IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Joint Learning of Body and Part Representation for Person Re-Identification

YUANYUAN WANG[1,3], ZHIJIAN WANG[1], WENJING JIA[2], XIANGJIAN HE[2], (SENIOR MEMBER, IEEE), AND MINGXIN JIANG[4]

[1]College of Computer and Information, Hohai University, Nanjing 211100, China
[2]Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia
[3]College of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China
[4]College of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian 223003, China

Corresponding author: Yuanyuan Wang (e-mail: zhfwyy@hyit.edu.cn).

This work was jointly supported by the Science and Technology Projects of Huaian (Grant HAG201602), the Major Program of Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (Grant 18KJA520002), the Six Talent Peaks Project in Jiangsu Province (Grant 2016XYDXXJS-012), the Natural Science Foundation of Jiangsu Province (Grant BK20171267), and the 533 Talents Engineering Project in Huaian (Grant HAA201738). It was mostly completed when Yuanyuan Wang was working at University of Technology Sydney as a Visiting Scholar.

ABSTRACT Person re-identification (ReID), aiming to identify people among multiple camera views, has attracted an increasing attention due to the potential of application in surveillance security. Large variations in subjects' postures, view angles and illuminating conditions as well as non-ideal human detection significantly increase the difficulty of person ReID. Learning a robust metric for measuring the similarity between different person images is another under-addressed problem. In this paper, following the recent success of part-based models, in order to generate a discriminative and robust feature representation, we first propose to learn global and weighted local body-part features from pedestrian images. Then, in the training phase, angular loss and part-level classification loss are employed jointly as a similarity measure to train the network, which significantly improves the robustness of the resultant network against feature variance. Experimental results on several benchmark datasets demonstrate that our method outperforms the state-of-the-art methods.

INDEX TERMS Person Re-identification, Metric Learning, Part-Based Model, Angular Loss

I. INTRODUCTION

PERSON re-identification (ReID) aims at matching person images obtained from non-overlapping camera views and finding the person-of-interest (query/probe) among a large gallery of pedestrian images [1]. In recent years, person ReID has attracted an increasing attention due to its wide applications in video surveillance, such as surveillance security and retrieval of suspects. Given a query image, person ReID spots his/her appearance at another time or from another camera view. Although the person ReID problem has been studied in recent years, applying person ReID in real world is still a challenging problem due to the variance of illumination, body pose, occlusion and resolution [2]. Moreover, problems such as misaligned detections and similar clothing among different people further increase its difficulty in real world applications. Recently, there has been a boom in the research community of interest with several deep learning based approaches [3]–[5] being reported and becoming the state of the arts with much better results than traditional approaches [6].

Existing solutions for person ReID mainly consist of two stages, i.e., 1) extracting features from input pedestrian images, and 2) computing the similarity among samples by comparing their features in order to find the matching ones. The first stage is mainly concerned with extracting discriminative features which can effectively describe persons under different camera views. The convolutional neural network (CNN) has now been widely used to automatically extract discriminative features from both the query and the gallery images [6], [7]. The second stage involves learning a robust distance metric, i.e., a mapping function which minimizes the same-person distance and maximizes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

FIGURE 1. Examples of misaligned detection in person ReID. (a) and (b) Excess background. (c) and (d) Missing body parts.

the different-person distance. Most of the existing person ReID methods focus on either robust feature extraction to describe the whole person images [1], [3], [8], [9] or distance similarity measurement for comparing features [10]–[12], or a combination of both [4], [13]. In this work, we propose innovative and effective solutions to address the above two aspects respectively, i.e., body-part based feature extraction and fusion, and angular-loss based deep metric learning.

A. JOINT LEARNING OF GLOBAL AND LOCAL BODY-PART FEATURES FROM ALIGNED PEDESTRIAN IMAGES

When it comes to extracting features, the deep learning based methods typically learn invariant global features without considering their spatial relationship and therefore do not distinguish different body parts. Recently, the effectiveness of adopting part-based CNN models for person retrieval has been verified in [2], [4], [14]–[17]. Some of these methods equally partition an input pedestrian image into several non-overlapping horizontal stripes and learn features from each part. These local features have been shown to be more discriminative than those global features extracted from the whole image. The premise of learning discriminative features using the part-based models is that the body parts are well aligned in different images. However, object misalignment problems such as excess background or missing parts are inevitable in machine-detected pedestrian images, as shown in Fig. 1. Such misalignment can result in mismatched body part regions across camera views and has become a critical issue for high-accuracy person ReID. Fig. 1(a) shows such an example, where in the first pair of images the upper body of the first image is aligned with the head of the second image. As its consequence, the feature descriptors extracted from these regions for different parts of a body cannot be compared directly using an effective similarity measurement. Therefore, equally partitioning horizontal stripes does not work well when severe misalignment occurs. Moreover, the noise caused by the excess background also significantly compromises the body feature learning and matching process. Both types of misalignment can result in identification failure and degrade the performance of person ReID.

The experiments reported in [13], [18], [19] have shown that the accuracy of person ReID with sample alignment and cropping preprocessing is usually higher than that without such preprocessing. In particular, for person ReID using the part-based methods, aligning pedestrian's body parts in the image is even more important. More effective methods need to be developed to address the deficiency of existing part-based methods in terms of sensitivity to partial occlusions and misalignment.

To resolve the above issues, following the success of the part-based models, we propose a solution to align body parts in detected pedestrian images and learn more discriminative local features and global features for person ReID. Our method first utilizes the OpenPose technique [20] to estimate human key points. Then, based on the locations of the estimated head and lower-body key points, we align the input images by removing excessive background or padding zeros to image borders. The aligned pedestrian images are then fed into ResNet [21] which learns feature descriptors. In this way, the features of different body regions can be better represented and aligned across images.

Moreover, since individual local and global feature learnings are suboptimal [13], [17], [18] but complementary with each other, we propose a two-branch framework to learn more discriminative global full-body features and local body-part features simultaneously.

When learning local body-part features, Wei *et al*. [2] proposed the framework of global-local-alignment descriptor (GLAD), which cropped body parts before extracting local features. However, in this way the local feature map of a cropped body part may become too small so that the discrimination of the extracted features becomes too weak, especially for the low-resolution head region. Therefore, different from the GLAD approach [2], we conduct partition on the convolutional layer (ResNet-50) for learning part-level features. To that end, we formulate the local and global branches on a shared convolution network to extract the feature map so that the model parameter size and overfitting risks are reduced [13].

Moreover, feature fusion, which has been demonstrated to be effective in image search [14], [22], [23], is able to capture complementary information. Existing part-based methods treat each body part equally. However, our experimental results have shown that different body parts contribute differently to the pedestrian ReID. Therefore, in our work, the body-part features extracted from each of the body part regions are concatenated with different weights to form the final feature representation. The effectiveness of this strategy will be shown in our experiments in Section IV.

