

Densely Semantically Aligned Person Re-Identification

Zhizheng Zhang^{1*}

Cuiling Lan^{2†}

Wenjun Zeng²

Zhibo Chen^{1†}

¹University of Science and Technology of China

²Microsoft Research Asia

zhizheng@mail.ustc.edu.cn

{culan, wezeng}@microsoft.com

chenzhibo@ustc.edu.cn

Abstract

We propose a densely semantically aligned person re-identification framework. It fundamentally addresses the body misalignment problem caused by pose/viewpoint variations, imperfect person detection, occlusion, etc. By leveraging the estimation of the dense semantics of a person image, we construct a set of densely semantically aligned part images (DSAP-images), where the same spatial positions have the same semantics across different images. We design a two-stream network that consists of a main full image stream (MF-Stream) and a densely semantically-aligned guiding stream (DSAG-Stream). The DSAG-Stream, with the DSAP-images as input, acts as a regulator to guide the MF-Stream to learn densely semantically aligned features from the original image. In the inference, the DSAG-Stream is discarded and only the MF-Stream is needed, which makes the inference system computationally efficient and robust. To the best of our knowledge, we are the first to make use of fine grained semantics to address the misalignment problems for re-ID. Our method achieves rank-1 accuracy of 78.9% (new protocol) on the CUHK03 dataset, 90.4% on the CUHK01 dataset, and 95.7% on the Market1501 dataset, outperforming state-of-the-art methods.

1. Introduction

Person re-identification (re-ID) aims to match a specific person across multiple camera views or in different occasions from the same camera view. It facilitates many important applications, such as cross-camera tracking [40].

This task is challenging due to large variations on person pose and viewpoint, imperfect person detection, cluttered background, occlusion, and lighting differences, etc. Many of these factors result in spatial misalignment of the human body as shown in Fig. 1, where the same spatial positions do not correspond to the same semantics. The misalignment is one of the key challenges [30, 33, 53, 37, 48, 34, 57], which compromises performance.



Figure 1. Examples to illustrate the challenges of spatial misalignment in person re-ID caused by (a) different camera viewpoints, (b) different poses, (c) imperfect person detection, (d) misalignment within a local part, (e) cluttered background, (f) occlusion.

Some paradigms employ the convolutional neural networks to learn global feature representation in an end-to-end manner [1, 43, 8, 4, 2]. However, the capability of the global representations is limited by: 1) the lack of emphasis on local differences [48], and 2) the absence of any explicit mechanism to tackle the misalignment [2].

In recent years, many efforts have been made to alleviate these problems [33, 53, 37, 48, 34, 19]. To make the features focus on some local details, some works make a straightforward partition of the person image into a few fixed rigid parts (e.g., horizontal stripes) and learn detailed local features [5, 37, 18, 3, 35, 39]. However, such a partition cannot well align the human body parts. Some works have attempted the use of pose (which identifies different types of parts, e.g., head, arm, etc.) to localize body parts for learning part-aligned features [47, 15, 49, 33, 53, 42]. However, the body part alignment based on pose is too coarse to have satisfactory alignment. As shown in Fig. 1 (d), even for the same type of parts, there is still spatial misalignment within the parts, where the human semantics are different for the same spatial positions. It becomes critical to design an architecture which enables the efficient learning of densely

*This work is done when Zhizheng Zhang is an intern at MSRA.

†Corresponding author

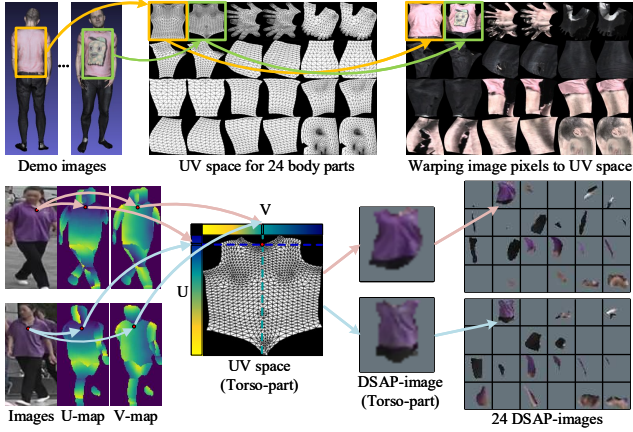


Figure 2. Illustration of the dense correspondences between a 2D person image and a surface-based canonical representation in the UV space. The person surface is partitioned into 24 body regions. Each region can be warped to a DSAP-image and the fine-grained (dense) semantics are spatially aligned for different person images.

semantically aligned features for re-ID.

In this paper, we propose a novel densely semantically aligned person re-ID framework, which fundamentally enables fine-grained semantic alignment and semantically aligned feature learning in re-ID.

First, we propose performing dense semantic alignment of the human body on a canonical space to address the misalignment challenges in person re-ID. We are inspired by the dense semantics estimation work of DensePose [9], which is capable of predicting the fine-grained pixel-level semantics of a person. Different from pose with only a limited number of coarse key joints, dense semantics establishes dense correspondences between a 2D person image and a 3D surface-based canonical representation of the human body [9, 10]. As illustrated in Fig. 2, the 3D surface of a person is segmented into 24 semantic body regions. Within a region, the semantics of each position is identified by a two-dimensional UV coordinate. Based on the estimated dense semantics in terms of UV coordinate values (on the U,V map), the original input image is warped to 24 *densely semantically aligned part images* (DSAP-images) in the UV space. In this way, person images with different viewpoints, poses, and backgrounds are semantically well aligned. Thus, such representation has the inherent merits for addressing misalignment challenges. Note that not only the coarse body part regions are aligned, but also the contents within a part are densely aligned at the pixel level.

