Concentrated Local Part Discovery With Fine-Grained Part Representation for Person Re-Identification

Chaoqun Wan, Yue Wu, Xinmei Tian, *Member, IEEE*, Jianqiang Huang, and Xian-Sheng Hua, *Fellow, IEEE*

***Abstract*—The attention mechanism for person re-identification has been widely studied with deep convolutional neural networks. This mechanism works as a good complement to the global features extracted from an image of the entire human body. However, existing works mainly focus on discovering local parts with simple feature representations, such as global average pooling. Moreover, these works either require extra supervision, such as labeling of body joints, or pay little attention to the guidance of part learning, resulting in scattered activation of learned parts. Furthermore, existing works usually extract local features from different body parts via global average pooling and then concatenate them together as good global features. We find that local features acquired in this way contribute little to the overall performance. In this paper, we argue the significance of local part description and explore the attention mechanism from both local part discovery and local part representation aspects. For local part discovery, we propose a new constrained attention module to make the activated regions concentrated and meaningful without extra supervision. For local part representation, we propose a statistical-positional-relational descriptor to represent local parts from a fine-grained viewpoint. Extensive experiments are conducted to validate the overall performance, the effectiveness of each component, and the generalization ability. We achieve a rank-1 accuracy of 95.1% on Market1501, 64.7% on CUHK03, 87.1% on DukeMTMC-ReID, and 79.9% on MSMT17, outperforming state-of-the-art methods.**

***Index Terms*—Person re-identification, local part learning, constraint attention mechanism, fine-grained representation.**

I. Introduction

PERSON re-identification (ReID) systems aim to identify people across different cameras and temporal periods. Given a query image of a person, a complete person ReID system should find all images of the same person that appeared in different scenarios and time periods. With the increasing construction of urban surveillance cameras and the demand for public security, person ReID provides the fundamental ability for real-time video monitoring and historical video analysis to detect pedestrian trajectories and discover criminal suspects. Person ReID has attracted considerable research attention in computer vision.

Recent years have witnessed the trend of discovering human body parts and incorporating the corresponding features as a complement to the global features for better distinctiveness. Global features are often extracted from the entire image in a handcrafted manner [4]–[7] or by feeding the image into a deep convolutional neural network (CNN) [8], [9]. Moreover, some methods based on metric learning have been proposed to enhance the feature representation [10]–[13]. Though dissimilar appearances can be easily distinguished, the global features are poorly distinctive for different pedestrians with similar appearances. Some works make a pixelwise or rigidwise comparison of original images or feature maps to exploit some details to compensate for the global features [14]–[16]. However, such local features are sensitive to large variations in illumination, occlusion, pose, scale and camera view. For example, in the first column of Fig. 1, the same body parts, such as the head, torso, arms or legs, are located in different positions in different images. Therefore, the pixelwise or rigidwise comparison is not effective in aligning the local parts or avoiding background interference.

To overcome these difficulties, recent works have focused on local part discovery to enhance the overall effectiveness. These approaches can be summarized into three main categories: image-level region proposal-based methods [18]–[21], feature-level rigid-based methods [22], [23] and feature-level attention-based methods [1]–[3], [24], [25]. The image-level region proposal-based methods discover local parts from the original images based on the human body structure, the feature-level rigid-based methods divide the feature maps into a group of horizontal rigid structures, and the feature-level attention-based methods employ attention mechanisms to generate attention maps that directly emphasize the local parts on feature maps. Nevertheless, the region proposal-based methods either depend on extra supervision, such as labeling of body joints [18], [20], [21], or employ complicated operations, such as affine transformation [19]. They require a high computational cost, especially for online systems. Regarding the rigid-based methods, the manual division cannot address the deformable local parts and the background noise. Regarding the attention-based methods, explicit guidance is hardly provided for learning attention maps,

Manuscript received February 22, 2019; revised July 1, 2019 and August 17, 2019; accepted September 24, 2019. Date of publication October 10, 2019; date of current version May 21, 2020. This work was supported in part by National Key Research and Development Program of China under Grant 2017YFB1002203 and in part by the National Natural Science Foundation of China under Grants 61872329 and 61572451. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marco Carli. *(Corresponding author: Xinmei Tian.)*

C. Wan and X. Tian are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: wancq14@mail.ustc.edu.cn; xinmei@ustc.edu.cn).