B. MEASURING SIMILARITY JOINTLY USING THE OVERALL ANGULAR LOSS AND PART-LEVEL CLASSIFICATION LOSS

In terms of measuring similarity, representation learning [3] and deep metric learning [10] are two types of methods commonly used for measuring the target loss function. Recently, different loss functions, such as triplet loss [24], quadruplet loss [10] and n-pair loss [25], have been proposed to measure the feature similarity among different samples. These methods mainly use distance-based metrics to measure the similarity, which are sensitive to scale changes. Moreover,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

setting a fixed margin threshold does not suit for samples with significantly different intra-class distributions [4].

Person ReID is mostly considered as a special task of image retrieval problem [23], both of which look for images from gallery containing the same pedestrians or objects according the probe image. Very recently, an angular loss [26] is proposed to solve the image searching task. Inspired by this work, we propose to use angular relationships to learn a more robust similarity metric for the person ReID task.

The main contributions of this work are three folded.

1) Firstly, we propose a two-module framework to jointly learn global and local body-part features from pedestrian images. In order to learn a more discriminative body-part features, input pedestrian images are aligned based on estimated body key points. Learned body-part features are then fused by taking into consideration their contributions to identification.
2) Secondly, for a more robust similarity measurement, we propose to use an overall angular loss and part-level classification loss jointly in metric learning, resulting in better convergence and performance. For this purpose, we encode the relation in terms of the angle inside triplet at the negative point.
3) Last but not the least, to demonstrate the superior performance over the state of the arts, our proposed method is compared with the state-of-the-art methods on three person ReID datasets, i.e., Market1501 [23], DukeMTMC-reID [27], [28] and CUHK03 [1].

The rest of the paper is organized as follows. In Section II, a more focused literature review on the recent, highly relevant works is first presented. Then, Section III details our proposed method. Comparative experiments and performance analysis are presented in Section IV. Finally, the paper is concluded in Section V.

II. RELATED WORKS

In this section, we review the recently published approaches on part-based deep learning methods for the person ReID and metric learning problem.

A. PART-BASED DETECTION FOR PERSON REID

Recently, deep learning has become the main method for person ReID showing remarkable performance [1], [6], [9]. The above-mentioned methods learn full-body features while ignoring the spatial local information of different body parts, so it is difficult for them to capture detailed information in pedestrian images. When the detection outcome is not well aligned or there are partial occlusions, these features' discriminability is compromised.

In most recent works, part-based methods have achieved much better performance for person ReID than just learning full-body discriminative features [2], [4], [15], [17], [18], [29]. Cheng *et al*. [4] proposed a multi-channel part-based network to independently learn the feature maps of four body parts. Similar to the approach in [4], there are several methods that equally partition the input pedestrian images into several stripes and learn part features independently [2], [17]. Li *et al*. [17] captured the local context knowledge by stacking multi-scale convolution layers on each layer. Lin *et al*. [30] introduced a multi-structure strategy to handle spatial misalignments. Their strategy of tackling the misalignment problem works well on relatively small datasets. However, in large-scale personal ReID datasets, their approach appears to be computationally expensive and time consuming. Wei *et al*. [2] coarsely divided a pedestrian image into three parts according to human body structure. However, in these works, image partitioning is carried out at the bottom of the network, which results in the areas being very small for a CNN and therefore the features generated are too weak. Moreover, when utilizing local information, these methods simply divide each image into several horizontal parts without considering alignment.

In addition, human pose estimation methods have been reported with encouraging improvement for pedestrian partition [14], [16], [31]. Besides, Zhao *et al*. [29], [32] embedded the attention mechanism in the network. Similarly, Bai *et al*. [33], [34] focused on body parts and combined CNN with long-short-term memory (LSTM) components. Qian *et al*. [35] used generative adversarial network (GAN) to generate images of different human body poses to solve the problem of large pose variations and insufficient cross-view pairing pictures of the target pose. They also proposed "pose normalization", which converts each input human image correspondingly to one of the eight pre-set standard poses. However, this work, same as most other existing works, has assumed equal contributions from different body parts to the ReID results. They did not make any effort to handle the detection errors. Instead, we use OpenPose to extract pedestrian's key points in order to address the issue with misaligned detection, which is one of our major contributions.

Feature fusion has been proven effective in person ReID [22], [36]. Generally, there are mainly two kinds of feature fusion, i.e., early fusion and late fusion. In early fusion, descriptors were combined at feature level [14], where the pretrained human pose was used to estimate person joint points and then features were extracted. The main contributors to the accuracy results in [14] come from the whole image, and the regional features of the body enhance the overall accuracy very little. This method tends to produce a large number of parameters at the fully connected layer. Furthermore, the fine-grained part extraction mentioned in the paper is computationally expensive [2]. On the other hand, late fusion refers to the fusion at the score or decision level [22] leading to a trade-off between the information contents. We adopt the adaptive late fusion method as proposed in [22] to calculate the weight of each body-part.

B. METRIC LEARNING APPROACHES

In terms of the loss function, traditional metric learning methods for person ReID learn a Mahalanobis distance metric in a Euclidean space to calculate the similarity of two pedestrian

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

images [6]. Recently, several deep metric learning methods have been reported, and they usually extract features from pedestrian images using CNNs or other deep models, and then compute a feature distance as the similarity measure of pedestrian images. The Triplet loss function [4], [12] has been used to investigate the relative similarity of people appearing in pairs of images and has been widely used in person ReID retrieval [3] and face recognition [24]. Cheng *et al*. [4] used an improved triplet loss function to train a CNN, jointly learning body and body-part features of people in images. However, the triplet loss needs mining hard samples for efficient mining of similar features; otherwise the training process will stagnate, take long time or unable to converge if the samples are too complicated [10].

To address the above problems, some variants of the triplet loss and hard negative/positive mining methods have been proposed [10]–[12], [26]. The performance of the quadruplet loss [10] on the testing set can be improved by further reducing the intra-class variations and enlarging the inter-class variations. Hermans *et al*. [11] proposed a generalization of the lifted embedding loss taking into account all anchor-positive pairs. The margin sample mining loss (MSML) [12] introduced the idea of hard sample mining, which only picks out the hardest positive sample pair and the hardest negative sample pair to calculate the loss.

Recently, Wang *et al*. [26] used a triangular geometry instead of a distance to capture the local structure of triplet loss. Angular in a triangle has rotation and scale invariance, and meanwhile can capture additional local structure of triplet triangles by imposing a geometric constraint.

In this work, we use the angular loss function to define the core component of a metric learning loss. We do not need to select a margin threshold, which has a large range of values like a triplet loss. Instead, we only need to determine the angular relationship between the samples in a triangular structure. Our experimental results demonstrate the benefits of this approach.

III. METHODOLOGY

A. OVERVIEW OF THE PROPOSED FRAMEWORK

Fig. 2 illustrates the framework of our proposed network. As shown in this figure, our proposed model addresses both discriminative feature generation and robust image similarity computation and ranking. During the training stage, the proposed network relates to two parallel and complementary branches, i.e., 1) Global full-body feature representation and local body-part feature representation, and 2) Softmax classification and ranking with the angular loss to predict the identities of training images. Weighted features of multiple body parts are concatenated to form the feature descriptor of an input image.

