2017 IEEE International Conference on Computer Vision

Pose-driven Deep Convolutional Model for Person Re-identification

Chi Su[1*†], Jianing Li[1*], Shiliang Zhang[1], Junliang Xing[2], Wen Gao[1], Qi Tian[3]
[1]School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249-1604, USA
`suchi@kingsoft.com,kaneiri1024@gmail.com,{slzhang.jdl,wgao}@pku.edu.cn,`
`jlxing@nlpr.ia.ac.cn,qi.tian@utsa.edu`

Abstract

Feature extraction and matching are two crucial components in person Re-Identification (ReID). The large pose deformations and the complex view variations exhibited by the captured person images significantly increase the difficulty of learning and matching of the features from person images. To overcome these difficulties, in this work we propose a Pose-driven Deep Convolutional (PDC) model to learn improved feature extraction and matching models from end to end. Our deep architecture explicitly leverages the human part cues to alleviate the pose variations and learn robust feature representations from both the global image and different local parts. To match the features from global human body and local body parts, a pose driven feature weighting sub-network is further designed to learn adaptive feature fusions. Extensive experimental analyses and results on three popular datasets demonstrate significant performance improvements of our model over all published state-of-the-art methods.

1. Introduction

Person Re-Identification (ReID) is an important component in a video surveillance system. Here person ReID refers to the process of identifying a probe person from a gallery captured by different cameras, and is generally deployed in the following scenario: given a probe image or video sequence containing a specific person under a certain camera, querying the images, locations, and time stamps of this person from other cameras.

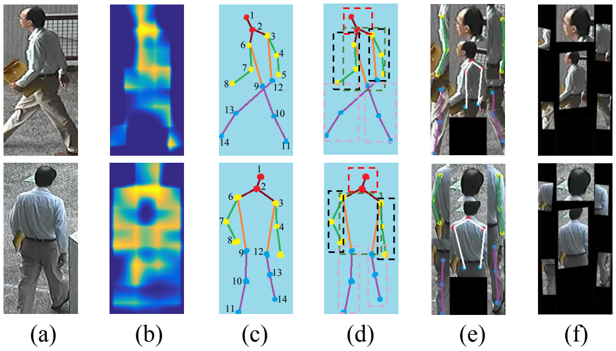

Figure 1. Illustration of part extraction and pose normalization in our Feature Embedding sub-Net (FEN). Response maps of 14 body joints (b) are first generated from the original image in (a). 14 body joints in (c) and 6 body parts in (d) can hence be inferred. The part regions are firstly rotated and resized in (e), then normalized by Pose Transform Network in (f).

Despite decades of studies, the person ReID problem is still far from being solved. This is mainly because of challenging situations like complex view variations and large pose deformations on the captured person images. Most of traditional works try to address these challenges with the following two approaches: (1) representing the visual appearance of a person using customized local invariant features extracted from images [11, 6, 33, 29, 60, 51, 64, 44] or (2) learning a discriminative distance metric to reduce the distance among features of images containing the same person [32, 9, 17, 36, 55, 23, 54, 30, 26, 65, 50, 3, 27, 4, 39, 28, 10, 37, 59]. Because the human poses and viewpoints are uncontrollable in real scenarios, hand-coded features may be not robust enough to pose and viewpoint variations. Distance metric is computed for each pair of cameras, making distance metric learning based person ReID suffers from the $\mathcal{O}^2$ computational complexity.

In recent years, deep learning has demonstrated strong model capabilities and obtains very promising performances in many computer vision tasks [24, 14, 31, 38, 8]. Meanwhile, the release of person ReID datasets like CUHK 03 [25], Market-1501 [63], and MARS [61], both of which contain many annotated person images, makes training deep

---

[*] indicates equal contribution.

[†] Chi Su finished this work when he was a Ph.d candiadate in Peking University, now he has got his Ph.d degree and is working in Beijing Kingsoft Cloud Network Technology Co.,Ltd, No.33,xiaoying Rd.W., HaiDian Dist., Beijing 100085, China

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.427

models for person ReID feasible. Therefore, many researchers attempt to leverage deep models in person ReID [1, 10, 53, 46, 42, 61, 13, 56, 43, 57]. Most of these methods first learn a pedestrian feature and then compute Euclidean distance to measure the similarity between two samples. More specifically, existing deep learning based person ReID approaches can be summarized into two categories: 1) use Softmax Loss with person ID labels to learn a global representation [1, 10, 53, 46, 42, 61, 13], and 2) first learn local representations using predefined rigid body parts, then fuse the local and global representations [5, 47, 40] to depict person images. Deep learning based methods have demonstrated significant performance improvements over the traditional methods. Although these approaches have achieved remarkable results on mainstream person ReID datasets, most of them do not consider pose variation of human body.

Because pose variations may significantly change the appearance of a person, considering the human pose cues is potential to help person re-identification. Although there are several methods [5, 47, 40] that segment the person images according to the predefined configuration, such simple segmentation can not capture the pose cues effectively. Some recent works [62, 16] attempt to use pose estimation algorithms to predict human pose and then train deep models for person ReID. However, they use manually cropped human body parts and their models are not trained from end to end. Therefore, the potential of pose information to boost the ReID performance has not been fully explored.

