Attention-Aware Compositional Network for Person Re-identification

Jing Xu[1] Rui Zhao[1,2*] Feng Zhu[1] Huaming Wang[1] Wanli Ouyang[3]
[1]SenseNets Technology Limited [2]SenseTime Group Limited [3]The University of Sydney
eudoraxj@gmail.com, zhaorui@sensetime.com, zhufengx@mail.ustc.edu.cn
huamingwang2014@gmail.com, wanli.ouyang@sydney.edu.au

Abstract

Person re-identification (ReID) is to identify pedestrians observed from different camera views based on visual appearance. It is a challenging task due to large pose variations, complex background clutters and severe occlusions. Recently, human pose estimation by predicting joint locations was largely improved in accuracy. It is reasonable to use pose estimation results for handling pose variations and background clutters, and such attempts have obtained great improvement in ReID performance. However, we argue that the pose information was not well utilized and hasn't yet been fully exploited for person ReID.

In this work, we introduce a novel framework called Attention-Aware Compositional Network (AACN) for person ReID. AACN consists of two main components: Pose-guided Part Attention (PPA) and Attention-aware Feature Composition (AFC). PPA is learned and applied to mask out undesirable background features in pedestrian feature maps. Furthermore, pose-guided visibility scores are estimated for body parts to deal with part occlusion in the proposed AFC module. Extensive experiments with ablation analysis show the effectiveness of our method, and state-of-the-art results are achieved on several public datasets, including Market-1501, CUHK03, CUHK01, SenseReID, CUHK03-NP and DukeMTMC-reID.

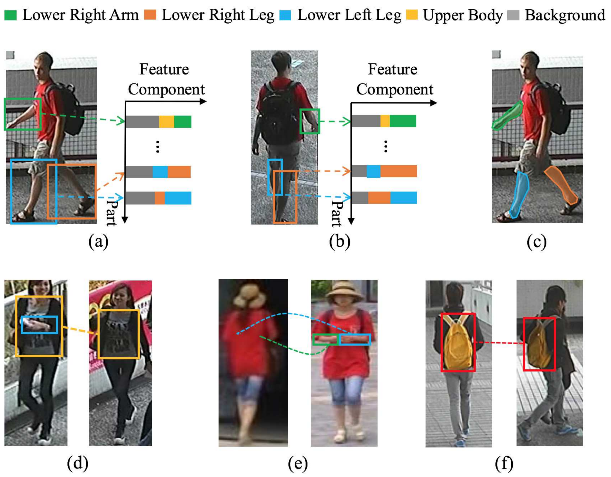

Figure 1. Part alignment challenges in person ReID. (a-c): Describing body parts by bounding boxes may introduce many irrelevant regions from background and other parts. Matching between features extracted from loose boxes in (a) and tight boxes in (b) would deteriorate the matching accuracy. A finer part region representation in (c) would help alleviate this problem. (d-f): The importance of different parts should be adaptively adjusted. Upper body is occluded by forearms in (d), the two forearms are all occluded in (e). Features from occluded part should be eliminated during matching, while salient visual cues like yellow backpack in (f) need to be emphasized.

1. Introduction

Person re-identification (ReID) targets on identifying the same individual across different camera views. Given an image containing a target person (as query) and a large set of images (gallery set), a ReID system is expected to rank the images from gallery according to visual similarity with the query image. It has many important applications in video surveillance by saving large amount of human efforts in exhaustively searching for a target person from large amount of video sequences. For example, finding missing elderly and children, and suspect tracking, *etc*.

Many research works have been proposed to improve the state-of-the-art performance of public ReID benchmarks. However, identifying the same individual across different camera views is still an unsolved task in intelligent video surveillance. It is difficult in that pedestrian images often suffer from complex background clutters, varying illumination conditions, uncontrollable camera settings, severe occlusions and large pose variations.

Viewpoint changes and pose variations cause uncontrolled misalignment between pedestrian images. As the improvement of human pose estimation [2, 6], recent works [37, 49, 54] utilized pose estimation results to align body

[*] Rui Zhao is the corresponding author.

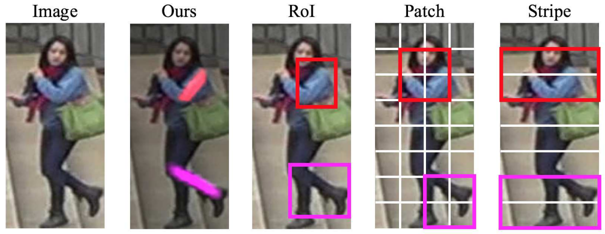

Figure 2. Our Pose-guided Part Attention precisely captures the target parts, excludes background clutter and adjacent part features, while pose-guided rectangular RoIs [49, 37, 54], patches [51, 25], and stripes [34, 1] include extensive noise features.

parts for better matching. Although great improvement in performance was obtained, there are still noticeable problems in these methods. These methods deal with misalignment by extracting features from patches, stripes, or pose-guided region of interest (RoI), where rectangular RoIs often introduce noise from adjacent parts or background in feature and lead to inaccurate matching. For example in Fig. 1(a), features of the right leg is extracted from its bounding box in orange, which includes extensive noise from left leg and background, as shown in the second bar of the histogram. Features of the right arm and left leg also consist of their adjacent parts and background. Furthermore, some body parts have large variations in shape and pose, and rectangular RoI would include inconsistent extent of background clutter and adjacent noise. For example, the right leg is loosely included in the Fig. 1(a) and tightly contained in Fig. 1(b). Matching between part features from loose box and tight box in two camera views would definitely deteriorate the matching accuracy. To deal with these problems, finer silhouettes contouring body parts like in Fig. 1(c) are needed, so that part features can be extracted more precisely, alleviating the influence from background clutters and adjacent noises.