In our proposed framework, these two branches share a convolutional network to extract a feature map. The features extracted from the bottom layer have some commonalities, which can greatly reduce the number of parameters. As pointed out in [13], [33], the global representation of full-body (i.e., pose, shape, background) and part-based local representation (i.e., head, upper body or lower body) are complementary to each other. Therefore, combining the full-body representation with the part-based representation gains the learned features more discriminative power and maximizes the same person identity matching. The global feature is extracted by applying global pooling directly on the feature map. For the local feature representation, one 1×1 convolutional layer is applied after an unequal horizontal stripe pooling. For model training, we utilize the cross-entropy classification loss function at the local layer and angular loss function at the global layer to optimize person identification and ranking task.

B. PEDESTRIAN ALIGNMENT

The outcomes of pedestrian detection may contain excess background due to non-ideal detection results. When using part-based models, the misalignment and occlusion of human body parts between pedestrian images become a critical factor affecting the person ReID accuracy. In this work, given originally detected pedestrian images, we first use the OpenPose toolkit [20] to crop images when excessive background exists or pad zeros to the corresponding image borders when there is part missing. The OpenPose technique is a real-time multi-person system in order to jointly detect human body. As illustrated in Fig. 3, it produces 18 key points for front-body estimation (the top-left figure) or 15 key points for back-body estimation (the bottom-left figure), and their connections for each input person image. The values of the positions where no key points are detected are set to zero. In the case of a pedestrian's backside image, the values of the 0-th, 14-th and 15-th key points are zero. Even in the case of partial occlusions, the key points of the pedestrian's main part can be detected at least, e.g., the 1-th, 8-th and 11-th key points.

Based on the detected key points values of the head and lower body, we first align the pedestrian images by removing excess background. An example is shown in the second column of Fig. 3, where the excess background above the person's head is removed. We then design a procedure to deal with image misalignment, as illustrated in Alg. 1, where α is a hyperparameter used to determine whether there is an excessive background or need padding border in the image. In our experiments, we empirically set the values of α to 0.1 according to the statistics of the detected key points.

As shown in the second row of Fig. 3, body parts are roughly aligned to their corresponding positions in the images. Thus, preprocessing the misaligned images reduces the scale and position variance and roughly aligns the parts of the human body, so it benefits the subsequent matching steps.

C. LOCAL BODY-PART FEATURE EXTRACTION

The effectiveness of using part-based features for person ReID has been reported in several latest works [2], [4], [15], [18], [33]. The existing part-based methods evenly divide an input pedestrian image into predefined parts without considering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

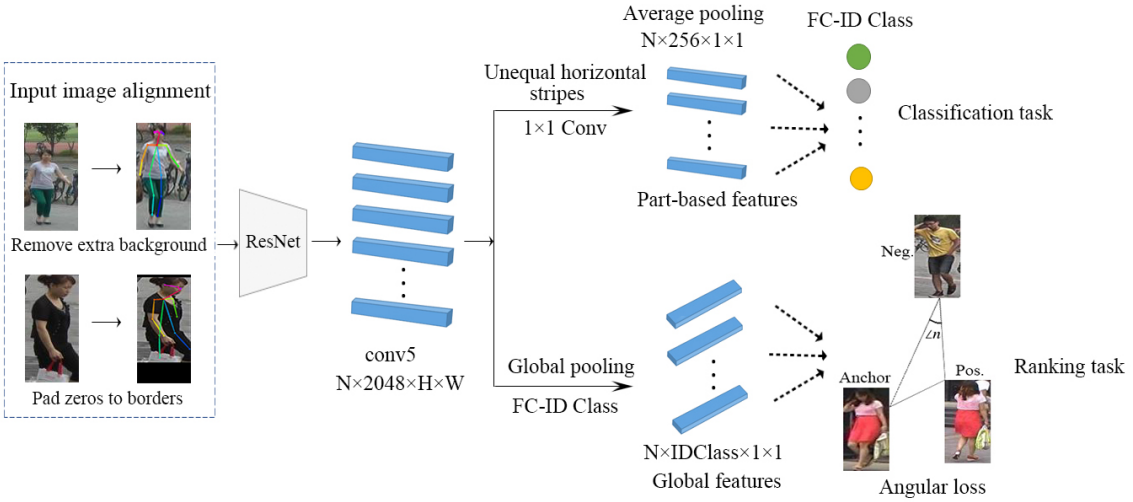

FIGURE 2. The framework of our proposed network. The OpenPose toolkit [20] is first used to estimate human key points and process input images. Excess background is removed and image borders are padded with zero based on the information of the key points. Then, the aligned images are fed into ResNet-50 to extract feature maps. In the body-part branch, the processed image is unequally partitioned into several horizontal stripes. After a 1 × 1 kernel-sized convolutional layer and a fully connected (FC) layer, a classifier is trained with cross-entropy loss. In another branch, ranking is carried out using the angular loss function.

Algorithm 1 Image misalignment preprocessing

Step 1. Use the OpenPose to produce 15 or 18 key points of the input pedestrian image.

Step 2. Determine whether the pedestrian image is with excess background or missing part based on the ordinates of the 16th, 17th, 9th, 12th, 10th and 13th key points:

if $y_{16} > \alpha \cdot H_{image}$ or $y_{17} > \alpha \cdot H_{image}$ **then**
The ordinate range of the cropped image is $y_{crop} \in [0, y_{16} - \alpha \cdot H_{image}]$.
else if $y_{16} = 0$ and $y_{17} = 0$ and $y_1 > 0$ **then**
Padding zeros to the image's upper border.
The height of padding is $2 \cdot \alpha \cdot H_{image}$.
end if
if $y_{10} < (1-\alpha) \cdot H_{image}$ and $y_{13} < (1-\alpha) \cdot H_{image}$ **then**
The ordinate range of the cropped image is $y_{crop} \in [y_{10} + \alpha \cdot H_{image}, H_{image}]$.
else if $y_{10} > (1-\alpha) \cdot H_{image}$ or $y_{13} > (1-\alpha) \cdot H_{image}$ **then**
Padding zeros to the image's lower border.
The height of padding is $\alpha \cdot H_{image}$.
else if $y_9 > (1-\alpha) \cdot H_{image}$ or $y_{12} > (1-\alpha) \cdot H_{image}$ **then**
Padding zeros to the image's lower border.
The height of padding is $3 \cdot \alpha \cdot H_{image}$.
else if $y_8 > (1-\alpha) \cdot H_{image}$ or $y_{11} > (1-\alpha) \cdot H_{image}$ **then**
Padding zeros to the image's lower border.
The height of padding is $5 \cdot \alpha \cdot H_{image}$.
end if

Step 3. Stretch the image keeping its height unchanged.