To better alleviate the challenges from pose variations, we propose a Pose-driven Deep Convolutional (PDC) model for person ReID. The proposed PDC model learns the global representation depicting the whole body and local representations depicting body parts simultaneously. The global representation is learned using the Softmax Loss with person ID labels on the whole input image. For the learning of local representations, a novel Feature Embedding sub-Net (FEN) is proposed to learn and readjust human parts so that parts are affine transformed and re-located at more reasonable regions which can be easily recognizable through two different cameras. In Feature Embedding sub-Net, each body part region is first automatically cropped. The cropped part regions are hence transformed by a Pose Transformation Network (PTN) to eliminate the pose variations. The local representations are hence learned on the transformed regions. We further propose a Feature Weighting sub-Net (FWN) to learn the weights of global representations and local representations on different parts. Therefore, more reasonable feature fusion is conducted to facilitate feature similarity measurement.

Some more detailed descriptions to our local representation generation are illustrated in Fig.1. Our method first locates the key body joints from the input image, *e.g.*, illustrated in Fig.1 (c). From the detected joints, six body parts are extracted, *e.g.*, shown in Fig.1(d). As shown in Fig.1(e), those parts are extracted and normalized into fixed sizes and orientations. Finally, they are fed into the Pose Transformation Network (PTN) to further eliminate the pose variations. With the normalized and transformed part regions, *e.g.*, Fig.1 (f), local representations are learned by training the deep neural network. Different parts commonly convey different levels of discriminative cues to identify the person. We thus further learn weights for representations on different parts with a sub-network.

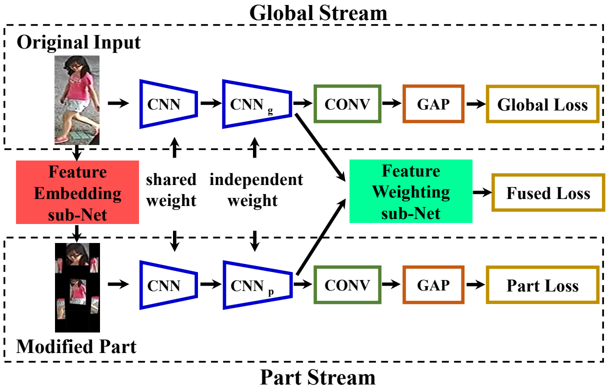

Figure 2. Flowchart of Pose-driven Deep Convolutional (PDC) model. Feature Embedding sub-Net (FEN) leverages human pose information and transforms a global body image into an image containing normalized part regions. Feature Weighting sub-Net (FWN) automatically learns the weights of the different part representations to facilitate feature similarity measurement.

Most of current deep learning based person ReID works do not consider the human pose cues and the weights of representation on different parts. This paper proposes a novel deep architecture that transforms body parts into normalized and homologous feature representations to better overcome the pose variations. Moreover, a sub-network is proposed to automatically learn weights for different parts to facilitate feature similarity measurement. Both the representation and weighting are learned jointly from end to end. Since pose estimation is not the focus of this paper, the used pose estimation algorithm, *i.e.*, Fully Convolutional Networks(FCN) [31] based pose estimation method is simple and trained independently. Once the FCN is trained, it is incorporated in our framework, which is hence trained in an end-to-end manner, *i.e.*, using images as inputs and person ID labels as outputs. Experimental results on three popular datasets show that our algorithm significantly outperforms many state-of-the-art ones.

2. Related Work

Traditional algorithms perform person re-identification through two ways: (a) acquiring robust local features visually representing a person's appearance and then encoding

Table 1. Detailed structure of the proposed Pose-driven Deep Convolutional (PDC) model.

type	share weight	patch size /stride	output size	depth	#1×1	#3×3 reduce	#3×3	double#3×3 reduce	double #3×3	pool proj
data	-	-	512× 256× 3	-	-	-	-	-	-	-
convolution	Yes	7× 7/2	256× 128× 64	1	-	-	-	-	-	-
max pool	-	3× 3/2	128× 64× 64	0	-	-	-	-	-	-
convolution	Yes	3× 3/1	128× 64× 192	1	-	64	192	-	-	-
max pool	-	3× 3/2	64× 32× 192	0	-	-	-	-	-	-
inception(3a)	Yes	-	64× 32× 256	3	64	64	64	64	96	avg+32
inception(3b)	Yes	-	64× 32× 320	3	64	64	96	64	96	avg+64
inception(3c)	Yes	stride 2	32× 16× 576	3	0	128	160	64	96	max+pass through
inception(4a)	Yes	-	32× 16× 576	3	224	64	96	96	128	avg+128
inception(4b)	Yes	-	32× 16× 576	3	192	96	128	96	128	avg+128
inception(4c)	Yes	-	32× 16× 576	3	160	128	160	128	160	avg+128
inception(4d)	Yes	-	32× 16× 576	3	96	128	192	160	192	avg+128
inception(4e)	Yes	stride 2	16× 8× 1024	3	0	128	192	192	256	max+pass through
inception(5a)	No	-	16× 8× 1024	3	352	192	320	160	224	avg+128
inception(5b)	No	-	16× 8× 1024	3	352	192	320	192	224	max+128
convolution	No	1× 1/1	16× 8× class num	1	-	-	-	-	-	-
ave pool	-	global pooling	1× 1× class num	0	-	-	-	-	-	-

them [11, 6, 33, 29, 60, 51, 64]; (b) closing the gap between a person's different features by learning a discriminative distance metric [32, 9, 17, 36, 55, 23, 54, 30, 26, 65, 50, 3, 27, 4, 39, 28, 10, 37, 59]. Some recent works [1, 10, 53, 46, 42, 61, 13, 5, 47, 40, 62, 16] have started to apply deep learning in person ReID and achieved promising performance. In the following, we briefly review recent deep learning based person ReID methods.