Y. Wu, J. Huang, and X.-S. Hua are with the Alibaba Group, Hangzhou 311121, China (e-mail: matthew.wy@alibaba-inc.com; jianqiang.hjq@alibaba-inc.com; xiansheng.hxs@alibaba-inc.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2946486

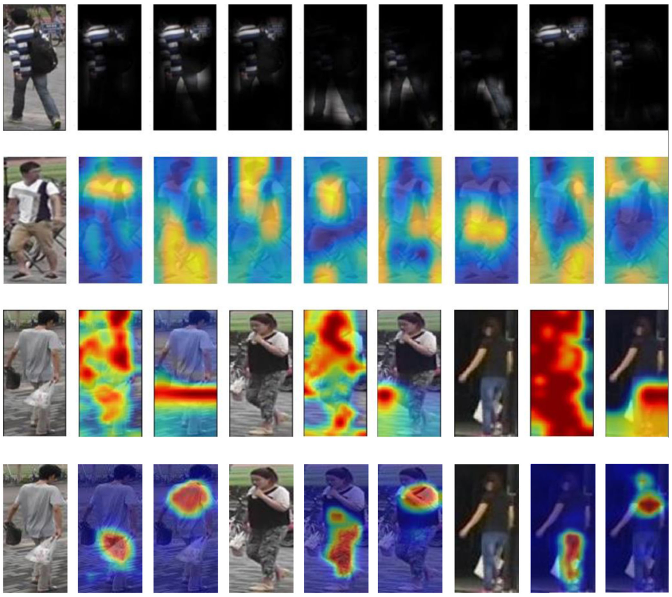


Fig. 1. Visualization of local part discovery. In the first two rows, there is 1 set of visualization results for each row, and in the last two rows, there are 3 sets. In each set, the first column is the original image, and the others are visualizations of learned attention maps. The first three rows are visualization results from DLPAN [1], HP-Net [2], and HA-CNN [3]. The last row is the visualization result obtained by the proposed concentrated attention module.

resulting in unconcentrated or meaningless activation. Some examples are shown in the first three rows of Fig. 1, where the activation regions on the attention maps are almost scattered with much noise on the background. These issues will limit the representation ability of local features and even produce an undesirable impact on the final results.

In addition to accurate local part discovery, distinctive local feature representation is another significant tool to enhance the feature ability. Most previous methods simply adopt global average pooling (GAP) to map convolutional feature maps into feature vectors. Although this operation has been proven to be effective after years of practice, some works [26], [27] still reveal drawbacks of such deep representations for fine-grained learning, especially for the features of highly related local parts. Moreover, according to our experiment, features of local parts obtained by global average pooling contribute little to the overall performance, particularly when global features are well learned. The reasons are twofold: 1) The result of global average pooling is highly concise, which neglects significant local cues, such as location, distribution and correlation. People who resemble each other only have slight differences in body parts, and global average pooling is too weak to capture such differences. 2) Local part description is meant to highlight a completely different viewpoint to avoid becoming an ensemble of global features. This description should have the ability to describe the detailed cues of body parts and a cooperative representation ability for their relations.

In this paper, we address the two key problems of the attention mechanism for person ReID: discovering precise local parts and exploring substantial local cues. We solve the problems by observing that 1) semantic local parts are always small concentrated regions, such as the head, arm, leg and torso, and 2) humans determine precise identifications from detailed comparisons and even the relation of body parts. For local part discovery, the traditional attention module produces unconcentrated or meaningless activations as previously mentioned, resulting in a limited local representation. Since the local parts are always small concentrated regions, we propose an iterative concentration process for the attention module. This process iteratively filters out the farthest point on learned local parts and forces the activation to be concentrated on these local parts. In addition, we redesign the attention module with a multiscale architecture to overextract more possible regions of body parts. The multiscale attention module with the iterative concentration process is termed the concentrated attention module. For local part description, we propose a statistical-positional-relational (SPR) descriptor to delve into the meticulousness and distinctiveness of local cues, as well as their cooperative representation ability. Specifically, we first extract 5 statistics as statistical features to describe the local activation, containing the values of the mean, variation, median, min and max for each feature map. Since different parts have relatively different ranges of position, we use the barycenter of attention maps as the positional feature to roughly avoid misalignment of body parts. In addition, we utilize the invariance to pose of human body parts [26], [28] and propose a relational feature to explore such invariance for cooperative representation. The statistical feature, positional feature and relational feature are fused as the SPR descriptor for fine-grained local part description. Fig. 2 illustrates the whole framework.