Second, we propose a new framework intending to fully exploit the densely semantically aligned representations for person re-ID. For dense semantics estimation, since the person in a 2D image is a projection from a 3D person, nearly half of the 3D surface is invisible and thus cannot be detected from the 2D image (see the examples of the 24

DSAP-images in Fig. 2, where many of the DSAP-images do not have valid information). Besides, there are usually estimation errors, including missing detection, especially on the images of the re-ID dataset which usually have low resolution and blurring artifacts. It remains challenging to design an effective network to fully exploit the semantically aligned information as there are loss of information and noise there.

In our design, we leverage the densely semantically aligned information to drive the main network to learn semantically aligned features from the original image. As shown in Fig. 3, our network consists of a main full image stream (*MF-Stream*) and a densely semantically aligned guiding stream (*DSAG-Stream*). For the *MF-Stream*, the full image is taken as the input. For the *DSAG-Stream*, the 24 DSAP-images obtained from the dense semantic alignment module are taken as the input. Rather than making the features of the two streams both have re-ID ability, the *DSAG-Stream* acts as a regulator to guide the *MF-Stream* to learn semantically aligned features. We achieve this by element-wise fusing of the *MF-Stream* features and the *DSAG-Stream* features, with supervisions added on the fused features. End-to-end joint training enables the interaction and joint optimization of the two streams.

In summary, we have made three main contributions.

- We propose making use of dense semantic alignment for person re-ID, addressing the misalignment challenges.
- A densely semantically aligned deep learning based framework is proposed for person re-ID. To the best of our knowledge, our proposed framework is the first one to make use of fine grained semantics to address the misalignment problems for effective person re-ID. We propose an effective fusion and supervision design to facilitate semantically aligned feature learning. It enables the interaction between the *DSAG-Stream* and the *MF-Stream* during the learning process. This greatly enhances the power of the *MF-Stream* even though its input images are not semantically aligned.
- The *DSAG-Stream*, as a regulator, can be removed during the inference without sacrificing the performance. This also removes the dependency on the performance of the dense semantics estimator during inference, making the inference model more computationally efficient and robust to dense semantics estimation errors.

We perform extensive ablation studies and the experimental results demonstrate that our proposed architecture with the dense semantic alignment are very powerful. We achieve state-of-the art performance on the Market-1501, CUHK03, and CUHK01 datasets, and competitive performance on DukeMTMC-reID. On the CUHK03 dataset, our performance significantly outperforms the previous methods, by at least **+10.9%/+7.8%** in Rank-1/mAP accuracy.

2. Related Work

Body part/pose-aligned approaches. Spatial misalignment is ubiquitous and is one of the key challenges in re-ID. In the early works, some patch-based methods perform patch-level matching to address patch-wise misalignment [23, 51, 52]. To avoid mismatched patches with similar appearances [30], human semantics of part/pose is introduced so that the similarity matching is performed between the semantically corresponding parts [6, 46]. In recent years, human semantics in terms of part/pose is widely used to localize body parts for part-aligned deep feature learning and matching [47, 15, 49, 33, 53, 42]. In [42], body poses/parts are first detected and deep neural networks are designed for representation learning on both the local parts and global region. Some works rely on constrained attention selection mechanisms from human mask/part/pose to implicitly calibrate misaligned images [32, 25, 45, 14, 34].

All the above works aim to address misalignment at the coarse body part level. However, there is still misalignment within each part. Our work intends to fundamentally address the misalignment problem. It differs from previous works in three main aspects. First, our approach intends to fully exploit the fine-grained semantically aligned representations. Second, we leverage the semantically aligned representations, which play the role of regulators, to guide the semantic feature learning from the original image. Third, during inference, we do not need the *DSAG-Stream*, making our model computationally efficient and robust.

Local and global based approaches. Many approaches make use of both the global and local feature to simultaneously exploit their advantages [3, 39, 18, 42, 33, 49, 48]. Global features learned from the full image intend to capture the most discriminative clues of appearance but may fail to capture discriminative local details. Thus, part-based features are exploited as a remedy. Wang *et al.* design a multiple granularity network, which consists of one branch for global features and two branches for local feature representations [39]. In [48], the image feature map is rigidly divided into local stripes and a shortest path loss is introduced to align local stripes. This aids the global feature learning by means of sharing weights of the backbone network. However, the alignment is still too coarse without considering person dense semantics. We leverage the densely semantically aligned representation to *guide* the learning of both the global features and part-aware features.

Approaches based on joint multi-loss learning. Zheng *et al.* suggest that person re-ID lies in between image classification and instance retrieval [55]. The classification task and ranking task are complementary to each other. Recently, some approaches [43, 21, 4, 39] optimize the network simultaneously with both classification loss and ranking loss, *e.g.*, triplet loss [29, 13]. Similarly, we leverage the complementary advantages of the two tasks.