Deep learning is commonly used to either learn a person's representation or the distance metric. When handling a pair of person images, existing deep learning methods usually learn feature representations of each person by using a deep matching function from convolutional features [1, 25, 53, 13] or from the Fully Connected (FC) features [58, 40, 61]. Apart from deep metric learning methods, some algorithms first learn image representations directly with the Triplet Loss or the Siamese Contrastive Loss, then utilize Euclidean distance for comparison [48, 5, 10, 46]. Wang *et al.* [48] use a joint learning framework to unify single-image representation and cross-image representation using a doublet or triplet CNN. Shi *et al.* [40] propose a moderate positive mining method to use deep distance metric learning for person ReID. Another novel method [40] learns deep attributes feature for ReID with semi-supervised learning. Xiao *et al.* [53] train one network with several person ReID datasets using a Domain Guided Dropout algorithm.

Predefined rigid body parts are also used by many deep learning based methods [5, 47, 40] for the purpose of learning local pedestrian features. Different from these algorithms, our work and the ones in [62, 16] use more accurate human pose estimation algorithms to acquire human pose features. However, due to the limited accuracy of pose estimation algorithms as well as reasons like occlusion and lighting change, pose estimation might be not accurate enough. Moreover, different parts convey different levels of discriminative cues. Therefore, we normalize the part regions to get more robust feature representation using Feature Embedding sub-Net (FEN) and propose a Feature Weighting sub-Net (FWN) to learn the weight for each part feature. In this way, the part with high discriminative power can be identified and emphasized. This also makes our work different from existing ones [62, 16], which do not consider the inaccuracy of human poses estimation and weighting on different parts features.

3. Pose-driven Deep ReID Model

In this section, we describe the overall framework of the proposed approach, where we mainly introduce the Feature Embedding sub-Net (FEN) and the Feature Weighting sub-Net (FWN). Details about the training and test procedures of the proposed approach will also be presented.

3.1. Framework

Fig.2 shows the framework of our proposed deep ReID model. It can be seen that the global image and part images are simultaneously considered during each round of training. Given a training sample, we use an human pose estimation algorithm to acquire the locations of human pose joints. These pose joints are combined into different human body parts. The part regions are first transformed using our Feature Embedding sub-Net (FEN) and then are combined to form a new modified part image containing the normalized body parts. The global image and the new modified part image are then fed into our CNN together. The two images share the same weights for the first several layers, then have their own network weights in the subsequent layers. At last, we use Feature Weighting sub-Net (FWN) to learn the weights of part features before fusing them with the global features for final Softmax Loss computation.

Considering that pedestrian images form differen-

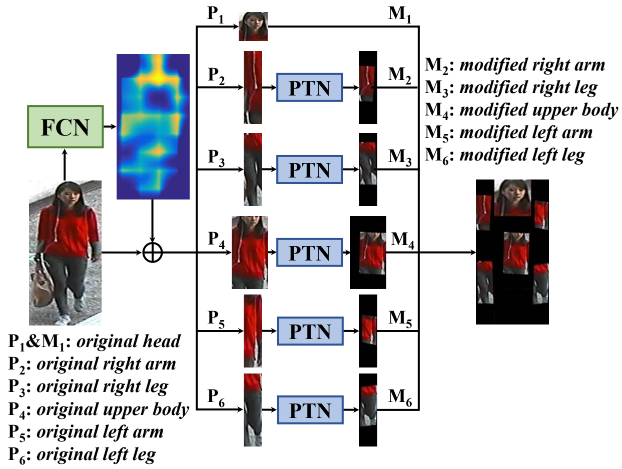

Figure 3. Illustration of Feature Embedding sub-Net (FEN). We divide the human image into 6 parts and apply an affine transformation on each part (except head part) by PTN, then we combine 6 transformed part regions together to form a new image.

t datasets have different sizes, it is not appropriate to directly use the CNN models pre-trained on the ImageNet dataset [7]. We thus modify and design a network based on the GoogLeNet [45], as shown in the Table 1. Layers from data to inception(4e) in Table 1 corresponds to the blue CNN block in Fig.2, CNNg and CNNp are inception(5a) and inception(5b), respectively. The green CONV matches the subsequent 1×1 convolution. The loss layers are not shown in Table 1. The Batch Normalization Layers [18] are inserted before every ReLU Layer to accelerate the convergence. We employ a Convolutional Layer and a Global Average Pooling Layer (GAP) at the end of network to let our network can fit different sizes of input images. In this work, we fix input image size as 512×256.