FIGURE 3. Examples of the estimated key body points. The first column shows the estimated key points of front body and back body respectively. The images on the right show the person samples with inaccurate detection (excess background and missing part in the first row) and aligned images according to key points information (in the second row on the right).

the alignment of different body parts. Moreover, the partition of images occurs at the bottom layer of a CNN, resulting in that each partition becomes too small to contain sufficient context information. This significantly reduces the discriminability of the produced features [2]. Besides, none of the existing works has considered the unequal contributions of different parts of a body for feature representation.

In our work, in order to retain the spatial contextual information, aligned person images are fed into ResNet-50 to extract features without segmentation. Moreover, the lower convolution layer intends to capture low-level features, such as object-oriented edges or corners, which are similar for the images containing the same person. Therefore, we construct a two-branch framework where the lower convolution layer is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

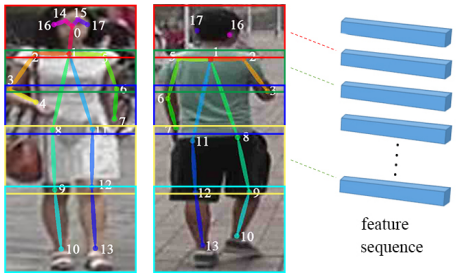

FIGURE 4. Each vector in the feature sequence describes the corresponding region in the original image, which is unequally partitioned into five horizontal stripes according to the statistics of human key points.

shared by the two branches, as shown in Fig. 2. This reduces the size of the model's parameters and the risk of over-fitting [13], [15]. Furthermore, we remove the last spatial down-sampling operation and the fully connected layer in the ResNet-50 to enrich the granularity of features, as suggested in [15], [37].

As depicted in Fig. 2, we use ResNet-50 for learning the shared low-level features, which have the same size of $N \times 2048 \times H \times W$, where N is the batch size, 2048 is the channel number and $H \times W$ is the image size. In the local representation layer, according to our observation, different body parts show different levels of discriminativity. Unequally partitioning an image according to the natural structure of the human body in the image can therefore produce better results than equally partitioning.

According to the statistics of estimated key points, we divide an input pedestrian image into five horizontal, overlapping parts. We denote the k-th key point of a body joint location by p_k for $k = 1, 2, \ldots, 17$. The 18 located body joints are assigned to five sets, i.e., head part $B_1 = \{p_0, p_1, p_{14}, p_{15}, p_{16}, p_{17}\}$, shoulder part $B_2 = \{p_1, p_2, p_3, p_5, p_6\}$, abdomen part $B_3 = \{p_3, p_4, p_6, p_7\}$, leg part $B_4 = \{p_8, p_9, p_{11}, p_{12}\}$ and foot part $B_5 = \{p_9, p_{10}, p_{12}, p_{13}\}$, as illustrated in the five overlapped, color coded boxes in Fig. 4. The corresponding sub-region $B_i \in \{B_1, B_2, B_3, B_4, B_5\}$, $(i = 1, 2, \ldots, 5)$ can be obtained based on the vertical coordinates of the key points in each part set.

Denote the vertical coordinate of the sub-region B_i by y_{B_i}, and the vertical coordinate of the key point p_k by y_{p_k},

$$y_{B_i} = \begin{cases} [0, (\mu_{y_{B_i}} + \xi)], & i = 1 \\ [(\min(\mu_{y_{B_i}}) - \xi), (\max(\mu_{y_{B_i}}) + \xi)], & i = 2, 3, 4, 5 \end{cases}, \tag{1}$$

where $\mu_{y_{B_i}} = \frac{1}{Num_{B_i}} \sum_i^{Num_{B_i}} y_{p_{k,i}}$ for $i = 1, 2, \ldots, 5$ and $k = 0, 1, \ldots, 17$, $\mu_{y_{B_i}}$ is the mean value of the key points $y_{p_{k,i}}$ and it reduces the inaccurate part detection results due to occlusion, and ξ is a parameter controlling the overlapping between neighboring part regions. ξ is set to 5 for the 128×64 sized pedestrian images in our experiments.

The learned features of each stripe then undergo an average pooling operation with a kernel size of $Stripe_{height}[i] \times W$, which generates the corresponding feature vector of a size of $[N, 2048, 1, 1]$. Then, 1×1 convolution is applied to decrease the channel number from 2048 to 256 (size of $[N, 256, 1, 1]$). As shown in Fig. 4, each local feature vector describes a horizontal region part of the original image. Different from the method in [2], our method divides each image at a higher level, instead of directly cropping the image.

Then, we use the late feature fusion method described in [22] to calculate the feature importance of each body-part. In this way, all the resultant feature vectors are fused together as a person's feature representation.

During training, the body-part classification sub-branch learns a softmax classifier with a cross-entropy loss for identity prediction. After the feature fusion, a softmax layer with T nodes are then connected, where T is the unique person number in the training set. We use the cross-entropy loss [15] (the "CrossEntropyLoss" function available in the PyTorch library) as a loss function L_{Cross}, which is defined as:

$$L_{Cross}(p, q) = -\sum_{t=1}^{T} p(t) \times \log(q(t)), \tag{2}$$

where T is the number of classes, and $q(t)$ is the predicted probability of the input belonging to label t, $q(t)$ is normalized by the softmax layer, $\sum_{t=1}^{T} q(t) = 1$, and $p(t)$ is the ground-truth distribution.

D. GLOBAL REPRESENTATION AND METRIC LEARNING

Part-based representation focuses on the discriminative pedestrian details. Global features, containing more high-level semantic information, can complement to part-based local features. Therefore, we extract a global representation directly by inserting a global average pooling and a fully connected layer after the shared low-level feature extraction. This feature is adopted as a global descriptor. More specifically, we adopt angular loss [26] for metric learning to further improve the performance.

The angular loss [26] used the relationship of angles rather than distance as a measure of similarity. Traditional triplet loss and its various variants are based on a distance measurement, which cannot address the problem of scale change. It is difficult to select an appropriate global distance margin γ in Eq. 3 due to that the intra-class distance may vary significantly. The angular loss constrains the angle n of the negative sample points as shown in Fig. 2 (the bottom-right part of the "Angular loss"). The main idea of angular loss is to encode the relationship in terms of the angle inside triplet at the negative point. By setting a upper bound for the angle, the method pushes the negative point away from the center of positive cluster and drags the positive points closer to each other. Angle is a similarity-transform-invariant metric, proportional to the relative comparison of triangle edges. The traditional triplet only takes two edges into account. The angular loss is scale-invariant and can improve the robustness of the objective function to counter the feature variations due to distance. The additional, third constraint improves the robustness and effectiveness of the

optimization. It essentially adds geometric constraints, which can capture additional local structures in comparison with the triplet loss or verification loss.

The goal of metric embedding learning is to learn a function $f(x) = \mathbb{R}^F \rightarrow \mathbb{R}^D$ that maps semantically similar instances from the data manifold in $\mathbb{R}^F$ onto metrically close points in $\mathbb{R}^D$ [4], [11]. Recently, the triplet loss has been proved to be effective in learning discriminative image features and has been widely used in person ReID [10], [12]. There are many methods proposed in the literature to improve the triplet loss in order to achieve better performance on a testing set.