3. Densely Semantically Aligned Person Re-ID

We propose a new framework aiming to fully exploit the densely semantically aligned representations for robust person re-ID. Fig. 3 shows the flowchart. The network consists of two streams: the main full image stream (*MF-Stream*), and the densely semantically aligned guiding stream (*DSAG-Stream*). Based on the dense semantic alignment module, from the input person image, we construct 24 densely semantically aligned part images (DSAP-images) as the input to the *DSAG-Stream*. Having the merits of being semantically aligned, the *DSAG-Stream* acts as a regulator to regularize the feature learning of the *MF-Stream* from the original image, through our fusion and loss designs. The entire network is trained in an end-to-end manner. We discuss the details in the following subsections.

3.1. Construction of DSAP-images

Dense semantics annotation/estimation on 2D images [9] establishes dense correspondences from 2D images to the human body surface. Each position on the surface has a different semantic meaning, which can be parameterized/represented by a two-dimensional UV coordinate value [10, 9]. The same UV coordinate value corresponds to the same semantics. Thus, in the UV space, the dense semantics are inherently aligned.

For the dense semantic alignment module, the original RGB image is warped to the representation in UV space to obtain 24 DSAP-images based on the estimated dense semantics.

Dense semantics estimation. We adopt the off-the-shelf DensePose model (trained on the DensePose-COCO dataset) [9] to estimate the dense semantics of a 2D image. It segments a person to 24 surface-based body part regions. For each detected body part, the semantics for each pixel is provided in terms of a coordinate value (u, v) in the UV space, where $u, v \in [0, 1]$. Please refer to [9] for more details.

Warping. For the i -th body part region, as illustrated in Fig. 2, based on the semantics, the pixel values on the person can be warped onto a DSAP-image of size $S \times S$ in the deformation-free UV space, where the rows and columns represent the U and V, respectively. The DSAP-images are initialized by the mean values of images before the warping. Note that the background and not detected semantic positions are not warped. We simply copy the pixel value (r, g, b) of the body parts with its semantics estimated as (u, v) to the $(\lfloor u \times S \rfloor, \lfloor v \times S \rfloor)$ position of the corresponding DSAP-image. $\lfloor x \rfloor$ is the function to get the greatest integer less than or equal to x and we set S to 32 in our experiments.

Discussion. For the DSAP-images of the i -th body part, the semantic identities on the same spatial positions are always the same. They are densely semantically aligned.

Such representations have three major advantages. 1) It overcomes spatial misalignment challenges resulting from

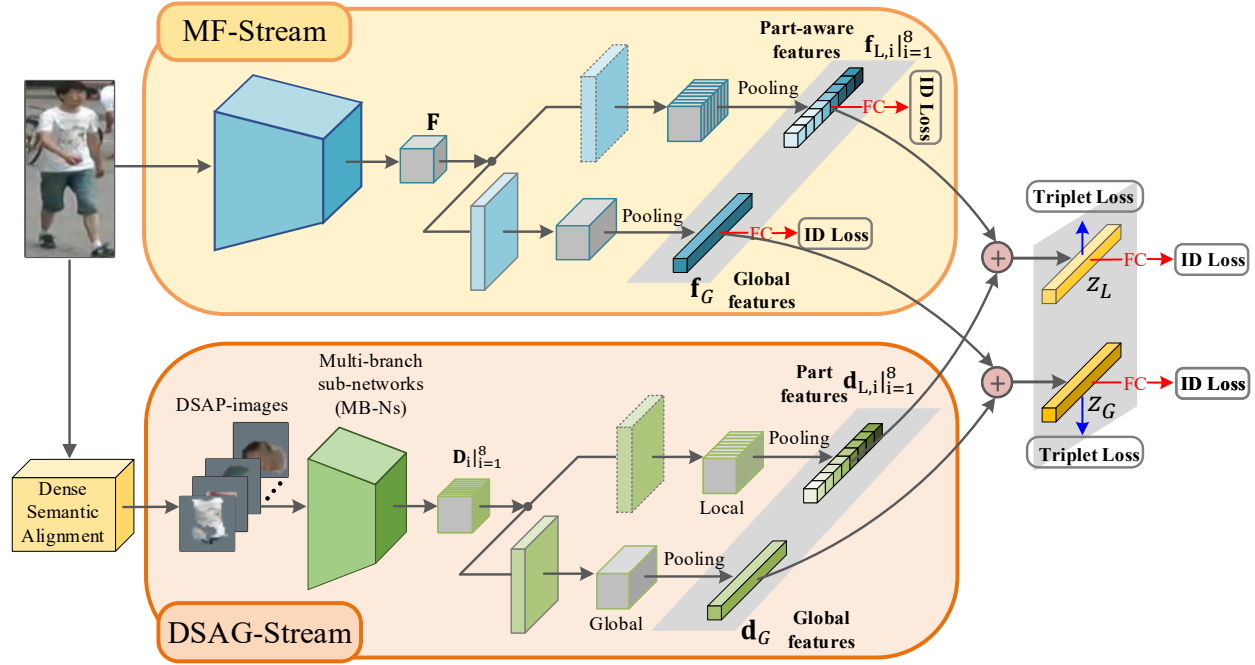


Figure 3. Flowchat of the proposed densely semantically aligned person re-ID (DSA-reID). It consists of two streams: *MF-Stream* and *DSAG-Stream*, which are jointly trained through our fusion and supervision design. The *DSAG-Stream*, with the input DSAP-images that are densely semantically aligned, plays the role of a regulator which facilitates the joint optimization of the entire network. In the inference, to be computationally efficient, the *DSAG-Stream* is discarded. The global features f_G and part-aware features $f_L = f_{L,i}|_{i=1}^8$ are used as the final features for re-ID. They are simultaneously exploited to make use of the global information and local detailed information.

diverse viewpoints and poses, and imperfect person detection. 2) It avoids the interference from diverse background clutters since only human body regions are warped to DSAP-images. 3) It is free from the appearance interference from occlusion since the semantics are not estimated over the occluding objects.