In this work, we propose to use Pose-guided Part Attention instead of rectangular RoI. Pose-guided Part Attention is a confidence map that could precisely capture the target part, and exclude background clutter and adjacent part features, as shown in Fig. 2. Attention-aware part features can be extracted by applying the part attention mask on feature maps, and feature alignment by part is naturally achieved. We will show in experiments that attention-aware part features are more accurate and robust, and the aligned pedestrian features are more discriminative than those proposed in conventional methods.

Occlusion is also a common and severe problem in practical ReID scenario. For example in Fig. 1(d-f), body part may be occluded by other body parts, adjacent persons or things like carrying baggage or trolley. Some observations can be concluded: 1) rigid body parts like head-shoulder, upper torso, lower torso are often partially occluded by adjacent non-rigid parts like upper arms, lower legs, *etc*. 2) non-rigid body parts suffer heavy self-occlusion and are often fully occluded. 3) occlusion by carrying things is not a bad situation, which should be considered as a special part to help re-identification. It would be ideal to weaken features for partially occluded rigid part like the upper body in Fig. 1(d), eliminate features for fully occluded non-rigid part like the forearms in Fig. 1(e), and retain features for carrying things like the backpack in Fig. 1(f). Based on above observations on the occlusion problem, we propose a pose-guided visibility score to measure the occlusion extent for each body part, and it provides image-specific part importance score to decide feature importance in matching. Experimental results show its usefulness in handling occlusion cases.

Based on above motivations, a new Attention-Aware Compositional Network for person re-identification is proposed. The contributions of our work can be summarized in several folds:

- A unified framework named Attention-Aware Compositional Network (AACN) is proposed to deal with misalignment and occlusion problem in person re-identification.
- Pose-guided Part Attention is introduced to estimate finer part attention to exclude adjacent noise. It is designed to capture both rigid and non-rigid body parts simultaneously in a unified framework.
- Visibility score is introduced to measure the occlusion extent for each body part. It provides image-specific part importance scores for Attention-aware Feature Composition.
- Extensive experiments demonstrate that our approach achieve superior performance on several public datasets, including CUHK03 [25], CUHK01 [24], Market-1501 [55], CUHK03-NP [60], DukeMTMC-reID [59] and SenseReID [49].

2. Related Work

2.1. Person Re-identification

There are two categories of methods addressing the problem of person re-identification, namely feature representation and distance metric learning. The first category mainly includes the traditional feature descriptors [53, 52, 51, 27, 7, 31, 35] and deep learning features [43, 41, 42, 12, 25, 40]. These approaches dedicate to design view-invariant representations for person images. The second category [28, 12, 27, 45, 19, 9, 23, 14, 33] mainly targets on learning a robust distance metric to measure the similarity between images.

Pedestrian alignment, matching two person images with their corresponding parts, is of non-trivial importance. Existing ReID methods mainly focus on extracting two types of features, namely global features extracted from the whole image [11, 44] and region features generated from local patches [51, 57, 22]. However, these approaches have not taken the accurate alignment of body regions into consideration. Recently, thanks to the great progress of pose estimation methods [6, 13] and RPN [32], reliable body parts are able to be acquired, which makes it possible to identify individuals via extracted region. For example, Zhao *et al.* [49] proposed Spindle Net, that extracted and fused three level part features. Parts were extracted by PRN. Su *et al.* [37] proposed a Pose-driven Deep Convolutional model (PDC) that utilized Spatial Transformer Network (STN) to localize and crop body regions based on pre-defined centers. Zheng *et al.* [54] introduced to extract Pose Invariant Embedding (PIE) through aligning pedestrians to standard pose. Alignment was done by applying affine transformation to pose estimation results. However, these methods are all based on rigid body regions, which cannot accurately localize human body regions. In our model, non-rigid parts are obtained based on the connectivity between human joints. Thus our model is capable of extracting more precise information for each body part, and handling occlusion issues.

2.2. Human Parsing

Human parsing [17, 10, 47, 26, 15] is related to our work in that parsing results can accurately localize body part. For example, Gong *et al.* [17] imposed joint structure loss to improve segmentation results. Dong *et al.* [15] explored pose information to guide human parsing. However, we choose to generate non-rigid parts based on connectivity of human keypoints rather than human parsing because of the following reason: Existing human parsing methods mainly focus on particular scenarios, such as fashion pictures, and the parsing models often show weak generalization on surveillance data. Human pose is easier to label than parsing, and it can be better generalized to surveillance scenario owing to large variance of the datasets [2, 29].