3.2. Feature Embedding sub-Net

The Feature Embedding sub-Net (FEN) is divided into four steps, including locating the joint, generating the original part images, PTN, and outputting the final modified part images.

With a given person image, FEN first locates the 14 joints of human body using human pose estimation algorithm [31]. Fig.1(c) shows an example of the 14 joints of human body. According to number, the 14 joints are $\{head,\ neck,\ rightshoulder,\ rightelbow,\ rightwrist,\ leftshoulder,\ leftelbow,\ leftwrist,\ lefthip,\ leftknee,\ leftankle,\ righthip,\ rightknee,\ rightankle\}$. Then we propose six rectangles to cover six different parts of human body, including the head region, the upper body, two arms and two legs.

For each human joint, we calculate a response feature map $V_i \in \mathbb{R}^{(X,Y)}$. The horizontal and vertical dimensions of the feature maps are denoted by X and Y, respectively. With the feature maps, the fourteen body joints $J_i = [X_i, Y_i], (i = 1, 2 \cdots 14)$, can be located by finding the center of mass with the feature values:

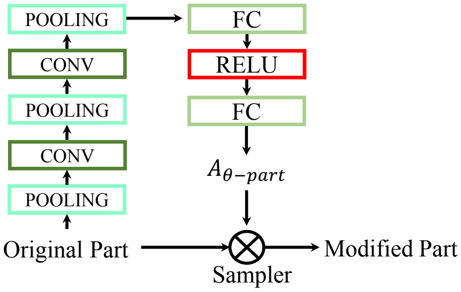

Figure 4. Detailed structure of the PTN subnet.

$$J_i = [X_i, Y_i] = [\frac{\sum V_i(x_j, y)x_j}{\sum V_i}, \frac{\sum V_i(x, y_j)y_j}{\sum V_i}], \quad (1)$$

where X_i, Y_i in Eq.1 are the coordinates of joints , and $V(x, y)$ is the value of pixels in response feature maps.

Different from [62, 16] , we do not use complex pose estimation networks as the pre-trained network. Instead, we use a standard FCN [31] trained on the LSP dataset [21] and MPII human pose dataset [2]. In the second step, the FEN uses the 14 human joints to further locate six sub-regions (head, upper body, left arm, right arm, left leg, and right leg) as human parts. These parts are normalized through cropping, rotating, and resizing to fixed size and orientation.

As shown in Fig.1(d), the 14 located body joints are assigned to six rectangles indicating six parts. The head part $P_1 = [1]$, the upper body part $P_2 = [2, 3, 6, 9, 12]$, the left arm part $P_3 = [6, 7, 8]$, the right arm part $P_4 = [3, 4, 5]$, the left leg part $P_5 = [9, 10, 11]$, and the right leg part $P_6 = [12, 13, 14]$, respectively.

For each body part set $P_i \in \{P_1, P_2, P_3, P_4, P_5, P_6\}$, The corresponding sub-region bounding box $H_i \in \{H_1, H_2, H_3, H_4, H_5, H_6\}$ can be obtained based on the location coordinates of all body joints in each part set:

$$H_i = \begin{cases} [x-30, x+30, y-30, y+30], & if \quad i = 1 \\ [x_{min}-10, x_{max}+10, y_{min}-10, y_{min}+10], & \\ & if \quad i = 2, 3, 4, 5, 6 \end{cases} \quad (2)$$

An example of the extracted six body sub-regions are visualized in Fig.1(d). As shown in Fig.1(e), these body sub-regions are normalized through cropping, rotating, and resizing to fixed sizes and orientations. All body parts are rotated to fixed vertical direction. Arms and legs are resized to 256×64, upper body is resized to 256×128 and head is resized to 128×128. Those resized and rotated parts are combined to form the body part image. Because 6 body parts have different sizes, black area is unavoidable in body part image.

Simply resizing and rotation can not overcome the complex pose variations, especially if the pose estimations are inaccurate. We thus design a PTN modified from Spatial Transformer Networks (STN) [19] to learn the angles required for rotating the five body parts.

STN is a spatial transformer module which can be inserted to a neural network to provide spatial transformation capabilities. It thus is potential to adjust the localizations and angles of parts. A STN is a small net which allows for end-to-end training with standard back-propagation, therefore, the introduction of STN doesn't substantially increase the complexity of training procedure. The STN consist of three components: localisation network, parameterised sampling grid, and differentiable image sampling. The localisation network takes the input feature map and outputs the parameters of the transformation. For our net, we choose affine transformation so our transformation parameter is 6-dimensional. The parameterized sampling grid computes each output pixel and the differentiable image sampling component produces the sampled output image. For more details about STN, please refer to [19].

As discussed above, we use a 6-dimensional parameter A_θ to complete affine transformation:

$$\begin{pmatrix} x^s \\ y^s \end{pmatrix} = A_\theta \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix}, \qquad (3)$$

where the $\theta_1, \theta_2, \theta_4, \theta_5$ are the scale and rotation parameters, while the θ_3, θ_6 are the translation parameters. The (x^t, y^t) in Eq.3 are the target coordinates of the output image and the (x^s, y^s) are the source coordinates of the input image.