DSAP-images, however, have three limitations with respect to its roles in the person re-ID task. 1) The valid contents on the DSAP-images are very sparse (see Fig. 2). As a 2D projection from the 3D surface, nearly half of the body regions are invisible on the 2D image and thus cannot be detected by DensePose. Besides, there are usually estimation errors, including missing detections, especially on images with low resolution and blurring artifacts. 2) The dense semantics estimator is not optimal. Since there is no labeled dense semantics for the re-ID datasets, we leverage the DensePose model trained on the COCO-DesenPose dataset. However, there is a gap between these datasets in resolutions, image quality, and pose distributions. 3) Since the background is removed, some discriminative contents, such as a red backpack, are also removed.

3.2. Joint Learning of Our Network

Due to the sparsity of valid contents and potential semantics estimation errors on the DSAP-images (as discussed in

subsection 3.1), it is very challenging to design an effective network to exploit the semantically aligned information from the DSAP-images alone. In fact, a few of our early attempts along this line (with only the DSAP-images as input) have failed to deliver good results. To exploit the merits of the DSAP-images while addressing the above mentioned challenges, in our design, we propose treating them as regulators in an end-to-end network to drive the semantically aligned feature learning from the original full image. One important advantage of this design is that during the inference, the regulators are not needed, making it computationally efficient. This also removes the dependency of the inference on the performance of the dense semantics estimator, making the system practically more robust.

Fig. 3 shows the flowchat. The *DSAG-Stream* plays the role of a regulator to assist the training of the *MF-Stream*. We achieve this through the corresponding feature fusion between the *DSAG-Stream* and the *MF-Stream*, and the supervision on the fused features. For the *DSAG-Stream*, the input DSAP-images are densely semantically aligned and thus the output features inherit the merits. We intend to leverage the *DSAG-Stream* to drive the *MF-Stream* to learn both global features and part-aware features. For each stream, a small head network with two branches are designed to focus on global and local information respectively.

3.2.1 DSAG-Stream

The *DSAG-Stream* consists of the multi-branch sub-networks (MB-Ns) and a small Head network formed by a global branch and a part branch as shown in Fig. 3. We show the detailed architectures in Table 1.

Multi-branch sub-networks (MB-Ns). Both global information and local details are important and complementary for re-ID [3, 39, 18, 42, 33, 49]. In order to learn local detailed features of the separate region part rather than mixing all parts together, we adopt multi-branch sub-networks (MB-Ns) to learn local feature maps $\mathbf{D}_i \in \mathbb{R}^{h \times w \times c}$ of size $h \times w$ with c channels, $i = 1, 2, \dots, N$, for N merged body part regions, respectively. Note that the N body part regions have no overlap. The N feature maps are concatenated along channels and we have $\mathbf{D} = \mathbf{D}_i|_{i=1}^N = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_N] \in \mathbb{R}^{h \times w \times c_A}$, where $c_A = N \times c$.

For the MB-Ns, we have two levels of merging to progressively merge features from correlated body parts, in order to exploit the symmetry of human body to be better viewpoint robust and to reduce the number of branches. We obtain 8 separate feature maps from MB-Ns as $\mathbf{D}_i|_{i=1}^N$, where $N = 8$. The semantics for a pair of left-right symmetric parts, are semantically aligned in the UV space and we element-wisely add the features in the first level merging. At the second level merging, similarly, we merge the two branches corresponding to the front-back symmetric parts and finally obtain 8 branches as illustrated in Fig. 4.

Head network. It consists of two separate branches which focus on global and local information respectively.

For the global branch, the output feature vector $\mathbf{d}_G \in \mathbb{R}^{2048}$ are obtained by

$$\mathbf{d}_G = \mathcal{P}(\mathcal{H}(\mathbf{D})), \quad (1)$$

where $\mathcal{H}(\cdot)$ denotes an underlying mapping consisting of a few stacked layers; $\mathcal{P}(\cdot)$ denotes the average spatial pooling operation. We take the network architecture of conv5_g as shown in Table 1 for this mapping of $\mathcal{P}(\mathcal{H}(\cdot))$.

For the part branch, the output feature vector $\mathbf{d}_L \in \mathbb{R}^{2048}$ is a concatenation of the feature vectors $\mathbf{d}_{L,i} \in \mathbb{R}^{256}$ of the 8 merged parts, *i.e.*, $\mathbf{d}_L = [\mathbf{d}_{L,1}, \mathbf{d}_{L,2}, \dots, \mathbf{d}_{L,8}]$, with $\mathbf{d}_{L,i}$ obtained by

$$\mathbf{d}_{L,i} = \mathcal{P}(\mathcal{F}(\mathbf{D}_i)), \quad (2)$$

where $\mathcal{F}(\mathbf{X})$ denotes an underlying mapping consisting of a few stacked layers. We take the network architecture of conv5_l as shown in Table 1 for this mapping of $\mathcal{P}(\mathcal{F}(\cdot))$.