The main contributions are summarized as follows:

- We propose to explore the attention mechanism from two aspects, and our method outperforms state-of-the-art approaches.
- We propose a new constrained attention module that can guide the attention maps to obtain concentrated local parts over body parts.
- We employ a novel SPR descriptor for fine-grained local part representation, which achieves better performance than the global features and significantly improves the overall performance.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III presents the proposed constrained attention module and the fine-grained local part representation in detail. Section IV presents extensive comparisons with state-of-the-art methods and analysis of the proposed method . Section V concludes the paper.

II. Related Work

A. Deep Learning-Based Methods

The success of deep learning in image classification has inspired a large number of studies on person ReID. Many approaches have been designed based on the convolutional neural network structure. For instance, Li *et al.* propose a filter-pairing neural network (FPNN) with a rigidwise patch-matching layer to solve the problem of the pose and viewpoint variance [15].

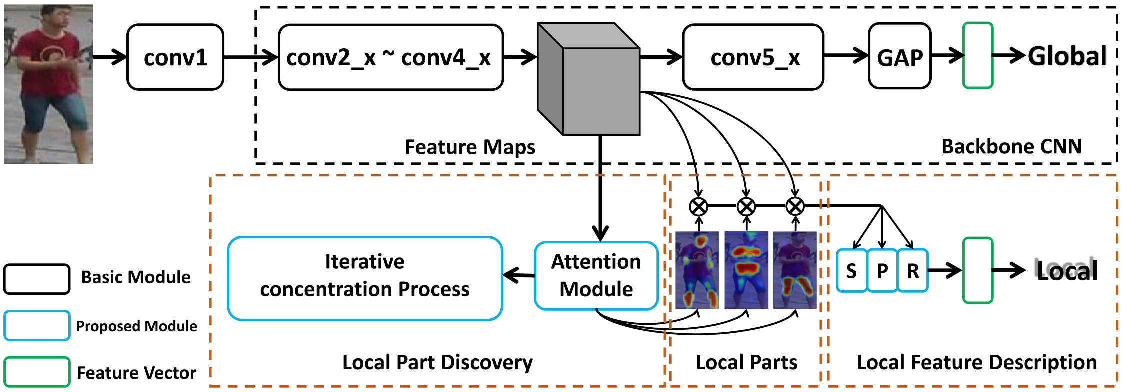


Fig. 2. Overview of the proposed network architecture based on ResNet-50 [17]. The black rectangles denote the basic modules in ResNet-50, where "GAP" means global average pooling. The blue rectangles denote the new proposed modules for local part discovery and local feature description, where "S," "P" and "R" represent the statistical, positional and relational feature, respectively. The green rectangles denote the global and local features. The proposed attention mechanism learns the local description from a middle layer of a backbone convolutional network through the multiscale attention module with an iterative concentration learning process to generate concentrated and meaningful outputs. The statistical-positional-relational descriptor then extracts corresponding features to obtain substantial local information. The global and local branches are learned separately for better discriminability.

Wu *et al.* design a double-column architecture with pixelwise neighborhood patch matching to extract features from image pairs [14]. However, these shallow networks can hardly exploit the full potential of a CNN. Current works are more inclined to take advantage of the classic deep neural networks, such as VGG [29], GoogLeNet [30] and ResNet [17]. With the model pretrained on ImageNet and fine-tuned for person ReID, the extracted features are more discriminative. Nevertheless, these methods represent the person from a global perspective or additionally embed pixelwise or rigidwise matching for a detailed comparison. As discussed in the previous section, these features are sensitive to variations in illumination, pose, scale and camera view.