Usually the STN computes one affine transform for the whole image, considering a pedestrian's different parts have various orientations and sizes from each other, STN is not applicable to a part image. Inspired by STN, we design a Pose Transformer Network (PTN) which computes the affine transformation for each part in part image individually and combines 6 transformed parts together. Similar to STN, our PTN is also a small net and doesn't substantially increase the complexity of our training procedure. As a consequence, PTN has potential to perform better than STN for person images. Fig.3 shows the detailed structure of PTN. Considering a pedestrian's head seldom has a large rotation angle, we don't insert a PTN net for the pedestrian's head part. Therefore, we totally have 5 independent PTN, namely $A_{\theta-larm}$, $A_{\theta-rarm}$, $A_{\theta-upperbody}$, $A_{\theta-lleg}$, $A_{\theta-rleg}$. Each PTN can generate a 6-dimensional transformation parameter $A_{\theta i}$ and use $A_{\theta i}$ to adjust pedestrian's part P_i, we can get modified body part M_i. By combining the five transformed parts and a head part together, we obtain the modified part image.

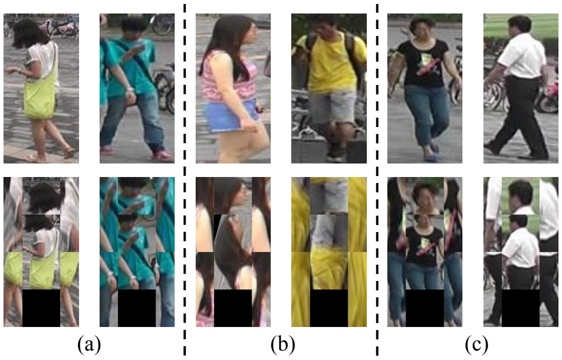

Figure 5. Illustration of some inaccurate part detection result. (a) Arms are obscured by upper bodies. (b) Upper bodies with large variation. (c) Miss detection on arms.

3.3. Feature Weighting sub-Net

The generated part features are combined with the global feature to generate a robust feature representation for precise person re-identification. As the poses generated by the pose detector might be affected by factors like occlusions, pose changes, etc. Then inaccurate part detection results could be obtained. Examples are shown in Fig.5. Therefore, the part features could be not reliable enough. This happens frequently in real applications with unconstrained video gathering environment. Simply fusing global feature and the part feature may introduces noises. This motivates us to introduce Feature Weighting sub-Net (FWN) to seek a more optimal feature fusion. FWN is consisted with a Weight Layer and a nonlinear transformation, which decides the importance of each dimension in the part feature vector. Considering that a single linear Weight Layer might cause excessive response on some specific dimensions of the part vector, we add a nonlinear function to equalize the response of part feature vector, and the fused feature representation is

$$F_{fusion} = [F_{global}, tanh(F_{part} \odot W + B)], \qquad (4)$$

where the F_{global} and the F_{part} are the global and part feature vectors. The W and B in Eq. 4 are the weight and bias vectors which have the same dimensions with F_{part}. The $\odot$ means the Hadamard product of two vectors, and the $[,]$ means concatenation of two vectors together. The $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ imposes the hyperbolic tangent nonlinearity. F_{fusion} is our final person feature generated by F_{global} and F_{part}.

To allow back-propagation of the loss through the FWN, we give the gradient formula:

$$\frac{\partial f_i}{\partial g_j} = \begin{cases} 1, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \neq j \end{cases} \qquad (5)$$

$$\frac{\partial f_i}{\partial p_k} = \begin{cases} w(1 - tanh^2(wp_j + b)), & \text{if} \quad i = k + m, \\ 0, & \text{if} \quad i \neq k + m. \end{cases} \qquad (6)$$

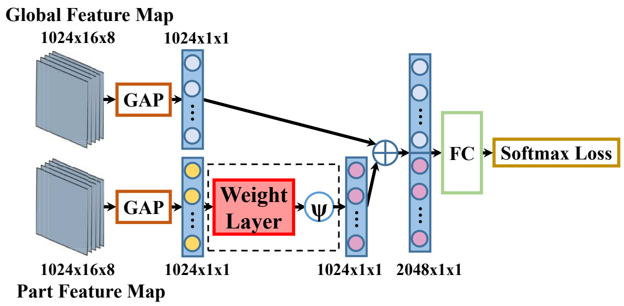

Figure 6. Illustration of the Feature Weighting sub-Net(FWN).

where $f_i \in F_{fusion}(i = 1,2\cdots m+n)$, $g_j \in F_{global}(j = 1,2\cdots m)$, $p_k \in F_{part}(k = 1,2\cdots n)$, $w_k \in W(k = 1,2\cdots n)$, $b \in B(k = 1,2\cdots n)$, m and n are the dimensions of F_{global} and F_{part}.

3.4. ReID Feature Extraction

The global feature and body-part features are learned by training the Pose-driven Deep Convolutional model. These two types of features are then fused under a unified framework for multi-class person identification. PDC extracts the global feature maps from the global body-based representation and learns a 1024-dimensional feature embedding. Similarly, a 1024-dimension feature is acquired from the modified part image after the FEN. The global body feature and the local body part features are compensated into a 2048-dimensional feature as the final representation. After being weighted by FWN, the final representation is used for Person ReID with Euclidean distance.