3.2.2 MF-Stream

We use the sub-network of ResNet-50 (conv1, conv2_x, conv3_x, and conv4_x) [11] to get the feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times c_A}$. To facilitate the joint learning with the corresponding features from the *DSAG-Stream*, with the feature map \mathbf{F} as input (see Fig. 3), the global features \mathbf{f}_G and

the part-aware features \mathbf{f}_L are learned by the two separate branches of a small Head network. This Head network architecture is similar to the Head network of the *DSAG-Stream*. Note that the features \mathbf{f}_G and \mathbf{f}_L are used for re-ID in our final scheme.

3.2.3 Two-Stream Fusion

We fuse the global features from the two streams by element-wise adding, *i.e.*, $\mathbf{z}_G = \mathbf{f}_G + \mathbf{d}_G$, which enables the joint optimization of the two streams. Similarly, the part-aware features and part features from the two streams are fused as $\mathbf{z}_L = \mathbf{f}_L + \mathbf{d}_L$.

3.2.4 Loss Designs

To train the network, we use the widely-used identification loss (ID Loss), *i.e.*, the cross entropy loss for identification classification, and the ranking loss of triplet loss with batch hard mining [13] (Triplet Loss) as our loss functions.

Considering the noises in the DSAP-images due to semantics estimation errors, and the high complexity of the DensePose model, in our design, we treat the DSAP-images as regulators to drive the semantically aligned feature learning from the original full image, expecting the *MF-Stream* alone to work in inference. We add supervision to the features $\mathbf{f}_G, \mathbf{f}_L$ from the *MF-Stream* and to the fused features $\mathbf{z}_G, \mathbf{z}_L$, respectively as illustrated in Fig. 3. Specifically, for the *MF-Stream*, we add the ID loss for the global feature vector \mathbf{f}_G , and each part-aware feature vector $\mathbf{f}_{L,i}, i = 1, 2, \dots, 8$. For the fused features $\mathbf{z}_G, \mathbf{z}_L$, both the ID loss and triplet loss are added. The loss computed using the fused features makes the gradient back-propagated to the *MF-Stream* be also influenced by the *DSAG-Stream* features, since they contribute to the fused features and the resulting loss. In this way, the *DSAG-Stream* plays the role of regularization by impacting the feature learning of the *MF-Stream* in the training.

To calculate each identification loss, a classifier constructed by two fully connected (FC) layers followed by a SoftMax function is applied to the feature vector to output the classification probability.

4. Experiments

4.1. Datasets and Evaluation Metrics

Market1501 [54] has 32,668 DPM-detected pedestrian image boxes of 1,501 identities, with 12,936 training, 3,368 query and 19,732 gallery images. 751 identities are used for training while the remaining 750 for testing.

CUHK03 [17] consists of 1,467 pedestrians. This dataset provides both manually labeled bounding boxes from 14,096 images and DPM-detected bounding boxes from 14,097 images. We adopt the new training/testing protocol following [58, 57, 12]. In this protocol, 767 identities are used for training and the remaining for testing.

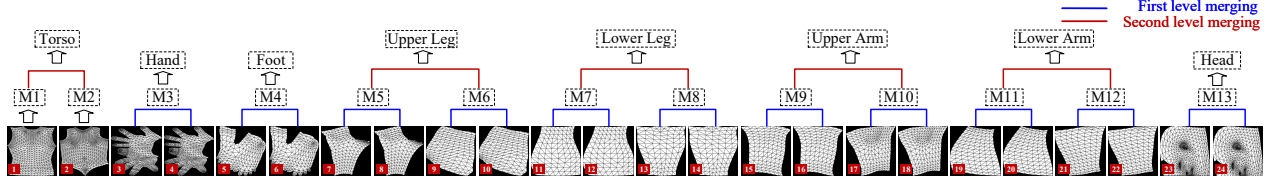


Figure 4. Illustration of two level merging corresponding to the 24 body parts.

Table 1. Detailed architecture of our *DSAG-Stream*. We construct it using similar convolutional layers and building blocks as in ResNet-18 [11]. For conv1, $5 \times 5, 32$ denotes the convolutional kernel size is 5×5 and output channel number is 32. Following the representation style in [11], building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed by conv3.1 and conv4.1 with a stride of 2. #Bran. denotes the number of sub-branches.

Layer name	Parameters	Output size	#Bran.
MB-Ns	conv1	$5 \times 5, 32$	32×32
	conv2	$3 \times 3, 64$	32×32
	conv3_x	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	16×16
	merging	element-wise add	16×16
	conv4_x	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	8×8
	merging	element-wise add	8×8
Head	conv5_g	$\begin{bmatrix} 3 \times 3, 2048 \\ 3 \times 3, 2048 \end{bmatrix} \times 2$	8×8
	Average Pooling	1×1	1×1
	conv5_l	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	8×8
	Average Pooling	1×1	1×1

CUHK01 [16] comprises 3884 images of 971 identities, captured in two disjoint camera views. We adopt the common experimental setting following [1, 5, 50].

DukeMTMC-reID [56] is a subset of Duke Dataset [27] for image-based re-ID. We use the standard training/testing split and evaluation setting following [56, 20]. It contains 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images.

Evaluation Metrics. Following the common practices, we use the cumulative matching characteristics (CMC) at Rank-1 (at least), Rank-5, Rank-10, and mean average precision (mAP) to evaluate the accuracy.

4.2. Implementation Details

Network settings. We take ResNet-50 [11] to build our baseline networks as in some re-ID systems [3, 35, 48, 2]. Similar to [35], the last spatial down-sample operation in the Conv5 layer is removed.