Every triplet loss $\{x_a, x_p, x_n\}$ contains an anchor x_a, a positive x_p and a negative x_n in the iteration of a batch, where x_a and x_p are images from the same person, and x_n is from a different person. The philosophy of the triplet loss function is to try to minimize the distance between an anchor and a positive person sample meanwhile maximizing the distance between the anchor and a negative sample. The triplet of ℓ_2-normalized features $\{f_a, f_p, f_n\}$ is used to calculate the distances. Thus, the commonly used triplet loss [10] can be formulated to:

$$L_{triplet} = \sum_{a,p,n}^{N} \left[\underbrace{\|f(x_a) - f(x_p)\|_2^2}_{minimize} - \underbrace{\|f(x_a) - f(x_n)\|_2^2}_{maxmize} + \gamma_{triplet} \right]_+, \tag{3}$$

where the threshold $\gamma_{triplet}$ is a distance margin distinguishing the positive pairs from the negative. $f(x_a)$, $f(x_p)$, $f(x_n)$ represent the normalized highly-embedded features and $[\cdot]_+ = max(\cdot, 0)$.

In the original triplet constraint (Eq. 3), it is difficult to choose a proper distance margin without meaningful reference. By comparison, setting an angle in the angular constraint is an easier task because it has concrete and interpretable meaning in geometry. According to triangle theory, the sum of the triangle's inner angles is 180°. In order to keep negative samples away from both the anchor and positive samples, the angle of n needs to be smaller than 60°, as shown in Fig. 2. We aim to make the angle n smaller. Using the tangent theorem and the definition of hyperparameter θ, the angular loss implies minimizing the following hinge loss $L_{Angular}$ [26] as shown in Eq. 4. The full derivation and proof can be found in [26]. It constrains the angle n to be less than a predefined upper bound θ. We determine the range of hyperparameter values through statistical data, and finally optimize the hyperparameter by a hyperparameter optimizer. The hyperparameter θ is set to be between 30° and 50°, which works well in our experiments.

$$L_{Angular} = \left[\|x_a - x_p\|_2^2 - 4\tan^2\theta \|x_n - x_c\|_2^2 \right]_+. \tag{4}$$

Different branches have complementary advantages for learning discriminative features. We jointly train the entire network to predict the identity of each input image for both part-based and global feature learning. Then, the final loss of the training network is given in Eq. 5.

$$L = L_{Angular} + \lambda L_{Cross}, \tag{5}$$

where λ is the hyperparameter, which is set to 2.0 in our experiments.

IV. EXPERIMENTS

We implement our proposed algorithm with the ResNet-50 architecture on PyTorch and evaluate its performance on three large-scale benchmark datasets for person ReID, i.e., Market1501 [23], DukeMTMC-reID [27], [28] and CUHK03 [1]. In this section, we report the results and compare our proposed method with the state-of-the-art methods.

A. DATASETS

The Market1501 dataset [23] is one of the most widely used datasets for person ReID. It contains in total 32,668 annotated bounding boxes of 1,501 identities collected from six cameras (five high-resolution cameras and one low-resolution camera), among which 19,732 images of 750 identities are for testing and 12,936 images of 751 identities are for training. On average, there are 17.2 images per identity in the training set. All images are automatically detected using the deformable part model (DPM) approach instead of using hand-drawn boxes, for a more realistic setting. The misalignment problem of body region across images is common in the person ReID dataset [8], [14]. There are two kinds of evaluation settings [23], i.e., single query and multiple query, using one or several images of one person under one camera for a query. In this work, both single and multiple query modes are used for the Market-1501 dataset.

The DukeMTMC-reID dataset is a subset of the newly-released DukeMTMC dataset [28] used for cross-camera tracking, which manifests itself as one of the largest and challenging pedestrian image datasets. We adopt its ReID version provided in [27]. Similar to the format of the Market-1501 dataset, the DukeMTMC-reID dataset contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images captured by eight high-resolution cameras. The pedestrian images are cropped using hand-drawn bounding boxes. Pedestrians are similarly dressed, so the DukeMTMC-reID dataset is more challenging.

The CUHK03 dataset [1] contains in total 13,164 images of 1,467 identities collected on the CUHK campus. Each identity is captured by two disjoint cameras and has 4.8 images on average for each view. The CUHK03 dataset contains two kinds of bounding boxes, i.e., the "detected" set produced by the DPM algorithm and the hand-drawn "labeled" set. In our work, we evaluate our model on the bounding boxes detected by the DPM algorithm, which is closer to the realistic setting but more challenging. Unless otherwise specified, CUHK03 indicates the detected set in this paper. In the training set, there are on average 9.6 images per identity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

Experiments on the CHUK03 and DukeMTMC-reID datasets are performed in single query mode.

B. IMPLEMENTATION DETAILS

The implementation details of the proposed method are described as follows.

We use ResNet-50 pre-trained on ImageNet as the base model where the average pooling and fully connected layer are discarded. All training and testing images are normalized to 128×64 pixels. We follow the common data augmentation strategies such as random horizontal flipping and cropping to alleviate the over-fitting problem. We use the L2 normalized person representation and Euclidean metric to measure the distance between two pedestrian images. A fine-tuning strategy is applied on the training set to avoid overfitting [38], [39]. We empirically set the dropout probability to 0.5 [8], [27], and use the SGD solver to train our model. All experiments are conducted on a server of Intel i7 CPU and equipped with two GeForce GTX 1080 GPU cards. We set batch size to 128. The increased training time of the model is mainly caused by the metric learning of angular loss. Adam optimizer [40] is used and the initial learning rate is set to 0.0001. Then, we reduce the learning rate to 0.00001 at 50 epochs until convergence is achieved. The hyperparameter optimizer sklearn is used to search on a hyperparameter space to find the most reasonable hyperparameter θ for the metric learning model. The initial range of θ in Eq. 4 is set to $25^\circ \sim 60^\circ$ in our experiments.

In terms of evaluation, we adopt the commonly used evaluation protocol [1], [29] for fair comparison with existing methods. Concretely, we evaluate our method with the cumulative matching characteristics (CMC) at rank-1, rank-5, rank-10 accuracies and mean average precision (mAP), which reflects the precision and recall rates of the retrieval process. The CMC shows the probability that a query identity appears in the ranking list. The rank-i accuracy indicates the probability that a query image is found within the top i ranks in the ranking list for $i = 1, 5$ and 10.

In addition, similar to other methods described in [18], we adopt the re-ranking method proposed in [41] and perform re-ranking to further improve the retrieval performance of the initial results. Zhong *et al.* [41] proposed a k-reciprocal encoding method to re-rank the results of person ReID. After obtaining the initial top-k using the normal person ReID method, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector. The re-ranking method with k-reciprocal encoding combines the original distance and Jaccard distance. The advantage of re-ranking method is that no labeled data is required and no human interaction. The re-ranking method effectively improves the person ReID performance on several large-scale benchmark datasets for person ReID.

C. COMPARISON WITH THE STATE OF THE ARTS

We compare the results of our method with those of the representative methods tested on several benchmark datasets.