4. Experiment

4.1. Datasets

We select three widely used person ReID datasets as our evaluation protocols, including the *CUHK 03* [25], *Market 1501* [63], and *VIPeR* [15]. Note that, because the amount of images in *VIPeR* is not enough for training a deep model, we combine the training sets of *VIPeR*, *CUHK 03* and *Market 1501* together to train the model for *VIPeR*.

CUHK 03: This dataset is made up of 14,096 images of 1,467 different persons taken by six campus cameras. Each person only appears in two views. This dataset provides two types of annotations, including manually labelled pedestrian bounding boxes and bounding boxes automatically detected by the Deformable-Part-Model (DPM) [12] detector. We denote the two corresponding subsets as labeled dataset and detected dataset, respectively. The dataset also provides 20 test sets, each includes 100 identities. We select the first set and use 100 identities for testing and the rest 1,367 identities for training. We report the averaged performance after repeating the experiments for 20 times.

Market 1501: This dataset is made up of 32,368 pedestrian images taken by six manually configured cameras. It has 1,501 different persons in it. On average, there are 3.6 images for each person captured from each angle. The images can be classified into two types, *i.e.*, cropped images and images of pedestrians automatically detected by the DPM [12]. Because *Market 1501* has provided the training set and testing set, we use images in the training set for training our PDC network and follow the protocol [63] to report the ReID performance.

Table 2. The results on the *CUHK 03*, *Market 1501* and *VIPeR* datasets by five variants of our approach and the complete PDC.

dataset	CUHK03		Market1501		VIPeR
	labeled	detected			
method	rank1	rank1	mAP	rank1	rank1
Global Only	79.83	71.89	52.84	76.22	37.97
Part Only	53.73	47.29	31.74	55.67	22.78
Global+Part	85.07	76.33	62.20	81.74	48.42
Global+Part+FEN	87.15	77.57	62.58	83.05	50.32
Global+Part+FWN	86.41	77.62	62.58	82.69	50.00
PDC	**88.70**	**78.29**	**63.41**	**84.14**	**51.27**

VIPeR: This dataset is made up of 632 person images captured from two views. Each pair of images depicting a person are collected by different cameras with varying viewpoints and illumination conditions. Because the amount of images in *VIPeR* is not enough to train the deep model, we also perform data augmentation with similar methods in existing deep learning based person ReID works. For each training image, we generate 5 augmented images around the image center by performing random 2D transformations. Finally, we combine the augmented training images of *VIPeR*, training images of *CUHK 03* and *Market 1501* together, as the final training set.

4.2. Implementation Details

The pedestrian representations are learned through multi-class classification CNN. We use the full body and body parts to learn the representations with Softmax Loss, respectively. We report rank1, rank5, rank10 and rank20 accuracy of cumulative match curve (CMC) on the three datasets to evaluate the ReID performance.As for Market-1051, mean Average Precision (mAP) is also reported as an additional criterion to evaluate the performance.

Our model is trained and fine-tuned on Caffe [20]. Stochastic Gradient Descent (SGD) is used to optimize our model. Images for training are randomly divided into several batches, each of which includes 16 images. The initial learning rate is set as 0.01, and is gradually lowered after each 2×10^4 iterations. It should be noted that, the learning rate in part localization network is only 0.1% of that in feature learning network. For each dataset, we train a model on its corresponding training set as the pretrained body-based model. For the overall network training, the network is initialized using pretrained body-based model. Then, we adopt the same training strategy as described above. We implement our approach with GTX TITAN X GPU, Intel i7

Table 3. Comparisons on *CUHK 03* detected dataset.

Methods	rank1	rank5	rank10	rank20
MLAPG [28]	51.15	83.55	92.05	96.90
LOMO + XQDA [27]	46.25	78.90	88.55	94.25
BoW+HS [63]	24.30	-	-	-
LDNS [59]	54.70	84.75	94.80	95.20
GOG [35]	65.50	88.40	93.70	-
IDLA [1]	44.96	76.01	84.37	93.15
SI+CI [48]	52.17	84.30	92.30	95.00
LSTM S-CNN [47]	57.30	80.10	88.30	-
Gate S-CNN [46]	61.80	80.90	88.30	-
EDM [40]	52.09	82.87	91.78	97.17
PIE [62]	67.10	92.20	96.60	98.10
PDC	**78.29**	**94.83**	**97.15**	**98.43**

CPU, and 128GB memory.

All images are resized to 512 × 256. The mean value is subtracted from each channel (B, G, and R) for training the network. The images of each dataset are randomized in the process of training stage.

4.3. Evaluation of Individual Components

We evaluate five variants of our approach to verify the validity of individual components in our PDC, *e.g*., components like Feature Embedding sub-Net (FEN) and Feature Weighting sub-Net (FWN). Comparisons on three datasets are summarized in Table 2. In the table, "Global Only" means we train our deep model without using any part information. "Global+Part" denotes CNN trained through two streams without FEN and FWN. Based on "Global+Part", considering FEN is denoted as "Global+Part+FEN". Similarly, "Global+Part+FWN" means considering FWN. In addition, "Part Only" denotes only using part features. PDC considers all of these components.