For the *MF-Stream*, we use a part of the ResNet-50 architecture (i.e., conv1, conv2_x to conv4_x) as the sub-network

to obtain the feature map **F**. The weights pretrained on ImageNet [7] are used for initialization. The Head network architecture is similar to the Head network of the *DSAG-Stream* and is randomly initialized. The difference is that the architecture of the global branch in the *MF-Stream* is the same as the network architecture of the conv5_x block in ResNet-50 rather than that in ResNet-18. Each local branch of the *MF-Stream* uses an architecture similar to the global branch but has only 1/8 of the number of channels on each layer. For the *DSAG-Stream*, the network is randomly initialized and trained from scratch.

Data augmentation. We use the commonly used data augmentation strategies of random cropping [41], horizontal flipping and random erasing [59, 41, 38] (with a probability of 0.5) in both the baseline schemes and our schemes.

Optimization. For the triplet loss with batch hard mining [13], we sample $P = 16$ identities and $K = 4$ images [39] for each identity as a mini-batch and the margin parameter is set to 0.3. The ID loss for the *MF-Stream* features, the triplet loss, and ID loss for the fused features are weighed by 0.5, 1.5 and 1.0 respectively. We adopt Adam optimizer with a weight decay of 5×10^{-4} to train our models. We warm up the models for 20 epochs with a linear growth learning rate from 8×10^{-6} to 8×10^{-4} . Then, the learning rate is decayed by a factor of 0.5 for every 40 epochs. We observe that the models converge after training of 320 epochs. All our models are implemented on PyTorch and trained in an end-to-end manner.

4.3. Comparison with State-of-the-Art

We compare our proposed Densely Semantically Aligned re-ID scheme (DSA-reID) with current state-of-the-art methods of four categories in Table 2. *Basic-CNN methods* have similar network structures with a commonly used baseline in deep re-ID systems [53, 48, 3, 39, 38], which learns a global descriptor. *Pose/Part-related methods* leverage the coarse pose/part semantic information to assist re-ID. *Stripe-based methods* divide the full RGB image/feature map into several horizontal stripes to exploit local details. MGN [39] combines the local features of multiple granularities and the global features. *Attention-based methods* [50, 19, 31, 38] jointly learn attention selection and feature representation. Note that we do not implement re-ranking [58] in all our models for clear comparisons.

Table 2. Performance (%) comparisons with the state of the art methods. Bold numbers denote the best performance, while numbers with underlines denote the second best. Superscript * indicates that model is pre-trained on CUHK03 and fine-tuned on CUHK01.

Method		Market1501 (SQ)		CUHK03				CUHK01		DukeMTMC-reID	
				Labeled		Detected					
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	Rank-5	Rank-1	mAP
Basic-CNN (ResNet-50)	IDE(ECCV18) [35]	85.3	68.5	43.8	38.9	-	-	-	-	73.2	52.8
	Gp-reid(Arxiv18) [2]	92.2	81.2	-	-	-	-	-	-	85.2	72.8
Pose/Part -related	Spindle(CVPR17) [49]	76.9	-	-	-	-	-	79.9	94.4	-	-
	PIE(Arxiv17) [53]	78.7	53.9	-	-	-	-	-	-	-	-
	MSCAN(CVPR17) [15]	80.8	57.5	-	-	-	-	-	-	-	-
	PDC(ICCV17) [33]	84.1	63.4	-	-	-	-	-	-	-	-
	Pose Transfer(CVPR18) [22]	87.7	68.9	33.8	30.5	30.1	28.2	-	-	68.6	48.1
	PN-GAN(ECCV18) [26]	89.4	72.6	-	-	-	-	-	-	73.6	53.2
	PSE(CVPR18) [28]	87.7	69.0	-	-	30.2	27.3	67.7	86.6	79.8	62.0
	MGCAM(CVPR18) [32]	83.8	74.3	50.1	50.2	46.7	46.9	-	-	-	-
	MaskReID(Arxiv18) [25]	90.0	75.3	-	-	-	-	84.3	-	78.9	61.9
	Part-Aligned(ECCV18) [34]	91.7	79.6	-	-	-	-	80.7*	94.4*	84.4	69.3
	AACN(CVPR18) [45]	85.9	66.9	-	-	-	-	88.1	96.7	76.8	59.3
	SPReID(CVPR18) [14]	92.5	81.3	-	-	-	-	-	-	84.4	71.0
Stripe -based	AlignedReID(Arxiv17) [48]	91.8	79.3	-	-	-	-	-	-	-	-
	Deep-Person(Arxiv17) [3]	92.3	79.6	-	-	-	-	-	-	80.9	64.8
	PCB+RPP(ECCV18) [35]	93.8	81.6	63.7	57.5	-	-	-	-	83.3	69.2
	MGM(MM18) [39]	95.7	86.9	68.0	67.4	66.8	66.0	-	-	88.7	78.4
Attention -based	DLPAP(ICCV17) [50]	81.0	63.4	-	-	-	-	76.5*	94.2*	-	-
	HA-CNN(CVPR18) [19]	91.2	75.7	44.4	41.0	41.7	38.6	-	-	80.5	63.8
	DuATM(CVPR18) [31]	91.4	76.6	-	-	-	-	-	-	81.8	64.6
	Manacs(ECCV18) [38]	93.1	82.3	69.0	63.9	65.5	60.5	-	-	84.9	71.8
Dense Semantics -based (Ours)	DSA-reID	95.7	87.6	78.9	75.2	78.2	73.1	90.4*	97.8*	86.2	74.3

Market-1501. DSA-reID achieves the best performance. Our method and the second best method MGN [39] have similar performance and both outperform the other methods by at least +1.9%/+4.6% in Rank-1/mAP accuracy. We only show the single query (SQ) results to save space, and a similar trend is observed for the multiple query setting.