Table 1. Comparison of results of single and multiple queries on Market1501 dataset. * denotes the use of deep learning methods for body-part features.

Methods	Single Query		Multiple Query	
	mAP	rank-1	mAP	rank-1
BoW + KISSME [23]	20.76	44.42	19.42	44.36
Gated S-CNN [43]	39.55	65.88	48.45	76.04
P2S (ResNet-50) [44]	44.27	70.72	55.73	85.78
CADL (CaffeNet) [45]	47.11	73.84	55.58	80.85
Spindle Net* [14]	-	76.90	-	-
GAN (ResNet-50) [27]	56.23	78.06	68.52	85.12
TOMM (ResNet-50) [3]	59.87	79.51	70.33	85.84
Quad (ResNet-50) [10]	61.10	80.00	-	-
MSCAN* (CaffeNet) [17]	57.53	80.31	66.7	86.79
PAR* (ResNet-50) [29]	63.40	81.00	-	-
SSM (ResNet-50) [46]	68.80	82.21	76.18	88.18
SVDNet (ResNet-50) [47]	62.10	82.30	-	-
PAN (ResNet-50) [8]	63.35	82.81	71.72	88.18
PDC* (CaffeNet) [16]	63.40	84.40	-	-
TriNet (ResNet-50) [11]	69.14	84.92	76.42	90.53
JLML (ResNet-39) [13]	65.50	85.10	74.50	89.70
Angular (GoogLeNet) [26]	69.73	85.53	77.02	91.29
MultiScale* (ResNet-50) [42]	73.10	88.90	80.70	92.30
GLAD* (GoogLeNet) [2]	73.90	89.90	-	-
Ours (ResNet-50)	**74.57**	**91.51**	**81.78**	**94.71**
Ours + re-rank (ResNet-50)	**88.50**	**92.96**	**91.75**	**95.31**

1) Evaluation on Market-1501

We evaluate the proposed method against the recently published works on the Market-1501 dataset. As shown in Table 1, our method achieves competitive results on the Market-1501 dataset. Specifically, our method attains 74.57% in terms of mAP and 91.51% matching rate at rank-1 under the single query setting, and outperforms all other person ReID methods, including hand-crafted methods [23], deep learning based methods and deep learning methods with part features [2], [14], [16], [17], [29], [42]. Combined with the re-ranking approach [41], our performance is further improved, reaching 92.96% matching rate at rank-1 with single query mode. Note that, in Table 1 other methods did not use the re-ranking method, which was published only recently in 2017 and therefore has only got widely used as a default technique in the last couple of years. Similar improvements are obtained using multiple query settings on the Market1501 dataset, gaining 81.78% mAP and 94.71% rank-1 matching rate without re-ranking, and 91.75% mAP and 95.31% rank-1 matching rate with re-ranking, respectively.

2) Evaluation on DukeMTMC-reID and CUHK03

The comparison with several existing models on the DukeMTMC-reID dataset and CUHK03 dataset are presented in Table 2. As shown in this table, our method outperforms all other methods with a large margin, achieving 81.14% and 60.84% in mAP after using re-ranking method on the DukeMTMC-reID dataset and CUHK03 dataset, respectively. Using the pedestrian bounding boxes detected by the DPM algorithm in the CUHK03 dataset, our method has achieved 61.57% rank-1 accuracy. These results demonstrate that our model has consistent superiority and robustness over existing methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

Table 2. Comparison of results of single query on DukeMTMC-reID and CUHK03 ("detected" set). * denotes the use of deep learning methods for body-part features. # denotes unpublished papers.

Methods	DukeMTMC-reID		CUHK03	
	mAP	rank-1	mAP	rank-1
BoW + KISSME [23]	12.17	25.13	-	24.30
GAN (ResNet-50) [27]	47.13	67.68	-	-
PAN# (ResNet-50) [8]	51.51	71.59	34.00	36.30
SVDNet (ResNet-50) [47]	56.80	76.70	37.30	41.50
SVDNet+Era# (ResNet-50) [48]	62.44	79.31	43.50	48.71
PCB (UP)*# (ResNet-50) [15]	66.10	81.80	54.20	61.30
PCB (UP)*#+re-rank (ResNet-50)	79.89	85.11	58.96	61.55
Ours (ResNet-50)	64.09	81.73	45.01	49.71
Ours + re-rank (ResNet-50)	**81.14**	**86.04**	**60.84**	**61.57**

Table 3. Effectiveness of using the OpenPose based misalignment correction on Market1501 dataset.

Method	Single Query		Multiple Query	
	mAP	rank-1	mAP	rank-1
Ours *without* Alignment	73.32	90.65	80.52	93.65
Ours *with* Alignment	**74.57**	**91.51**	**81.78**	**94.71**

D. PERFORMANCE ANALYSIS

We further evaluate the impact of several important parameters used in our method to demonstrate the effectiveness of each of them. Since the Market1501 dataset allows the implementation of person ReID for a pedestrian retrieval task, the evaluation is performed on this dataset under single and multiple query settings.

1) Effectiveness of Alignment

Table 3 compares the mAP and rank-1 accuracies obtained with our model without using the OpenPose technique ("Ours *without* Alignment"), where the misaligned pedestrian images are not corrected. Using the OpenPose ("Ours *with* Alignment"), images' excess background is removed and body parts are aligned before being input into the training model. Clearly, the performance using alignment is better than not using alignment, with an improvement of 1.25% mAP and 0.86% rank-1 accuracy, respectively, with single query mode. After a further investigation, it is found there are approximately 2% of images that are misaligned before using alignment in the Market1501 dataset.

2) Effectiveness of Different Body Parts and Their Weighted Fusion

In order to understand the contributions of different body parts to the accuracy of person ReID, we conduct experiments on the Market1501 dataset with separated body-parts to empirically show how the various parts of a body influence the overall performance differently.

We train five different network models, corresponding to the five different parts of the body from top to bottom as shown in Fig. 4. Table 4 compares the mAP, and rank-1 and rank-5 accuracies obtained with different models trained with

Table 4. The comparison of mAP and rank-1, rank-5 accuracies of person ReID obtained on Market1501 dataset using models trained with the features extracted from the five different body parts. B refers to the features extracted from the whole body. B_i $(i = 1, 2, 3, 4, 5)$ denotes the features extracted from each of the five body parts. "*B without* B_1" refers to the features calculated from the whole body without the head region. "*B without* B_5" refers to the features calculated from the whole body without the foot region.

Feature	Single Query			Multiple Query		
	mAP	rank-1	rank-5	mAP	rank-1	rank-5
B_1	35.98	48.28	67.40	41.90	60.54	79.56
B_2	49.89	65.26	78.11	61.40	78.24	87.59
B_3	52.48	68.35	78.56	60.09	75.18	87.50
B_4	54.95	72.76	81.62	66.68	79.16	89.25
B_5	35.89	53.92	70.40	53.59	69.74	83.49
B	**88.50**	**92.96**	**95.69**	**91.75**	**95.31**	**97.24**
B without B_1	86.41	91.92	95.55	90.16	93.97	97.00
B without B_5	84.73	91.15	94.83	89.12	93.65	96.35

Table 5. Effectiveness of using the complementary advantages of different features on Market1501 dataset.