2.3. Attention based Image Analysis

Since the attention mechanism is effective in understanding images, it has been widely used in various tasks, including machine translation [4], visual question answering [46], object detection [3], semantic segmentation [8], pose estimation [13] and person re-identification [30]. Bahdanau [4] and Ba [3] adopted recurrent neural networks (RNN) to generate the attention map for an image region at each step, and combined information from different steps overtime to make the final decision. Chen *et al.* [8] introduced an attention mechanism that learned to softly weight multi-scale features at each pixel location. Chu *et al.* [13] proposed a multi-context attention model for pose estimation. Inspired by the methods mentioned above, we propose to learn attention map to capture human body part, and align features across different person images by masking with part attentions. Our attention map is learned guided by pose estimation, and it can contour the shape of part more precisely than rectangular RoI. Furthermore, the intensity of part attention infers the visibility of each part, which helps to deal with part occlusion issues.

3. Attention-Aware Compositional Network

The framework of our Attention-Aware Compositional Network (AACN) is illustrated in Fig. 3. AACN consists of two main components: 1) Pose-guided Part Attention (PPA) and 2) Attention-aware Feature Composition (AFC). Given one person image, the proposed PPA module aims to estimate an attention map and a visibility score for each pre-defined body part. Then, part feature alignment and weighted fusion are performed in AFC module, given attention maps and visibility scores from PPA. PPA and AFC are tightly integrated in our framework during both training and testing phases.

The PPA module considers two types of pre-defined body parts, namely, non-rigid parts and rigid parts. Due to the variations in appearance, attentions of these two types of parts are estimated separately. The PPA module is constructed by a two-stage three-branch neural network, which predicts confidence maps of keypoints, attention maps of non-rigid parts, and attention maps of rigid parts in the three branches, respectively. A visibility score is further estimated for each part based on part attention maps.

The AFC module applies the estimated part attention maps to mask the global feature map produced by a base network (GoogleNet [39] is used in this work). The resulting attention-aware part features are then weightedly fused with the guidance from part visibility scores. The final 1024-dimensional feature vector is adopted as the representation of the input person image.

3.1. Pose-guided Part Attention

Part attentions are denoted by normalized part confidence maps, which highlight specific regions of human body in the image. As shown in Fig. 4(a), there are two types of human body parts: rigid parts and non-rigid parts. Limb regions including upper arms, lower arms, upper legs, and lower legs are called non-rigid parts because of drastic pose variations they could occur, while trunk parts of human body including head-shoulder, upper torso, and lower torso are considered to be rigid. Attention maps of the two types of parts are simultaneously learned in a unified form through our proposed Pose-guided Part Attention network.

Inspired by the multi-stage CNN [6] for human pose estimation, we utilize a two-stage network to learn part attentions. The first stage individually predicts non-rigid part attentions $\mathbf{N}$, rigid part attentions $\mathbf{R}$, and keypoint confidence