CUHK03. DSA-reID outperforms others by a large margin, at least +10.9%/+7.8% in Rank-1/mAP for the labeled setting, and +11.4%/+7.1% in Rank-1/mAP for the detected setting. The images are less blurred than those in other datasets. The semantics estimation is more accurate which greatly helps the training of our networks.

CUHK01. Our method outperforms the current best result by +2.3%/+1.1% in Rank-1/Rank-5 accuracy. Similar to the methods in [5, 50, 34], this result is obtained with pre-training on CUHK03 and fine-tuning on CUHK01. For fair comparisons, we also test our model without a pre-training on CUHK03, it achieves 88.6%/97.1% in Rank-1/Rank-5 respectively, which are also the best.

DukeMTMC-reID. DSA-reID achieves the second best results. The semantics estimation on this dataset is error prone. More than 20% persons cannot be detected on the training images. DSA-reID outperforms all the other approaches except MGN [39] which ensembles local features at multiple granularities. We believe training a better DensePose estimator can further improve the performance.

4.4. Ablation Study

We perform comprehensive ablation studies on the Market-1501 dataset (single query).

Ours vs. baselines. In Table 3, “Baseline” and “Baseline(RE)” denote our baseline schemes without and with random erasing (RE) [2, 38], respectively. Label smoothing regularization [36], which acts as a mechanism to regularize the classifier layer by changing the ground-truth label distribution, has been demonstrated to be effective in recognition [24, 44]. We add label smoothing (LS) to the classification sub-task in the re-ID and denote this baseline as “Baseline(RE+LS)”. It improves the Rank-1/mAP accuracy over “Baseline(RE)” by +1.1%/+2.6%. Besides, we also take our *MF-Stream* only scheme that is built based on “Baseline(RE+LS)” but with a Head network of two branches as our baseline, referred to as “Baseline (Two branches)”.

We denote the proposed densely semantically aligned (DSA) re-ID schemes under different settings/designs with the prefix of “DSA”. “DSA(Two streams fused)” denotes our two stream scheme which takes \mathbf{z}_G and \mathbf{z}_L as the matching features for inference. In inference, the *DSAG-Stream* can be discarded and we refer to it as “DSA-reID(Only *MF-Stream*)”, which takes \mathbf{f}_G and \mathbf{f}_L as the matching features and is our final scheme, also named as “DSA-reID”.

We have the following observations/conclusions.

- 1) Our final scheme achieves significant performance

Table 3. Performance (%) comparisons of baselines and our schemes on the Market-1501 dataset.

Model	mAP	Rank-1	Rank-5	Rank-10
Baseline	76.4	91.2	96.5	97.9
Baseline (RE)	78.6	92.3	97.6	98.3
Baseline (RE+LS)	81.2	93.4	97.8	98.5
Baseline (Two branches)	83.4	94.0	98.0	98.7
DSA-Global(Single)	84.7	94.8	98.2	98.9
DSA-Local(Single)	83.2	94.0	97.9	98.6
DSA-Global(Joint)	87.4	95.6	98.6	99.1
DSA-Local(Joint)	86.5	95.2	98.4	99.0
DSA(Two streams fused)	87.5	95.8	98.4	99.1
DSA-reID(Only MF-Stream)	87.6	95.7	98.4	99.1

improvement, outperforming “Baseline (RE+LS)” by **+2.3%/+6.4%** and “Baseline(Two branches)” by **+1.7%/+4.2%** in Rank-1/mAP accuracy respectively. 2) “DSA-reID(Only *MF-Stream*)” has very similar performance as “DSA(Two streams fused)” but much lower computational complexity.

Global and part-aware/part features. For each stream, we have two branches which focus on global features and part features respectively. We show the analysis in Table 3. 1) “DSA-Global(Single)”/“DSA-Local(Single)” denotes the design with only the global/local branch in our two stream framework for both training and inferencing. “DSA-Global(Single)” outperforms “Baseline(RE+LS)” by **+1.4%/+3.5%** in Rank-1/mAP accuracy. “DSA-Local(Single)” outperforms “Baseline(RE+LS)” by **+0.6%/+2.0%** in Rank-1/mAP accuracy. These demonstrate that our semantic alignment design is very efficient. 2) Since global and part-aware/part features are complementary, our scheme with both the global and part-aware/part branches, “DSA(Two streams fused)”, achieves additional **+1.0%/+2.8%**, and **+1.8%/+4.3%** gain in comparison with “DSA-Global(Single)” and “DSA-Local(Single)” in Rank-1/mAP accuracy respectively. 3) “DSA-Global(Joint)” or “DSA-Local(Joint)” denotes that the inference is based on the features of the global branch or part-aware branch of our scheme “DSA(Two stream fused)”, *i.e.*, \mathbf{z}_G or \mathbf{z}_L . Thanks to the joint training, “DSA-Global/Local(Joint)” significantly outperforms “DSA-Global/Local(Single)”.