Feature Type	Single Query		Multiple Query	
	mAP	rank-1	mAP	rank-1
Ours *without* overlap or body-weight	69.62	88.66	77.81	92.71
Ours + 10% overlap	70.38	88.91	78.43	92.82
Ours + 20% overlap	71.49	89.90	79.32	93.47
Ours + 25% overlap	71.37	89.29	79.04	93.53
Ours + equal-weight	72.02	89.78	79.27	93.35
Ours + bodyweight	73.15	90.53	80.41	93.76
Ours + 20% overlap + body-weight	74.57	91.51	81.78	94.71
Ours + re-rank	**88.50**	**92.96**	**91.75**	**95.31**

different body parts and their combination. As detailed in this table, the discriminative degrees of the feature descriptors calculated from the head (i.e., B_1 in the table) and foot regions (i.e., B_5 in the table) are weaker than those obtained from the upper and lower body regions (i.e., B_2, B_3 and B_4 in the table), for both single query and multiple query modes. Apparently, different body parts (i.e., head, shoulder, abdomen, leg, foot and full-body) contribute differently to the person ReID task. The head and foot regions provide less reliable features, so they are not as significant for differentiating different persons as other body parts.

The experimental results (corresponding to "*B without* B_1" and "*B without* B_5" in Table 4) show the limited impact on the final accuracy when the head region or foot region is removed from the pedestrian images. Examining the samples in this dataset, we have observed that many faces are not with frontal view and the resolution of faces is too low. Therefore, their discriminative degree is very limited.

In our work, to reflect the contributions of different body parts to the overall performance, we propose to use a weighted fusion of feature descriptors, and set weights empirically. To do this, we use the method in [22] to normalize the results of Table 4 and empirically obtain five normalized weights for different parts of a body, i.e., 0.1645, 0.2096, 0.2373, 0.2405 and 0.1481 on Market-1501 dataset, respectively. According to our experiments, the weights for different parts of a body are nearly the same for different datasets.

Moreover, we observe that a certain percentage of over-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

Table 6. Effectiveness of using different loss functions on Market1501 dataset (without using re-ranking).

Loss Types	Single Query		Multiple Query	
	mAP	rank-1	mAP	rank-1
Classification Loss	73.32	90.65	80.52	93.65
Triplet Loss	64.38	82.46	72.96	86.43
Angular Loss	72.66	90.83	80.19	93.82
Classification + Angular Loss	**74.57**	**91.51**	**81.78**	**94.71**

lapping between body part regions also helps to improve the performance. Table 5 shows the comparison results of using weighted fusion of body-part features (indicated as "bodyweight" in the table) with and without overlapping ("overlap" in the table) between body parts. In this table, "Ours *without* overlap or bodyweight" refers to the case when neither of the overlapping nor weighted body-part feature strategies is used. "Ours + 10% overlap", "Ours + 20% overlap" and "Ours + 25% overlap" represents there is 10%, 20% and 25% overlapping between neighboring body part regions respectively. "Ours + equal-weight" represents equal weighted fusion of features obtained from different body parts. "Ours + bodyweight" represents weighted fusion of features obtained from different body parts. "Ours + 20% overlap + bodyweight" means our method with both overlapping and bodyweight strategies. "Ours + re-rank" indicates using re-ranking [41] method on the basis of "Ours + 20% overlap + bodyweight". As shown in Table 5, 20% overlapping between body parts get better performance.

It can be seen from Table 5 that using the weighted fusion of feature descriptors produces better accuracy than using the equal-weight fusion. This means that, with proper weighting, features of different body parts can be fused in a more effective way and are helpful in improving person ReID performance. Table 5 also shows that the feature descriptors are helpful in improving performance with proper overlapping. Simultaneously using the overlapped and weighted body parts, we have observed a consistent improvement on the Market1501 dataset. Our method gains 74.57% on mAP and 91.51% on rank-1 matching rate under the single query mode. When combined with an effective re-ranking approach, the performance has reached to a rank-1 rate of 92.96%.

3) Effectiveness of Loss Selection

In addition, using the complementary advantages of classification and angular loss is another important aspect of our approach. As mentioned earlier in Section III, they can be combined to capture different pedestrian characteristic from the aligned images and improve person ReID accuracy. We follow the settings in Section III, and compare the mAP and rank-1 accuracy obtained with different loss functions on the Market1501 dataset. The results are shown in Table 6.

As shown in this table, using classification or angular loss alone can achieve a rank-1 accuracy of 90.65% and 90.83%, respectively. Using the traditional triplet loss, the performance of rank-1 accuracy is much worse. Without

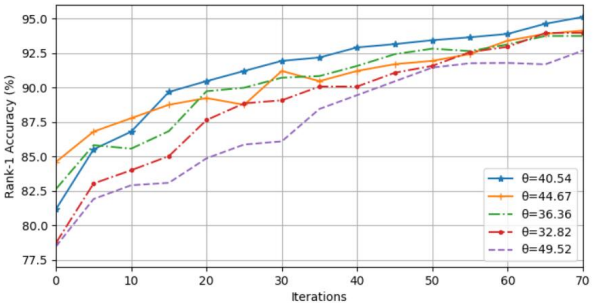

FIGURE 5. Comparison of rank-1 accuracies with different values of θ for angular loss during training on Market1501 dataset.

Table 7. Comparison of different values θ of angular loss during training on Market1501 dataset.

Angle θ	mAP	rank-1	rank-5	rank-10
$\theta = 49.52°$	85.62	92.68	96.44	97.49
$\theta = 44.67°$	88.41	94.13	98.04	98.78
$\theta = 40.54°$	**88.87**	**95.12**	**97.80**	**99.27**
$\theta = 36.36°$	87.13	93.74	98.53	98.14
$\theta = 32.82°$	85.49	93.97	96.58	97.21

combining part-based methods, the angular loss method alone is also comparable to most of the recently methods, as depicted in Table 1. Note that, without combination, classification loss preforms comparably with angular loss, and with the combination, the classification loss and angular loss together perform better than using each of them alone. Both classification and ranking information are important to learn discriminative features for person ReID. Using the global feature representation combined with the part-based representation enriches some fine details.

The angular loss function involves one hyperparameter θ (Eq. 4), which determines the degree of the constraint. Table 7 shows the effect of hyperparameter θ on the overall accuracy. θ is set by the hyperparameter optimizer sklearn. Setting θ to be 40.54° for the Market1501 dataset leads to the best performance for our method. Fig. 5 shows the comparison of the rank-1 accuracies with different values of parameter θ for angular loss during training on the Market1501 dataset. In the experiments, our method always performs stably well when the value of parameter θ is set to be between 30° and 50°.