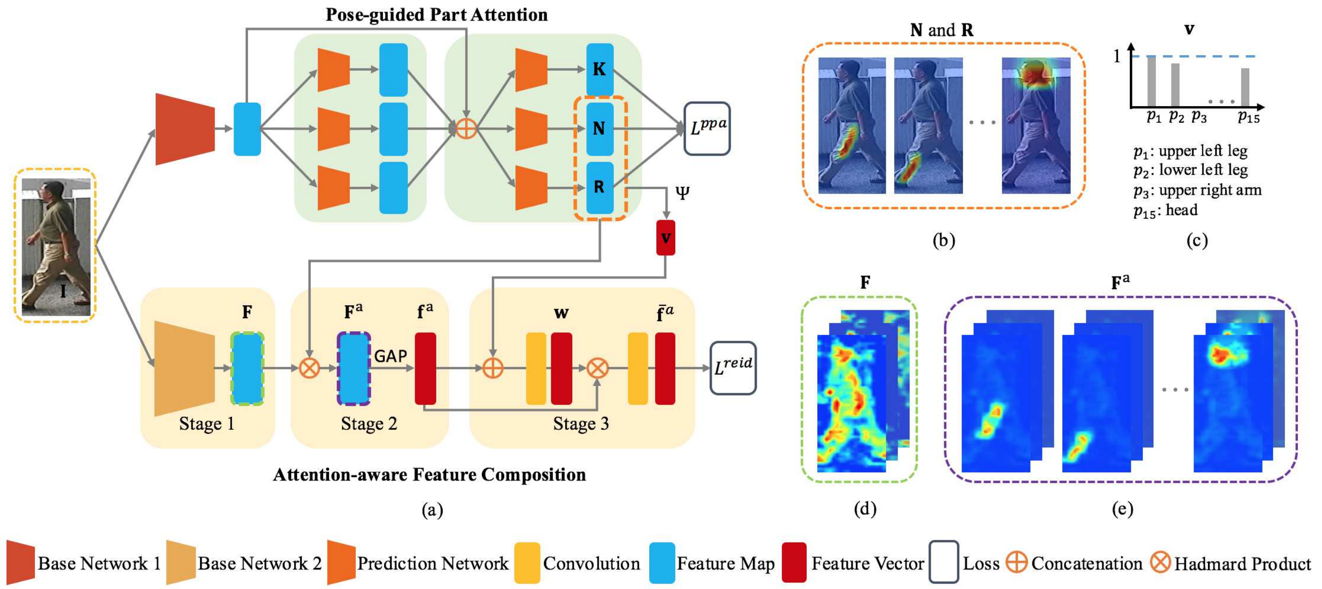

Figure 3. (a) Attention-Aware Compositional Network (AACN). Our framework consists of two main components: Pose-guided Part Attention (PPA) and Attention-aware Feature Composition (AFC). PPA aims to produce attention maps for locating non-rigid parts $\mathbf{N}$ and rigid parts $\mathbf{R}$. AFC is a 3-stage network that aims to extract robust features for pedestrian images. The first stage generates global context feature maps through a base network. Then, the attention-aware feature maps $\mathbf{F}^a$ for body parts are extracted in stage 2 with the guidance from part attentions learned in PPA. In stage 3, the part features are further re-weighted by jointly considering part visibility scores $\mathbf{v}$ and feature salience, resulting in the final compositional weighted feature vector $\bar{\mathbf{f}}^a$. Some visualization are shown in (b) attention maps, (c) visibility scores, (d) global context feature maps, and (e) attention-aware feature maps $\mathbf{F}^a$. (Best viewed in color.)

maps $\mathbf{K}$ by three independent prediction networks,

$$\mathbf{N}^1 = \rho^1(\mathbf{F}^{ppa}),\ \mathbf{R}^1 = \phi^1(\mathbf{F}^{ppa}),\ \mathbf{K}^1 = \psi^1(\mathbf{F}^{ppa}), \quad (1)$$

where $\mathbf{F}^{ppa}$ is the feature map at the 10-th layer of VGG-19 [36]. Keypoint estimation is introduced as an auxiliary task to improve part attention learning in a multi-task learning manner. Then, the second stage refines the attention maps by considering all previous predictions,

$$\begin{aligned} \mathbf{N}^2 &= \rho^2(\mathbf{F} \mid \mathbf{N}^1, \mathbf{R}^1, \mathbf{K}^1), \\ \mathbf{R}^2 &= \phi^2(\mathbf{F} \mid \mathbf{N}^1, \mathbf{R}^1, \mathbf{K}^1), \\ \mathbf{K}^2 &= \psi^2(\mathbf{F} \mid \mathbf{N}^1, \mathbf{R}^1, \mathbf{K}^1). \end{aligned} \quad (2)$$

For network training, supervision is imposed in both stages. The overall objective is

$$L^{ppa}(\rho, \phi, \psi) = \sum_{t=1,2} L^k(\mathbf{K}^t) + \mu_1 L^n(\mathbf{N}^t) + \mu_2 L^r(\mathbf{R}^t), \quad (3)$$

where L^k, L^n and L^r denote the loss function of keypoint confidence map, non-rigid part attention, and rigid part attention, respectively. μ_1 and μ_2 balance the importance of different losses.

Loss for Keypoint Confidence Map $L^k(\mathbf{K})$. Following the definition in MPII dataset [2], 14 keypoints (as shown in Fig. 4(a)) of human body are utilized to guide the learning of part attentions. The i-th channel $\mathbf{K}_i \in \mathbb{R}^{H \times W}$ of keypoint confidence maps $\mathbf{K} \in \mathbb{R}^{H \times W \times C^k}$ predicts the coordinates of the i-th keypoint by giving high confidence values to the true location. The difference between confidence maps $\mathbf{K}$ and ground truth maps $\mathbf{K}_i^*$ are measured by Mean-Square Error (MSE),

$$L^k(\mathbf{K}) = \frac{1}{C^k} \sum_{i=1}^{C^k} ||\mathbf{K}_i^* - \mathbf{K}_i||^2, \quad (4)$$

where, $\mathbf{K}_i^*$ is generated by applying a Gaussian kernel centered at the true location of the i-th keypoint. $C^k = 14$ is the number of keypoints.

Loss for Non-Rigid Part Attention $L^n(\mathbf{N})$. Non-rigid part attentions aim to highlight the corresponding limb parts. Inspired by the Part Affinity Field (PAF) in [6], we define the ground truth non-rigid parts as the connection area of two keypoints to approximate the target limb part. As shown in Fig. 4(b), the p-th non-rigid part is defined as a rectangle area $\mathcal{R}_p^n$ connecting two keypoints with bandwidth σ, and the ground truth non-rigid part attention is represented as

$$\mathbf{N}_p^*(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{R}_p^n, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

where, $\mathbf{x}$ indicates the location on the attention map. The

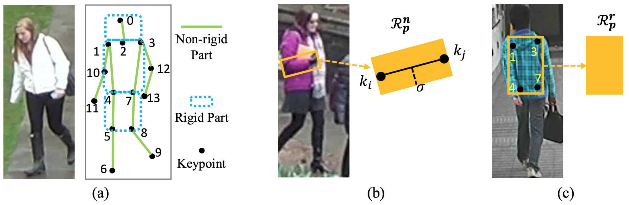

Figure 4. Illustration of Pose-guided Part Attention. (a) The 14 keypoints, 11 non-rigid parts, and 3 rigid parts defined in our work. (b) The ideal non-rigid part attention $\mathbf{R}_p^n$ for right elbow. (c) The ideal rigid part attention $\mathbf{R}_p^r$ for upper body.

errors of non-rigid part attention are measured by MSE

$$L^n(\mathbf{N}) = \frac{1}{C^n}\sum_{p=1}^{C^n} ||\mathbf{N}_p^* - \mathbf{N}_p||^2, \tag{6}$$

where, $C^n = 11$ is the number of non-rigid parts. $\mathbf{N}_p \in \mathbb{R}^{H\times W}$ is the predicted attention map for the p-th part.