From the experimental results, it can be observed that, fusing global features and part features achieves better performance than only using one of them. Compared with "Global Only", considering extra part cues, *i.e*., "Global+Part", largely improves the ReID performance and achieves the rank1 accuracy of 85.07% and 76.33% on *CUHK 03* labeled and detected datasets, respectively. Moreover, using FEN and FWN further boosts the rank1 identification rate. This shows that training our model using PTN and Weight Layer gets more competitive performance on three datasets.

The above experiments shows that each of the components in our method is helpful for improving the performance. By considering all of these components, PDC exhibits the best performance.

4.4. Comparison with Related Works

CUHK 03: For the *CUHK 03* dataset, we compare our PDC with some recent methods, including distance metric learning methods: MLAPG [28], LOMO + XQDA [27], BoW+HS [63], WARCA [22], LDNS [59], feature extraction method: GOG [35] and deep learning based methods: IDLA [1], PersonNet [52], DGDropout [53], SI+CI [48], Gate S-CNN [46], LSTM S-CNN [47], EDM [40], PIE [62] and Spindle [16]. We conduct experiments on both the detected dataset and the labeled dataset. Experimental results are presented in Table 3 and Table 4.

Table 4. Comparisons on *CUHK 03* labeled dataset.

Methods	rank1	rank5	rank10	rank20
MLAPG [28]	57.96	87.09	94.74	96.90
LOMO + XQDA [27]	52.20	82.23	94.14	96.25
WARCA [22]	78.40	94.60	-	-
LDNS [59]	62.55	90.05	94.80	98.10
GOG [35]	67.30	91.00	96.00	-
IDLA [1]	54.74	86.50	93.88	98.10
PersonNet [52]	64.80	89.40	94.90	98.20
DGDropout [53]	72.58	91.59	95.21	97.72
EDM [40]	61.32	88.90	96.44	99.94
Spindle [16]	88.50	97.80	98.60	99.20
PDC	**88.70**	**98.61**	**99.24**	**99.67**

Table 5. Comparison with state of the art on Market 1501.

Methods	mAP	rank1	rank5	rank10	rank20
LOMO + XQDA [27]	22.22	43.79	-	-	-
BoW+Kissme [63]	20.76	44.42	63.90	72.18	78.95
WARCA [22]	-	45.16	68.12	76.00	84.00
TMA [34]	22.31	47.92	-	-	-
LDNS [59]	29.87	55.43	-	-	-
HVIL [49]	-	78.00	-	-	-
PersonNet [52]	26.35	37.21	-	-	-
DGDropout [53]	31.94	59.53	-	-	-
Gate S-CNN [46]	39.55	65.88	-	-	-
LSTM S-CNN [47]	35.30	61.60	-	-	-
PIE [62]	55.95	79.33	90.76	94.41	96.65
Spindle [16]	-	76.90	91.50	94.60	96.70
PDC	**63.41**	**84.14**	**92.73**	**94.92**	**96.82**

Experimental results show that our approach outperforms all distance metric learning methods by a large margin. It can be seen that PIE [62], Spindle [16] and our PDC which all use the human pose cues achieve better performance than the other methods. This shows the advantages of considering extra pose cues in person ReID. It is also clear that, our PDC achieves the rank1 accuracy of 78.29% and 88.70% on detected and labeled datasets, respectively. This leads to 11.19% and 0.20% performance gains over the reported performance of PIE [62] and Spindle [16], respectively.

Market 1501: On *Market 1501*, the compared works that learn distance metrics for person ReID include LOMO + XQDA [27], BoW+Kissme [63], WARCA [22], LDNS [59], TMA [34] and HVIL [49]. Compared works based on deep learning are PersonNet [52], Gate S-CNN [46], LSTM S-CNN [47], PIE [62] and Spindle [16]. DGDropout [53] does not report performance on Market1501. So we implemented DGDroput and show experimental results in Table 5.

It is clear that our method outperforms these compared works by a large margin. Specifically, PDC achieves rank1 accuracy of 84.14%, and mAP of 63.41% using the single query mode. They are higher than the rank1 accuracy and

Table 6. Comparison with state of the art on VIPeR dataset.

Methods	rank1	rank5	rank10	rank20
MLAPG [28]	40.73	-	82.34	**92.37**
LOMO + XQDA [27]	40.00	67.40	80.51	91.08
BoW [63]	21.74	-	-	-
WARCA [22]	40.22	68.16	80.70	91.14
LDNS [59]	42.28	71.46	82.94	92.06
IDLA [1]	34.81	76.12	-	-
DGDropout [53]	38.6	-	-	-
SI+CI [48]	35.80	67.40	83.50	-
LSTM S-CNN [47]	42.40	68.70	79.40	-
Gate S-CNN [46]	37.80	66.90	77.40	-
MTL-LORAE [41]	42.30	72.20	81.60	89.60
Spindle [16]	**53.80**	**74.10**	83.20	92.10
PDC	51.27	74.05	**84.18**	91.46

Table 7. Performance of five variants of FWN on *CUHK 03*, *Market 1501* and *VIPeR*, respectively.

dataset	CUHK03		Market1501		VIPeR
	labeled	detected			
type	rank1	rank1	mAP	rank1	rank1
W_0	88.18	77.58	62.58	83.05	42.09
W_1	**88.70**	**78.29**	**63.41**	**84.14**	**43.04**
W_2	88.14	77.48	62.20	82.72	41.77
W_3	87.97	77.29	61.99	82.48	41.77
W_4	87.69	77.17	61.67	82.42	41.14

mAP of PIE [62], which performs best among the compared works. This is because our PDC not only learns pose invariant features with FEN but also learns better fusion strategy with FWN to emphasize the more discriminative features.

