existing solutions. In the case of detected boxes, the state-of-the-art accuracy is achieved. With manual bounding boxes, our method also achieves the best accuracy.

CUHK01. We compare the results with two evaluation settings (i.e., 100 and 486 test IDs) in Table 3. For 486 test IDs, the proposed method shows the best result. For 100 test IDs, our method achieves the second best result, following [16]. Note that [16] fine-tuned the model which is learned from the CUHK03+Market1501, whereas we trained the model using 871 training IDs of the CUHK01 dataset, following the settings in previous works [1, 7, 10, 76].

DukeMTMC. We follow the setting in [31] to conduct the experiments. Table 4 reports the results. The proposed method achieves the best result for both with and without re-ranking.

MARS. We also evaluate our method on one video-based person re-identification dataset [78]. We use our approach to extract the representation for each frame and aggregate the representations of all the frames using temporal average pooling, which shows similar accuracy to other aggregation schemes (RNN and LSTM). Table 5 presents the com-

[9] https://github.com/yokattame/SpindleNet

Table 1. Accuracy comparison on Market-1501

	Sinlge Query					Multi Query				
Rank	1	5	10	20	mAP	1	5	10	20	mAP
Varior et al. 2016 [58]	61.6	-	-	-	35.3	-	-	-	-	-
Zhong et al. 2017 [86]	77.1	-	-	-	63.6	-	-	-	-	-
Zhao et al. 2017 [76]	80.9	91.7	94.7	96.6	63.4	-	-	-	-	-
Sun et al. 2017 [53]	82.3	92.3	95.2	-	62.1	-	-	-	-	-
Geng et al. 2016 [16]	83.7	-	-	-	65.5	89.6	-	-	-	73.8
Lin et al. 2017 [31]	84.3	93.2	95.2	97.0	64.7	-	-	-	-	-
Bai et al. 2017 [2]	82.2	-	-	-	68.8	88.2	-	-	-	76.2
Chen et al. 2017 [9]	72.3	88.2	91.9	95.0	-	-	-	-	-	
Hermans et al. 2017 [19]	84.9	94.2	-	-	69.1	90.5	96.3	-	-	76.4
+ re-ranking	86.7	93.4	-	-	81.1	91.8	95.8	-	-	87.2
Zhang et al. 2017 [74]	87.7	-	-	-	68.8	91.7	-	-	-	77.1
Zhong et al. 2017 [87]	87.1	-	-	-	71.3	-	-	-	-	-
+ re-ranking	89.1	-	-	-	83.9	-	-	-	-	-
Chen et al. 2017 [8] (MobileNet)	90.0	-	-	-	70.6	-	-	-	-	-
Chen et al. 2017 [8] (Inception-V3)	88.6	-	-	-	72.6	-	-	-	-	-
Ustinova et al. 2017 [57] (Bilinear)	66.4	85.0	90.2	-	41.2	-	-	-	-	-
Zheng et al. 2017 [79] (Pose)	79.3	90.8	94.4	96.5	56.0	-	-	-	-	-
Zhao et al. 2017 [75] (Pose)	76.9	91.5	94.6	96.7	-	-	-	-	-	-
Su et al. 2017 [50] (Pose)	84.1	92.7	94.9	96.8	65.4	-	-	-	-	-
Proposed (Inception-V1, R-CPM)	88.8	95.6	97.3	98.6	74.5	92.9	97.3	98.4	99.1	81.7
Proposed (Inception-V1, OpenPose)	**90.2**	**96.1**	**97.4**	**98.4**	**76.0**	93.2	97.5	98.4	99.1	82.7
+ dilation	**91.7**	**96.9**	**98.1**	**98.9**	**79.6**	**94.0**	**98.0**	**98.8**	**99.3**	**85.2**
+ re-ranking	**93.4**	**96.4**	**97.4**	**98.2**	**89.9**	**95.4**	**97.5**	**98.2**	**98.9**	**93.1**

Table 2. Accuracy comparison on Market-1501+500k.

	Gallery size				
	metric	19732	119732	219732	519732
Zheng et al. 2017 [84]	rank-1	79.5	73.8	71.5	68.3
	mAP	59.9	52.3	49.1	45.2
Linet al. 2017 [31]	rank-1	84.0	79.9	78.2	75.4
	mAP	62.8	56.5	53.6	49.8
Hermans et al. 2017 [19]	rank-1	84.9	79.7	77.9	74.7
	mAP	69.1	61.9	58.7	53.6
Proposed (Inception V1, OpenPose)	rank-1	**91.7**	**88.3**	**86.6**	**84.1**
	mAP	**79.6**	**74.2**	**71.5**	**67.2**