Loss for Rigid Part Attention $L^r(\mathbf{R})$. Rigid part attention is introduced to capture body parts that take rigid transformations during changes of view or pose. Three rigid parts are defined in our work, namely head-shoulder, upper torso and lower torso. As shown in Fig. 4(c), each rigid part is defined by a neat rectangle $\mathcal{R}_p^r$, which tightly contains a set of specified keypoints. The set of keypoints for each rigid part are selected as $S_1 = \{0, 1, 3\}$ for head shoulder, $S_2 = \{1, 3, 4, 7\}$ for upper torso, and $S_3 = \{4, 5, 7, 8\}$ for lower torso. Then the ground truth attention map of rigid part p is defined as

$$\mathbf{R}_p^*(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{R}_p^r, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The loss for rigid part attention $L^r(\mathbf{R})$ is computed by accumulating all part losses,

$$L^r(\mathbf{R}, \mathbf{N}) = \frac{1}{C^r}\sum_{p=1}^{C^r} ||\mathbf{R}_p^* - \hat{\mathbf{R}}_p||^2, \tag{8}$$

where $C^r = 3$ is the number of rigid parts.

Part Visibility Score. The intensities in an attention map indicate the visibility of the part at each location. Motivated by this observation, we can define a global visibility score for each part as

$$v_p = \sum_{x,y} |\mathbf{R}_p(x, y)|, \text{ or } v_p = \sum_{x,y} |\mathbf{N}_p(x, y)|, \tag{9}$$

where (x, y) indicates the location on the attention map. Global visibility scores help to balance the importance between different body parts for person identification.

3.2. Attention-aware Feature Composition

Based on human part attentions introduced in Sec. 3.1, in this section, we propose an Attention-aware Feature Composition (AFC) which learns to align and re-weight features of body parts. AFC comprises of three main stages, namely Global Context Network (GCN), Attention-Aware Feature Alignment, and Weighted Feature Composition. At the first stage, a pedestrian image is input in GCN to extract global features, which are further fed into stage 2 together with part attentions estimated from the same input image. Stage 2 generates part-attention-aware features and concatenates them in the same order for all images. The aligned features are then re-weighted by visibility scores in stage 3 to generate the final compositional feature vector. The whole flowchart of AFC is shown in Fig. 3.

Stage 1: Global Context Network (GCN). GCN serves as a base network for global pedestrian feature extraction. Following [16], we build GCN based on the standard GoogleNet [39]. To reduce the computation cost for following stages of AFC, we add one more 256-channel 3×3 convolution layer after the layer "inception_5b/output" of GoogleNet, and then feed the 256-channel feature maps to the next stage. To better fit the aspect ratio of pedestrian images, the input image size is changed from 224×224 to 448×192, and the resulting feature maps at the last convolution layer changes from 7×7 to 14×6 accordingly.

GCN is independently trained first, and then jointly fine-tuned with following stages of AFC. In independent training, GCN is initialized with the GoogleNet model pretrained on ImageNet, and the newly added convolution layer is randomly initialized. Subsequent to the global average pooling layer at the end of GCN, identification loss and verification loss similar to [16] are applied to guide the global context learning. In joint fine-tuning, these two losses are retained to preserve high-quality global context features of pedestrian images.

Stage 2: Attention-Aware Feature Alignment. Global context features suffer from body part misalignment between pedestrian images. Based on human part attention introduced in Section 3.1, we propose a simple and effective scheme to achieve attention-aware feature alignment. Specifically, we extract attention-aware feature maps by applying Hadamard Product between global feature maps and each human part attention map, and the resulting feature maps are globally average pooled and concatenated to produce an aligned feature vector. Formally, we formulate the attention-aware feature alignment scheme as:

$$\mathbf{f}^a = Concat(\{\mathbf{f}_p\}_{p=1}^P), \ \mathbf{f}_p = \sigma_{gap}(\mathbf{F}_p^a), \tag{10}$$

$$\mathbf{F}_p^a = \mathbf{F} \circ \bar{\mathbf{M}}_p, \ \bar{\mathbf{M}}_p = \frac{\mathbf{M}_p}{\max(\mathbf{M}_p)}, \tag{11}$$

where, $\mathbf{M}_p \in \{\mathbf{N}_p, \mathbf{R}_p\}$ is the attention map for body parts,

$\bar{\mathbf{M}}_p$ is the normalized attention map, $\max(\mathbf{M}_p)$ indicates the maximum value in $\mathbf{M}_p$, $\mathbf{F}$ is the 256-channel global feature map produced by GCN, $\circ$ denotes the Hadamard Product operator which performs element-wise product on two matrices or tensors, and $\mathbf{F}_p^a$ denotes the attention-aware feature map for part p. $\sigma_{gap}(\cdot)$ is the global average pooling function, and $Concat(\cdot)$ represents concatenation operation on part feature vectors. Global feature maps $\mathbf{F}$ are masked with attention maps for P times, in which manner P sets of attention-aware feature maps are produced, *i.e.* $\{\mathbf{F}_p^a\}_{p=1}^P$. Each set of attention-aware feature maps $\mathbf{F}_p^a$ possess the feature information of the corresponding body part while preserving global context information of the image.