VIPeR: We also evaluate our method by comparing it with several existing methods on *VIPeR*. The compared methods include distance metric learning ones: MLAPG [28], LOMO + XQDA [27], BoW [63], WARCA [22] and LDNS [59], and deep learning based ones: IDLA [1], DGDropout [53], SI+CI [48], Gate S-CNN [46], LSTM S-CNN [47], MTL-LORAE [41] and Spindle [16].

From the results shown in Table 6, our PDC achieves the rank1 accuracy of 51.27%. This outperforms most of compared methods except Spindle [16] which also considers the human pose cues. We assume the reason might be because, Spindle [16] involves more training sets to learn the model for *VIPeR*. Therefore, the training set of Spindle [16] is larger than ours, *i.e.*, the combination of *Market 1501*, *CUHK03* and *VIPeR*. For the other two datasets, our PDC achieves better performance than Spindle [16].

4.5. Evaluation of Feature Weighting sub-Net

To test the effectiveness of Feature Weighting sub-Net (FWN), we verify the performance of five variants of FWN, which are denoted as W_k, $k = \{0,1,2,3,4\}$, where k is the number of Weight Layers in FWN with nonlinear transformation. For example, W_2 means we cascade two Weight Layers with nonlinear transformation, W_0 means we only have one Weight Layer without nonlinear transformation.

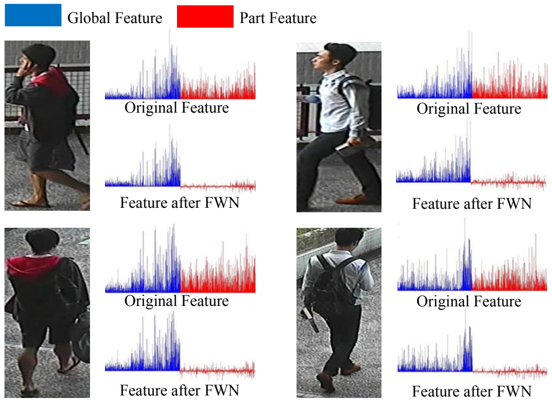

Figure 7. Examples of fused features before and after Feature Weighting sub-Net (FWN). The two images on the left side contains the same person. The other two images contains another person. FWN effectively keeps the discriminative feature and suppresses the noisy feature.

The experimental results are shown in Table 7. As we can see that one Weight Layer with nonlinear transformation gets the best performance on the three datasets. The ReID performance starts to drop as we increase of the number of Weight Layers, despite more computations are being brought in. It also can be observed that, using one layer with nonlinear transformation gets better performance than one layer without nonlinear transformation, *i.e.*, W_0. This means adding one nonlinear transformation after a Weight Layer learns more reliable weights for feature fusion and matching. Based on the above observations, we adopt W_1 as our final model in this paper. Examples of features before and after FWN are shown Fig. 7.

5. Conclusions

This paper presents a pose-driven deep convolutional model for the person ReID. The proposed deep architecture explicitly leverages the human part cues to learn effective feature representations and adaptive similarity measurements. For the feature representations, both global human body and local body parts are transformed to a normalized and homologous state for better feature embedding. For similarity measurements, weights of feature representations from human body and different body parts are learned to adaptively chase a more discriminative feature fusion. Experimental results on three benchmark datasets demonstrate the superiority of the proposed model over current state-of-the-art methods.

Acknowledgments This work is partly supported by National Science Foundation of China under Grant No. 61572050, 91538111, 61620106009, 61429201, 61672519, and the National 1000 Youth Talents Plan. Dr. Qi Tian is supported by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

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

[4] Y.-C. Chen, W.-S. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015.

[5] 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.

[6] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 2, page 6, 2011.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[8] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[9] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2011.

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

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

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

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

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.

[15] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007.

[16] Z. Haiyu, T. Maoqing, S. Jing, S. Shuyang, Y. Junjie, Y. Shuai, W. Xiaogang, and T. Xiaoou. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[17] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*. 2012.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[21] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMCV*, 2010.

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

[23] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *CVPR*, 2012.

[24] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 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] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[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] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012.

[30] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013.

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[32] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013.

[33] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

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

[35] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016.

[36] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

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

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

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

[40] 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.

[41] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

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

[43] C. Su, S. Zhang, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 2017.

[44] C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, and L. S. Davis. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition*, 66:4–15, 2017.

[45] 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.

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

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

[48] 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.

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

[50] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*. 2014.

[51] X. Wang and R. Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. 2014.

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

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

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

[55] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 29, pages 40–51, 2007.

[56] H. Yao, S. Zhang, D. Zhang, Y. Zhang, J. Li, Y. Wang, and Q. Tian. Large-scale person re-identification as retrieval. In *ICME*, 2017.

[57] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017.

[58] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. In *ICPR*, 2014.

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

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

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

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

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

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

[65] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. In *CVPR*, 2013.