Table 3. Accuracy comparison on CUHK03 and CUHK01

	CUHK03								CUHK01							
	Detected				Manual				100 test IDs				486 test IDs			
Rank	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
Shi et al. [70]	52.1	84.0	92.0	96.8	61.3	88.5	96.0	99.0	69.4	90.8	96.0	-	-	-	-	-
SIR-CIR [60]	52.2	-	-	-	-	-	-	-	71.8	91.6	96.0	98.0	-	-	-	-
Varior et al. [58]	68.1	88.1	94.6	98.8	-	-	-	-	-	-	-	-	-	-	-	-
Bai et al. [2]	72.7	92.4	96.1	-	76.6	94.6	98.0	-	-	-	-	-	-	-	-	-
Zhang et al. [72]	-	-	-	-	80.2	97.7	99.2	99.8	89.6	97.8	98.9	99.7	76.5	94.2	**97.5**	-
Sun et al. [53]	81.8	95.2	97.2	-	-	-	-	-	-	-	-	-	-	-	-	-
Zhao et al. [76]	81.6	97.3	98.4	**99.5**	85.4	97.6	99.4	**99.9**	88.5	**98.4**	**99.6**	**99.9**	74.7	92.6	96.2	98.4
Geng et al. [16]	84.1	-	-	-	85.4	-	-	-	**93.2**	-	-	-	77.0	-	-	-
Chen et al. [9]	87.5	97.4	98.7	**99.5**	-	-	-	-	-	-	-	-	74.5	91.2	94.8	97.1
Ustinova et al. [57] (Bilinear)	63.7	89.2	94.7	97.5	69.7	93.4	98.9	99.4	-	-	-	-	52.9	78.1	86.3	92.6
Zheng et al. [79] (Pose)	67.1	92.2	96.6	98.1	-	-	-	-	-	-	-	-	-	-	-	-
Zhao et al. [75] (Pose)	-	-	-	-	88.5	97.8	98.6	99.2	-	-	-	-	79.9	**94.4**	97.1	98.6
Su et al. [50] (Pose)	78.3	94.8	97.2	98.4	88.7	98.6	99.2	99.7	-	-	-	-	-	-	-	-
Proposed	**88.0**	**97.6**	**98.6**	**99.0**	**91.5**	**99.0**	**99.5**	**99.9**	90.4	97.1	98.1	98.9	**80.7**	**94.4**	**97.3**	**98.6**

Table 4. Accuracy comparison on DukeMTMC

Rank	1	5	10	20	mAP
Zheng et al. [85]	67.7	-	-	-	47.1
Tong et al. [67]	68.1	-	-	-	-
Lin et al. [31]	70.7	-	-	-	51.9
Schumann et al. [47]	72.6	-	-	-	52.0
Sun et al. [53]	76.7	86.4	89.9	-	56.8
Chen et al. [8] (MobileNet)	77.6	-	-	-	58.6
Chen et al. [8] (Inception-V3)	79.2	-	-	-	60.6
Zhun et al. [87]	79.3	-	-	-	62.4
+ re-ranking	84.0	-	-	-	78.3
Proposed (Inception V1, OpenPose)	82.1	90.2	92.7	95.0	64.2
+ dilation	**84.4**	**92.2**	**93.8**	**95.7**	**69.3**
+ re-ranking	**88.3**	**93.1**	**95.0**	**96.1**	**83.9**

Table 5. Accuracy comparison on MARS

Rank	1	5	10	20	mAP
Xu et al. [68] (Video)	44	70	74	81	-
McLaughlin et al. [40] (Video)	45	65	71	78	27.9
Zheng et al. [78] (Video)	68.3	82.6	-	89.4	49.3
Liu et al. [33] (Video)	68.3	81.4	-	90.6	52.9
Zhou et al. [88]	70.6	90.0	-	97.6	50.7
Li et al. [23]	71.8	86.6	-	93.1	56.1
+ re-ranking	83.0	93.7	-	97.6	66.4
Liu et al. [35]	73.7	84.9	-	91.6	51.7
Hermans et al. [19]	79.8	91.4	-	-	67.7
+ re-ranking	81.2	90.8	-	-	77.4
Proposed (Inception V1, OpenPose)	83.0	92.8	95	96.8	72.2
+ dilation	**84.7**	**94.4**	**96.3**	**97.5**	**75.9**
+ re-ranking	**85.1**	**94.2**	**96.1**	**97.4**	**83.9**

parison with the competing methods. Our method shows the highest accuracy over both image-based and video-based approaches.

7 Conclusions

We propose a new method for person re-identification. The key factors that contribute to the superior performance of our approach are as follows. (1) We adopt part maps where parts are not pre-defined but learned specially for person re-identification. They are learned to minimize the re-identification loss with the guidance of the pre-trained pose estimation model. (2) The part map representation provides a fine-grained/robust differentiation of the body part depending on their usefulness for re-identification. (3) We use part-aligned representations to handle the body part misalignment problem. The resulting approach achieves superior/competitive person re-identification performances on the standard image and video benchmark datasets.