The attention-aware feature maps $\mathbf{F}_p^a$ are passed through a global average pooling $\sigma_{gap}(\cdot)$ to generate summarized feature vector $\mathbf{f}_p$ for part p. These summarized part features are further concatenated to produce final attention-aware aligned feature vector $\mathbf{f}^a$ for the input pedestrian image.

Stage 3: Weighted Feature Composition. Since pedestrians vary in pose, suffer from occlusions, and may contain some relatively salient parts, the importance of each part should be adaptively adjusted in matching. Motivated by these observations, we introduce a weight vector $\mathbf{w}$ to measure the part importance. The weight vector is estimated by jointly considering part visibility and feature salience. As shown in Fig. 3, visibility scores and the attention-aware aligned feature vector are concatenated, and fed into a fully connected layer (implemented as 1×1 convolutions) to generate $\mathbf{w}$. Then the compositional weighted feature vector is generated as $\bar{\mathbf{f}}^a = Conv(Concat(\{\mathbf{w}_p \cdot \mathbf{f}_p\}_{p=1}^P))$, where $Conv$ indicates a convolution layer.

Overall, our framework integrated the PPA and ACF to extract feature for each input person image. In person ReID applications, given a query image, its feature is matched with that of each image in gallery set to generate distance score. Gallery images are sorted according to ascending order of the distance scores. Then the target person can be found among top-ranked gallery images.

3.3. Implementation Details

In AFC, GoogleNet is utilized as base network for global context feature extraction, and two additional "1x1" convolution layers are used for part weight estimation and final feature fusion, respectively. AACN is trained progressively. First, PPA and GCN are trained independently. PPA is trained with losses for part attention and pose estimation. GCN is trained with reid loss. Then, by fixing PPA and GCN, the parameters for feature weighting and composition in AFC are trained with reid loss. Finally, all modules are jointly fine-tuned.

Dataset	#ID	#train/test IDs	det./lab.
CHUK03 [25]	1467	1160/100	det.&lab.
CUHK01 [24]	971	486/485	lab.
Market-1501 [55]	1501	751/750	det.
CUHK03-NP [60]	1467	767/700	det.&lab.
DukeMTMC-reID [59]	1812	702/702	lab.
SenseReID [49]	1717	0/1717	det.

Table 1. Details of the datasets evaluated in our experiments. Bounding box labels of these datasets can be detected (det.) or manually labeled (lab.).

CUHK03 (labeled)	R-1	R-5	R-10	R-20
NFST [48]	62.55	90.05	94.80	98.10
JSTL [43]	75.30	-	-	-
Transfer [16]	85.40	-	-	-
SVDNet [38]	81.80	-	-	-
Quadruplet [9]	75.53	95.15	99.16	-
PAR [50]	85.40	97.60	99.40	99.90
Spindle [49]	88.50	97.80	98.60	99.20
PDC [37]	88.70	98.61	99.24	99.67
AACN (Ours)	**91.39**	**98.89**	**99.48**	**99.75**

Table 2. Comparison results on CUHK03 (labeled).

CUHK03 (detected)	R-1	R-5	R-10	R-20
NFST [48]	54.70	84.75	94.80	95.20
Transfer [16]	84.10	-	-	-
DPFL [11]	82.00	-	-	-
PAR [50]	81.60	97.30	98.40	99.50
PIE [54]	67.10	92.20	96.60	98.10
PDC [37]	78.29	94.83	97.15	98.43
AACN (ours)	**89.51**	**97.68**	**98.77**	**99.34**

Table 3. Comparison results on CUHK03 (detected).

CUHK01	R-1	R-5	R-10	R-20
NFST [48]	69.09	86.90	91.77	95.39
JSTL [43]	66.60	-	-	-
Transfer [16]	77.00	-	-	-
Quadruplet [9]	62.55	83.44	89.71	-
PAR [50]	75.00	93.50	95.70	97.70
Spindle [49]	79.90	94.40	97.10	98.60
AACN (Ours)	**88.07**	**96.67**	**98.16**	**99.10**

Table 4. Comparison results on CUHK01.

4. Experiments

In this section, the performance of AACN is compared with state-of-the-art methods on several public datasets. And then detailed ablation analysis is conducted to validate the effectiveness of AACN components.

4.1. Datasets and Protocols

Our proposed AACN framework is evaluated on several public person ReID datasets, as listed in Table 1. For fair comparison, we follow the official evaluation

Market-1501	Single Query		Multiple Query	
	R-1	mAP	R-1	mAP
NFST [48]	61.02	35.68	71.56	46.03
PAN [58]	82.81	63.35	88.18	71.72
SVDNet [38]	82.30	62.10	-	-
PAR [50]	81.00	63.40	-	-
Spindle [49]	76.90	-	-	-
PIE [54]	78.65	53.87	-	-
PDC [37]	84.14	63.41	-	-
AACN (Ours)	85.90	66.87	89.78	75.10
AACN+R.E. (Ours)	**88.69**	**82.96**	**92.16**	**87.32**

Table 5. Comparison results on Market-1501. Rank-1 accuracy (%) and mAP (%) are shown. R.E. : re-ranking method from [60].