B. Local Part Discovery

Studies on local part discovery can be summarized into three categories: image-level region proposal-based methods [18]–[21], feature-level rigid-based methods [22], [23] and feature-level attention-based methods [1]–[3], [24], [25]. Regarding region proposal-based methods, Zhao *et al.* use a region proposal network to localize 14 body joints and obtain 7 body parts [18]. Wei *et al.* use the pose estimation method to locate four key points and induce three local parts [21]. To avoid the dependence on auxiliary labels, such as body joints, Li *et al.* employ a transformer network, similar to the spatial transformer network [31], to automatically locate three predefined body parts (head, torso and legs) [19]. By not locating body parts with region proposals on the image level, another solution is to directly discover local parts from the feature level for simplicity. One practical approach is to divide the feature maps into a series of horizontal rigid structures. Zhang *et al.* horizontally divide the feature maps into several stripes and design an alignment algorithm to avoid mismatch [22]. Sun *et al.* conduct an identical division and further investigate the texture of each stripe [23]. They propose a refinement algorithm to generate soft boundaries. However, the horizontal division still suffers from noise regions on the background. To address this problem, the attention module is an alternative approach to learn activation masks that emphasize distinctive regions. Zhao *et al.* design a simple attention module on the last convolutional layer to adaptively discover local parts from the feature maps [1]. Liu *et al.* consider more spatial information and propose an HP-Net to multidirectionally feed the multilevel attention maps into the corresponding layers [2]. Li *et al.* consider hard local regions and propose a harmonious attention model for joint learning of the hard regional attention and soft pixel attention [3]. These methods extract local parts without guidance, which easily results in a scattered distribution of attention maps.

C. Local Part Representation

Traditional handcrafted methods design the local features from various aspects for better representation. Gray and Tao extract color and texture features on horizontal stripes [32]. Liao *et al.* propose the local maximal occurrence (LOMO) to analyze the horizontal occurrence of local features for a stable description against viewpoint change [33]. Chen *et al.* adopt polynomial kernels in the linear similarity function to obtain a robust local feature description [34]. Most recent methods with convolutional neural networks adopt global average pooling or global max pooling for the local description. However, this concise description will discard most of the original information, which is significant for the discrimination of local parts of humans. Carreira *et al.* design a second-order pooling representation that is better for local description [27]. Lin *et al.* further confirm the previous viewpoint and propose a bilinear convolutional neural network for fine-grained object detection [26]. As discussed in the previous section, local part representation is expected to delve into the meticulousness and distinctiveness of local cues, as well as their cooperative representation ability.

III. Approach

In this section, we introduce the proposed constrained attention module for local part discovery and the statistical-positional-relational descriptor for fine-grained local part