As depicted in Fig. 6, we further visualize some retrieval results on the three datasets, i.e., Market1501 [23], DukeMTMC-reID [27], [28] and CUHK03 [1]. The images in the first column are the query images. The retrieved images are sorted and shown in the second to the eleventh columns according to the similarity scores in the order of high to low. The correct and false matches are shown in green and red bounding boxes (best viewed in color), respectively. As shown in this figure, most candidate images can be retrieved correctly. The DukeMTMC-reID and CHUK03 datasets are more challenging, which contain pedestrians with occlusions and similar appearance. Therefore, the proposed model has retrieved some incorrect candidates.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2864588, IEEE Access

FIGURE 6. Examples of pedestrian retrieval results on three datasets using the proposed method in single query mode. The images in the first column are the query images. The top-10 retrieved images are sorted and shown in the second column to the eleventh column according to the similarity scores from high to low. The correct and false matches are shown in green and red bounding boxes (best viewed in color), respectively.

V. CONCLUSION AND FUTURE WORK

This paper has proposed a two-branch deep architecture which leverages the human part cues to learn highly discriminative features and similarity measurements for person ReID. The OpenPose toolkit has been employed to mitigate the body misalignment problem of pedestrian images. For the feature representation, weighted body-part feature fusion and global full-body feature descriptors are jointly employed for better performance. We have shown that the weighted and overlapping body-part feature representation are informative to capture discriminative details of pedestrian images. The classification loss and angular loss have been applied to simultaneously learn discriminative similarity measurement in a unified framework. Extensive comparative evaluations on three benchmark datasets have demonstrated the superiority of the proposed method over the state of the arts.

Even if deep learning features, such as the one extracted using CNN are very powerful, still they are not very powerful at extracting similar features regardless of the viewpoint. In our future work, we hope to extract the features of pedestrian using the latest Capsule Networks [49].

REFERENCES

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 152-159.

[2] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-localalignment descriptor for pedestrian retrieval," in Proceedings of the 2017 ACM on Multimedia Conference, 2017, pp. 420-428.

[3] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 14, no. 1, p. 13, 2017.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person reidentification by multi-channel parts-based cnn with improved triplet loss function," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1335-1344.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification." in AAAI, vol. 1, no. 2, 2017, p. 3.

[6] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv preprint arXiv:1610.02984, 2016.

[7] Y. Yang, L. Wen, S. Lyu, and S. Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification." in AAAI, vol. 1, 2017, p. 2.

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

[9] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 3376-3385.

[10] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 403-412.

[11] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv:1703.07737, 2017.

[12] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," arXiv preprint arXiv:1710.00478, 2017.

[13] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in Proc. Int. Joint Conf. on Artif. Intell. (IJCAI), 2017.

[14] 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 Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1077-1085.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," arXiv preprint arXiv:1711.09349, 2017.

[16] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in Proc. Int. Conf. Comput. Vis. (ICCV), 2017, pp. 3980-3989.

[17] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 384-393.

[18] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," arXiv preprint arXiv:1711.08184, 2017.

[19] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in Proc. Int. Conf. Comput. Vis. (ICCV), 2015, pp. 2399-2406.

[20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proc. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, no. 2, 2017, p. 7.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.

[22] 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 Proc. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1741-1750.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1116-1124.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 815-823.

[25] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in Advances in Neural Inf. Process. Syst., 2016, pp. 1857-1865.

[26] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in Proc. Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2593-2601.
[27] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in Proc. Int. Conf. Comput. Vis. (ICCV), 2017.
[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 17-35.
[29] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in Proc. Int. Conf. Comput. Vis. (ICCV), 2017.
[30] W. Lin, Y. Shen, J. Yan, M. Xu, J. Wu, J. Wang, and K. Lu, "Learning correspondence structures for person re-identification," IEEE Transactions on Image Processing, vol. 26, no. 5, pp. 2438-2453, 2017.
[31] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 34-50.
[32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3492-3506, 2017.
[33] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," arXiv preprint arXiv:1711.10658, 2017.
[34] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in Proc. Eur. Conf. Comput. Vis. (ECCV). Springer, 2016, pp. 135-153.
[35] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y. G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," arXiv preprint arXiv:1712.02225, 2017.
[36] J. Guo, Y. Zhang, Z. Huang, and W. Qiu, "Person re-identification by weighted integration of sparse and collaborative representation," IEEE Access, vol. 5, pp. 21632-21639, 2017.
[37] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via regionbased fully convolutional networks," in Advances in Neural Inf. Process. Syst., 2016, pp. 379-387.
[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Int. Conf. on Machine Learning (ICML), 2015.
[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.
[40] D. P. Kingma and J. B. Adam, "A method for stochastic optimization, 2014," in Int. Conf. on Learning Representations (ICRL), 2015.
[41] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person reidentification with k-reciprocal encoding," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 3652-3661.
[42] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2590-2600.
[43] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 791-808.
[44] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in Proc. Comput. Vis. Pattern Recognit. (CVPR), vol. 6, 2017.
[45] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in Proc. Comput. Vis. Pattern Recognit. (CVPR), vol. 6, 2017.
[46] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in Proc. Comput. Vis. Pattern Recognit. (CVPR), vol. 6, 2017, p. 7.
[47] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in Proc. Int. Conf. Comput. Vis. (ICCV), 2017.
[48] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," arXiv preprint arXiv:1708.04896, 2017.
[49] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in Neural Information Processing Systems. (NIPS), 2017.

YUANYUAN WANG is currently pursuing her Ph.D degree in Computer Science and Technology from College of Computer and Information, Hohai University of China. She received her M.S. degree in Computer Technology from Nanjing University of Science and Technology in 2010. She is currently a Lecturer with the College of Computer and Software Engineering of Huaiyin Institute of Technology of China. Her current research focuses on computer vision and person re-identification.

ZHIJIAN WANG received the M.S. and Ph.D. degree in Computer Science from Nanjing University, China. He is currently a Professor at the College of Computer and Information, Hohai University, China. His research interests include machine learning and computer application.

WENJING JIA received her Ph.D. degree in Computing Science from the University of Technology Sydney (UTS) in 2007. She is currently a Senior Lecturer at the Faculty of Engineering and IT and a Core Research Member at the Global Big Data Technologies Centre, UTS. She has authored over 100 quality journal articles and conference papers. Her research interests include image / video analysis, computer vision and pattern recognition.

XIANGJIAN HE (M'99-SM'05) received his Ph.D. degree in Computer Science from the University of Technology Sydney (UTS), Australia, in 1999. He is currently a Full Professor and the Director of the Computer Vision and Pattern Recognition Laboratory with the Global Big Data Technologies Centre (GBDTC), UTS.

MINGXIN JIANG received the Ph.D. degree in Signal and Information Processing, Dalian University of Technology, China, in 2013. She was a post-doctoral researcher with the Department of Electrical Engineering in Dalian University of Technology from 2013 to 2015. She is currently an Associate Professor in College of Electronic Information Engineering at Huaiyin Institute of Technology. Her research interests include multi-object tracking and vision sensors for robotics.