CUHK03-NP	labeled		detected	
	R-1	mAP	R-1	mAP
PAN [58]	36.86	35.03	36.29	34.00
DPFL [11]	43.00	40.50	40.70	37.00
SVDNet [38]	40.93	37.83	41.50	37.30
AACN (Ours)	**81.86**	**81.61**	**79.14**	**78.37**

Table 6. Comparison results on CUHK03-NP.

DukeMTMC-reID	R-1	mAP
OIM [44]	68.10	-
LSRO [59]	67.68	47.13
PAN [58]	71.59	51.51
SVDNet [38]	76.70	56.80
AACN (Ours)	**76.84**	**59.25**

Table 7. Comparison results on DukeMTMC-reID.

SenseReID	R-1	R-5	R-10	R-20
JSTL [43]	23.00	34.80	40.60	46.30
Spindle [49]	34.60	52.70	59.90	66.70
AACN (Ours)	**41.37**	**58.65**	**64.71**	**72.16**

Table 8. Cross-dataset evaluation on SenseReID.

protocols of each dataset. For CUHK03, CUHK01 and SenseReID, Cumulated Matching Characteristics (CMC) at rank-1, rank-5, rank-10 and rank-20 are compared between different approaches. For Market-1501, CUHK03-NP and DukeMTMC-reID, rank-1 identification rate and mean Average Precision (mAP) are reported.

4.2. Comparisons with State-of-the-Arts

The proposed AACN is compared with recent approaches with state-of-the-art performance. These approaches are categorized into two sets according to whether human pose information is utilized. One set is pose-irrelevant, which includes the null space semi-supervised learning method (NFST) [48], the domain guided dropout method (JSTL) [43], the deep transfer learning method (Transfer) [16], the Singular Vector Decomposition method (SVDNet) [38], the Online Instance Matching (OIM) method [44], the quadruplet loss method (Quadruplet) [9], the multi-scale representation (DPFL) [11], the pedestrian alignment network (PAN) [58], the Part-Aligned Representation (PAR) [50]. The other set introduces explicit pose estimation results into ReID, which includes the Spindle Net (Spindle) [49], the Pose-driven Deep Convolutional model (PDC) [37], and the Pose Invariant Embedding (PIE) [54].

The experimental results are presented in Table 2, 3, 4, 5, 6 and 7. It shows that our proposed AACN outperforms state-of-the-art approaches on all datasets. Specifically, when compared with the second best approach on each dataset, our AACN achieves 2.69%, 5.41%, 8.17%, 4.55% and 40.93% rank-1 accuracy improvement on CUHK03 (labeled), CUHK03 (detected), CUHK01, Market-1501 and CUHK03-NP (labeled), respectively. Though our AACN is very close to SVDNet [38] in rank-1 accuracy on DukeMTMC-reID dataset, the improvement in mAP metric (+2.45%) is still significant.

We also evaluate the generalization ability of our AACN on SenseReID dataset [49]. Following Spindle [49], we merge the training set of Market-1501 [55], CUHK01 [24], CUHK02 [23], CUHK03 [25], PSDB [43], Shinpuhkan [21], PRID [20], VIPeR [18], 3DPeS [5] and i-LIDS [56] for training, and then test on SenseReID. As shown in Table 8, AACN achieves 41.37% accuracy at rank-1, significantly outperforms Spindle which has an accuracy of 34.60%.

4.3. Ablation Analysis

Base Network. The performance of ReID approaches is influenced by base network structures, and different approaches may choose different backbones. As listed in Table 9, our approach is comparable to Spindle [49] and PDC [37] in base network size, but much smaller in overall model size. To better compare with previous works on exploiting pose information, we also experiment by replacing base network of our AACN with the one used by Spindle. It shows that our AACN still outperforms Spindle under the same base network structure.

Pose-guided Part Attention. The localization accuracy of PPA is compared with rectangular RoI [49] method on PASCAL-Person-Part dataset [10]. Since some body parts are not available on this dataset, we choose head, left arm (L.Arm), right arm (R.Arm), left leg (L.Leg), right leg (R.Leg) for comparison. Localization accuracy is measured as the Intersection-over-Union (IOU) between predictions and ground truth parsing labels. Both methods are trained on the same MPII dataset. As shown in Table 10, our proposed PPA is more accurate than RoI in part localization.

Furthermore, we evaluate the performance of different part localization methods on the person ReID task. Specifically, RoI [49] and parsing results from Parsing [17] are compared by replacing the PPA module in our AACN framework. The results are shown in Table 11. On

Method	Base model	# inception	# param (base)	# param (overall)	CUHK03(labeled)		CUHK03(detected)		Market-1501(SQ)	
					base	overall	base	overall	base	overall
PIE [54]	AlexNet	-	57M	114M	-	-	58.80	62.60	55.49	65.68
	ResNet-50	-	23M	46M	-	-	54.80	61.50	73.02	78.65
PDC [37]	GoogleNet-PDC	10	10M	14M	79.83	88.70	71.89	78.29	76.22	84.14
Spindle [49]	GoogleNet-Spindle	6	6M	44M	-	88.50	-	-	72.10	76.90
AACN (Ours)	GoogleNet-Spindle	6	6M	8M	84.01	89.16	81.70	86.65	71.41	81.95
	GoogleNet	9	6M	8M	86.11	91.39	83.78	89.51	79.63	85.90

Table 9. Comparison with human pose based approaches. Rank-1 accuracy (%) is reported.