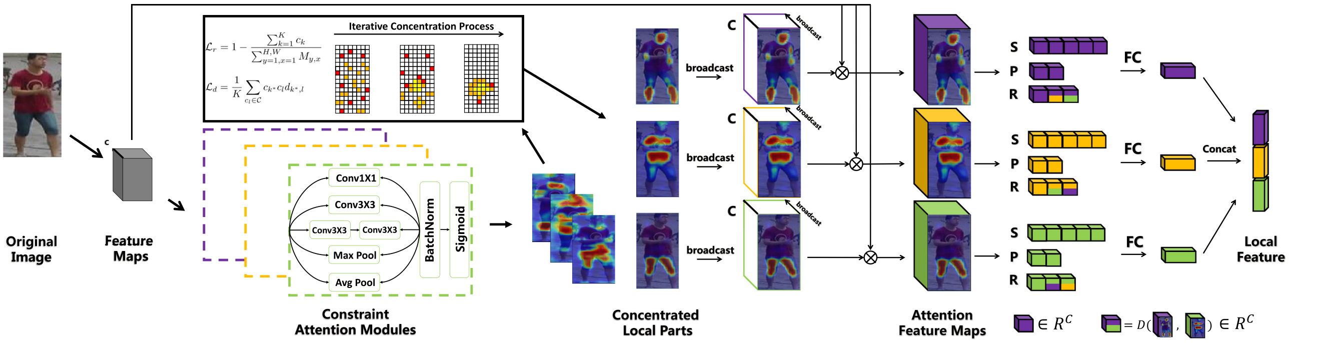

Iterative Concentration Process
$\mathcal{L}_r = 1 - \frac{\sum_{k=1}^{K} c_k}{\sum_{y=1,x=1}^{H,W} M_{y,x}}$
$\mathcal{L}_d = \frac{1}{K}\sum_{c_l \in \mathcal{C}} c_{k^*} c_l d_{k^*,l}$
Original Image
c
Feature Maps
Conv1X1
Conv3X3
Conv3X3
Conv3X3
Max Pool
Avg Pool
BatchNorm
Sigmoid
Constraint Attention Modules
broadcast
C
broadcast
Concentrated Local Parts
Attention Feature Maps
S
P
R
FC
Concat
Local Feature
$\in R^C$
$= D(\ ,\ ) \in R^C$

Fig. 3. The architecture of the concentrated attention mechanism. The constrained attention modules take the middle-level feature maps as the input and produce concentrated local parts as outputs with the aid of an iterative concentration process. Then, the concentrated local parts are broadcast and multiplied with the middle-level feature maps to form the part-based attention feature maps. Finally, the SPR descriptor extracts local cues to form the local feature.

representation in detail. Fig. 2 shows the framework, which consists of a backbone network for global feature extraction and the local attention mechanism for body part discovery and representation.

A. Constrained Attention Module

For local part discovery, we propose a constrained attention module for better part localization. The entire architecture is shown in Fig. 3, including the multiscale attention module and the iterative concentration process. Compared with the traditional methods of local part discovery, there are two differences in the proposed approach:

- We adopt a multiscale architecture for comprehensive perception. The output will overextract regions of interest of various scales.
- We design an unsupervised iterative concentration process to collect the activation of attention maps for learning concentrated and meaningful local parts.

Both of these components are necessary to obtain concentrated and even meaningful local parts. The multiscale architecture ensures that most of the possible regions are perceived; thereby, the iterative concentration process can gradually eliminate the invalid regions and produce concentrated activation. More details are discussed as follows.

1) Multiscale Attention Module: We design a multiscale perception architecture as shown in Fig. 3. This architecture is composed of five types of operations, including max pooling, average pooling and three convolutional operations with kernel sizes of 1×1, 3×3 and 5×5. Unlike the traditional inception architecture [30], we additionally use max pooling to reserve more possible regions with large responses. Intuitively, max pooling is able to describe the maximum outline and implies as many candidate regions of body parts as possible. The proposed architecture is more complex than others, such as [1]–[3], because we use feature maps after a single convolutional layer from the middle of the backbone network. Such middle-level features contain more spatial cues but less semantic information. A relatively complex attention module should be able to compensate for the semantic loss to some extent.

2) Iterative Concentration Process: Without explicit guidance, the above multiscale architecture learns freely. There is no guarantee that the resulting activation maps are meaningful and concentrated around body parts, leading to a mixture of both true positive regions and false positives, as shown in Fig. 4(b). To solve the concentration problem, we propose constraints from two aspects.

Specifically, given an image $\mathcal{I}$, the feature maps of a middle layer are denoted $X \in \mathcal{R}^{C\times H\times W}$. We denote the five types of kernels as $\mathcal{K}_{\max}, \mathcal{K}_{avg}, \mathcal{K}_{1\times 1}, \mathcal{K}_{3\times 3}$, and $\mathcal{K}_{5\times 5}$. $\mathcal{K}_*(X)$ represents the corresponding convolutional operation $* \in \{max, avg, 1\times 1, 3\times 3, 5\times 5\}$ on input X. The learned attention map from X is:

$$\begin{aligned} M &= \mathcal{F}_A(X) \\ &= \sigma(\mathcal{F}_{BN}(\mathcal{R}_{1\times 1}(\mathcal{K}(X)))), \end{aligned} \tag{1}$$

$$\begin{aligned} \mathcal{K}(X) = \mathcal{F}_{concat}(&\mathcal{K}_{\max}(X), \mathcal{K}_{avg}(X), \mathcal{K}_{1\times 1}(X), \\ &\mathcal{K}_{3\times 3}(X), \mathcal{K}_{5\times 5}(X)), \end{aligned} \tag{2}$$

where $\mathcal{F}_A : \mathcal{R}^{C\times H\times W} \Rightarrow \mathcal{R}^{H\times W}$ represents the function of the attention module, $\mathcal{R}_{1\times 1} : \mathcal{R}^{5\times H\times W} \Rightarrow \mathcal{R}^{H\times W}$ is a dimension reduction mapping, $\mathcal{F}_{BN}$ is a batch normalization transform, $\mathcal{F}_{concat}$ is a concatenation operation along the first dimension, and σ is the sigmoid function.