Dense vs. coarse semantic alignment. Thanks to the densely semantically aligned representation and our architecture design, our scheme achieves excellent performance. We take the DSAP-images as input to the *DSAG-Stream*. One may wonder about the performance if the cropped body parts without internal fine grained alignment are taken as input to our framework. We conduct an experiment by replacing the 24 DSAP-images by 24 cropped part images (without alignment within a part region) and refer to this scheme as coarsely semantically aligned re-ID, CSA. Table 4 shows the performance comparisons. 1) Our densely semantically aligned scheme significantly outperforms the coarsely

Table 4. Performance (%) comparisons between dense and coarse semantic alignment in our framework on the Market-1501 dataset.

Model	mAP	Rank-1	Rank-5	Rank-10
Baseline (RE+LS)	81.2	93.4	97.8	98.5
CSA(Only <i>MF-Stream</i>)	84.1	94.1	98.1	98.8
DSA(Only <i>MF-Stream</i>)	87.6	95.7	98.4	99.1

semantically aligned scheme by **+1.6%/+3.5%** in Rank-1/mAP accuracy. 2) Our coarsely semantically aligned scheme still outperforms the baselines by a large margin, demonstrating the effectiveness of our architecture design.

Two stream fusion designs. We investigate how to make the *MF-Stream* and the *DSAG-Stream* interact efficiently for joint training and show the comparisons in Table 5. “Concatenation+fc” denotes that for either branch, the features from the *MF-Stream* and the *DSAG-Stream* are concatenated followed by a fully connected layer. “Elem-add” denotes that the features from the *MF-Stream* and the *DSAG-Stream* are element-wisely added. “Concatenation+fc” has poor performance. In contrast, our fusion with element-wise add achieves excellent performance.

Table 5. Performance (%) comparisons on the designs of the two-stream fusion, on the Market-1501 dataset.

Fusion method	mAP	Rank-1	Rank-5	Rank-10
Concatenation+fc	81.6	93.0	97.6	98.6
Elem-add	87.6	95.7	98.4	99.1

5. Conclusion

In this paper, we propose a densely semantically aligned person re-ID framework, intending to address the ubiquitous misalignment problems. Thanks to the estimated dense semantics, it becomes possible to construct the densely semantically aligned part images (DSAP-images) from the 2D image. We design a two stream network consisting of the *MF-Stream* and the *DSAG-Stream*. Considering that the DSAP-images have the inherent densely semantically aligned merits, but are noisy due to semantics estimation error, we treat the *DSAG-Stream* as a regulator to assist the feature learning of the *MF-Stream*, through our fusion and supervision designs. In the inference, only the *MF-Stream* is needed, making the system more computationally efficient and robust. Our scheme achieves the best performance on Market-1501, CUHK03, and CUHK01. On CUHK03, our scheme significantly outperforms the previous methods, by at least **+10.9%/+7.8%** in Rank-1/mAP accuracy.

Acknowledgements

This work was partly supported by the National Key Research and Development Program of China under Grant No. 2016YFC0801001, the National Program on Key Basic Research Projects (973 Program) under Grant 2015CB351803, NSFC under Grant 61571413, 61390514.

References

- [1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 6
- [2] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018. 1, 6, 7
- [3] Xiang Bai, Mingkun Yang, Tengpeng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. *arXiv preprint arXiv:1711.10658*, 2017. 1, 3, 5, 6, 7
- [4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 1, 3
- [5] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 1, 6, 7
- [6] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 3
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6
- [8] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 1
- [9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. *CVPR*, 2018. 2, 3
- [10] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017. 2, 3
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6
- [12] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018. 5
- [13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3, 5, 6
- [14] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 3, 7
- [15] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 1, 3, 7
- [16] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 6
- [17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 5
- [18] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, 2017. 1, 3, 5
- [19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1, 6, 7
- [20] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017. 6
- [21] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, pages 3492–3506. 3
- [22] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018. 7
- [23] Omar Oreifej, Ramin Mehran, and Mubarak Shah. Human identity recognition in aerial images. In *CVPR*, 2010. 3
- [24] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 7
- [25] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018. 3, 7
- [26] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018. 7
- [27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 6
- [28] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 7
- [29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3
- [30] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 1, 3
- [31] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 6, 7
- [32] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 3, 7
- [33] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 1, 3, 5, 7
- [34] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 1, 3, 7
- [35] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. 2018. 1, 6, 7

- [36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 7
- [37] Rahul Rama Vior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 1
- [38] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018. 6, 7
- [39] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. *ACM Multimedia*, 2018. 1, 3, 5, 6, 7
- [40] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1
- [41] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 6
- [42] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM Multimedia*, pages 420–428, 2017. 1, 3, 5
- [43] Lin Wu, Chunhua Shen, and Anton van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 1, 3
- [44] Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. Disturblabel: Regularizing cnn on the loss layer. In *CVPR*, 2016. 7
- [45] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018. 3, 7
- [46] Yuanlu Xu, Liang Lin, Wei-Shi Zheng, and Xiaobai Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 3
- [47] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017. 1, 3
- [48] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Aligned: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 1, 3, 6, 7
- [49] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 1, 3, 5, 7
- [50] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 6, 7
- [51] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 3
- [52] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 3
- [53] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017. 1, 3, 6, 7
- [54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5
- [55] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3
- [56] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 6
- [57] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *TCSVT*, 2018. 1, 5
- [58] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 5, 6
- [59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6