Part	Head	L.Arm	R.Arm	L.Leg	R.Leg
RoI[49]	**26.53**	13.30	13.25	13.89	14.08
PPA	23.51	**25.19**	**23.12**	**16.63**	**16.21**

Table 10. Part localization accuracy. Part IoUs are given.

CUHK03	labeled		detected	
	R-1	R-5	R-1	R-5
AFC+RoI[49]	89.88	97.97	86.44	97.33
AFC+Parsing[17]	85.49	97.38	82.92	95.66
AFC+PPA	**90.58**	**98.65**	**87.98**	**97.64**

Table 11. Comparison of part localization methods for ReID.

CUHK03	labeled		detected	
	R-1	R-5	R-1	R-5
GCN	86.11	98.18	83.78	96.86
AFC_rigid	87.35	97.98	84.88	96.70
AFC_non-rigid	89.89	98.64	86.87	97.19
AFC_PPA	**90.58**	**98.65**	**87.98**	**97.64**

Table 12. Effectiveness of Attention-aware Feature Composition. "GCN" uses global features for person ReID. "AFC_rigid" only extracts features from rigid parts.

Rank-1 accuracy (%)	CUHK03		Market-1501	
	labeled	detected	SQ	MQ
AACN-w/o-v	90.58	87.98	86.58	90.29
AACN-v	**91.39**	**89.51**	**88.69**	**92.16**

Table 13. Effectiveness of visibility score.

CUHK03 (detected) set, "AFC+RoI" is 1.54% lower than "AFC+PPA" in rank-1 accuracy since it includes noise from adjacent areas, and "AFC+Parsing" is 5.06% lower due to domain difference.

Attention-aware Feature Composition. AFC is a key module in our proposed AACN framework. As shown in Table 12, "GCN" extracts features globally over the image, and achieves 86.11% and 83.78% rank-1 accuracy on CUHK03 (labeled) and CUHK03 (detected), respectively. When using AFC, our proposed "AFC_PPA" improves the accuracies to 90.58% and 87.98%. Using rigid parts only ("AFC_rigid") or non-rigid parts only ("AFC_non-rigid") still outperforms "GCN", and these two types of body parts are complementary to each other. More qualitative results are shown in Fig. 5. Even though a query image looks similar to an imposter globally, body part alignment and local feature aggregation in Attention-aware Feature Composition could effectively distinguish them.

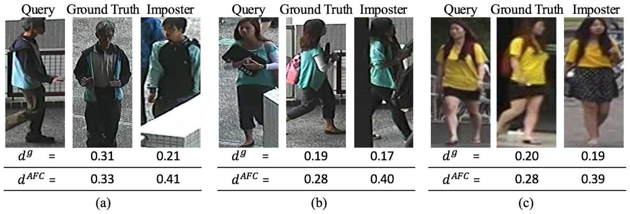

Figure 5. Comparison between global features from GCN and aligned part features from AFC. d^g is the distance computed on global features, and d^{AFC} is the distance computed on the compositional features produced by AFC on Pose-guided Part Attentions. The query images are more similar with the imposters in the global context feature space, but AFC effectively distinguishes them by (a) hair color, (b) upper arm, (c) shorts.

Visibility Score. The effectiveness of visibility score is evaluated on CUHK03 and Market-1501 dataset. As shown in Table 13, weighting part features with visibility scores significantly increase the rank-1 accuracy by 1.53% and 2.11% on CUHK03 (detected) and Market-1501 (SQ).

5. Conclusion

In this paper, we propose an Attention-Aware Compositional Network (AACN) to deal with the misalignment and occlusion problem in person re-identification. AACN is composed of two main components, namely, Pose-guided Part Attention (PPA) and Attention-aware Feature Composition (AFC), where PPA is to estimate finer part attention for preciser feature extraction. Also, visibility score is introduced to measure the occlusion extent, and to guide AFC to learn more robust feature for matching. Extensive experiments with ablation analysis demonstrate that our AACN achieves superior performance than state-of-the-art methods on several public datasets.

Acknowledgements. This work is supported by SenseTime Group Limited.

References

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

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[3] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv:1412.7755*, 2014.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[5] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2016.

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

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[9] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

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

[11] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *CVPR*, 2017.

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

[13] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *arXiv:1702.07432*, 2017.

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

[15] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014.

[16] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv:1611.05244*, 2016.

[17] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017.

[18] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007.

[19] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. In *ICCV*, 2017.

[20] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[21] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, volume 5, 2014.

[22] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[23] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

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

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

[26] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015.

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

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

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[30] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 2017.

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

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

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

[34] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li. Constrained deep metric learning for person re-identification. *arXiv:1511.07545*, 2015.

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

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[37] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[38] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

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

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

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

[42] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv:1601.07255*, 2016.

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

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[45] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[46] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.

[47] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

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

[49] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[50] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

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

[52] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

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

[54] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv:1701.07732*, 2017.

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

[56] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, volume 2, 2009.

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

[58] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv:1707.00408*, 2017.

[59] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[60] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.