Acknowledgement This work was partially supported by Microsoft Research Asia and the Visual Turing Test project (IITP-2017-0-01780) from the Ministry of Science and ICT of Korea.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. arXiv:1703.08359 (2017)
3. Bak, S., Corvée, E., Brémond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: AVSS (2010)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
5. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: CVPR (2016)
6. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: CVPR (2015)
7. Chen, S.Z., Guo, C.C., Lai, J.H.: Deep ranking for person re-identification via joint representation learning. IEEE TIP **25**(5), 2353–2367 (2016)
8. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: CVPR Workshop (2017)
9. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE TPAMI (2017)
10. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR (2016)
11. Cheng, D.S., Cristani, M.: Person re-identification by articulated appearance matching. In: Person Re-Identification. Springer (2014)
12. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC (2011)
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
14. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: CVPR (2016)
15. Garcia, J., Martinel, N., Micheloni, C., Gardel, A.: Person re-identification ranking optimisation by discriminant context information analysis. In: ICCV (2015)
16. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv:1611.05244 (2016)
17. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS (2010)
18. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV (2008)
19. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
20. Jing, X.Y., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., Hu, R., Xu, B.: Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: CVPR (2015)
21. Kim, J.H., On, K.W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: ICLR (2017)
22. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Person re-identification by unsupervised ℓ_1 graph learning. In: ECCV (2016)
23. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR (2017)
24. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: ACCV (2012)

25. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
26. Li, X., Zheng, W.S., Wang, X., Xiang, T., Gong, S.: Multi-scale learning for low-resolution person re-identification. In: ICCV (2015)
27. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
28. Liao, S., Li, S.Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: ICCV (2015)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
30. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: ICCV (2015)
31. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv:1703.07220 (2017)
32. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE TPAMI **5**(33), 978–994 (2011)
33. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S.: Video-based person re-identification with accumulative motion context. arXiv:1701.00193 (2017)
34. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. arXiv:1709.09930 (2017)
35. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR (2017)
36. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: ECCV (2012)
37. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR **9**(Nov), 2579–2605 (2008)
38. Martinel, N., Das, A., Micheloni, C., Roy-Chowdhury, A.K.: Temporal model adaptation for person re-identification. In: ECCV (2016)
39. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: CVPR (2016)
40. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR (2016)
41. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
42. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR (2015)
43. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR (2016)
44. Pham, N., Pagh, R.: Fast and scalable polynomial kernels via explicit feature maps. In: SIGKDD (2013)
45. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV workshop on Benchmarking Multi-Target Tracking (2016)
46. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
47. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: CVPR workshops (2017)
48. Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., Wang, J.: Person re-identification with correspondence structure learning. In: ICCV (2015)
49. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: CVPR (2015)

50. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
51. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: ICCV (2015)
52. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: ECCV (2016)
53. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV (2017)
54. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
55. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. In: CVPR (2015)
56. Tang, S., Andriluka, M., Andres, B., Schiele, B.: Multi people tracking with lifted multicut and person re-identification. In: CVPR (2017)
57. Ustinova, E., Ganin, Y., Lempitsky, V.: Multiregion bilinear convolutional neural networks for person re-identification. In: AVSS (2017)
58. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV (2016)
59. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: ECCV (2016)
60. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: CVPR (2016)
61. Wang, H., Gong, S., Zhu, X., Xiang, T.: Human-in-the-loop person re-identification. In: ECCV (2016)
62. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
63. Weinrich, C., Gross, M.V.H.M.: Appearance-based 3d upper-body pose estimation and person re-identification on mobile robots. In: ICSMC. IEEE (2013)
64. Wu, L., Shen, C., van den Hengel, A.: Personnet: Person re-identification with deep convolutional neural networks. arXiv:1601.07255 (2016)
65. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR (2016)
66. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv:1604.01850 (2016)
67. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR (2017)
68. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: ICCV (2017)
69. Xu, Y., Lin, L., Zheng, W., Liu, X.: Human re-identification by matching compositional template with cluster sampling. In: ICCV (2013)
70. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: ICLR (2014)
71. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR (2016)
72. Zhang, Y., Li, X., Zhao, L., Zhang, Z.: Semantics-aware deep correspondence structure learning for robust person re-identification. In: IJCAI (2016)
73. Zhang, Y., Li, B., Lu, H., Irie, A., Ruan, X.: Sample-specific svm learning for person re-identification. In: CVPR (2016)
74. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Dual mutual learning. arXiv:1706.00384 (2017)

75. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR (2017)
76. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
77. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: CVPR (2014)
78. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: A video benchmark for large-scale person re-identification. In: ECCV (2016)
79. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. arXiv:1701.07732 (2017)
80. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
81. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: CVPR (2015)
82. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv:1610.02984 (2016)
83. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: ICCV (2015)
84. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person re-identification. arXiv:1611.05666 (2016)
85. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. ICCV (2017)
86. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. CVPR (2017)
87. Zhong, Z., Zheng, L., Kang, G., Shaozi, L., Yi, Y.: Random erasing data augmentation. arXiv:1708.04896 (2017)
88. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: CVPR (2017)