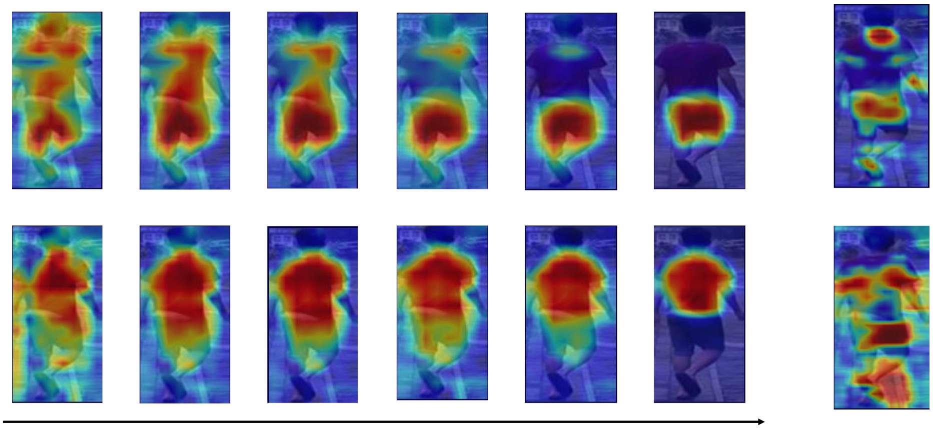
Suppose there are N multiscale attention modules, denoted $\{\mathcal{F}_A^1, \mathcal{F}_A^2, \ldots, \mathcal{F}_A^N\}$. The N learned single channel attention maps are:

$$\mathcal{M} = \{M^i = \mathcal{F}_A^i(X) | i \in \mathbb{N}\}. \tag{3}$$

Each attention map M^i is then broadcast to C channels $\tilde{M}^i$ and multiplied with the original feature maps X to generate the attention feature maps:

$$\mathcal{Z} = \{Z^i = \tilde{M}^i \otimes X | i \in \mathbb{N}\},\ Z^i \in \mathcal{R}^{C\times H\times W}. \tag{4}$$

Activation ratio constraint. For simplicity, we denote $M \in \mathcal{R}^{H\times W}$ as an arbitrary attention map. Since a body part occupies small local regions, only a portion of the attention map should be activated with large values. Suppose there are at most K activated points for a body part. The value of K may be different for different parts. We select K activated points with the largest activation values as the candidate points of the body part. The



(a) Iterative Concentration Process

(b)

Fig. 4. Visualization of the iterative concentration process. (a) Examples of how activation maps are concentrated on local parts through the iterative concentration process. (b) Results without the constrained attention module.

candidate points are denoted:

$$\mathcal{C} = \{c_k = \mathcal{T}_k(M) | k \in \mathbb{K}\}, \tag{5}$$

where c_k represents the kth candidate point as well as the corresponding value for simplicity. $\mathcal{T}_k(M)$ denotes the function of finding the k-th maximum value in M. We expect that the majority of activation is focused on these candidates to constrain the ratio of local parts on attention maps. The activation ratio constraint is defined by the following loss function:

$$\mathcal{L}_r = 1 - \frac{\sum_{k=1}^{K} c_k}{\sum_{y=1,x=1}^{H,W} M_{y,x}}. \tag{6}$$

The gradient for backpropagation is:

$$\frac{\partial \mathcal{L}_r}{\partial M_{y,x}} = \begin{cases} \dfrac{\sum_{c_k \in \mathcal{C}} c_k - \sum_{y=1,x=1}^{H,W} M_{y,x}}{\left(\sum_{y=1,x=1}^{H,W} M_{y,x}\right)^2}, & M_{y,x} \in \mathcal{C}, \\ \dfrac{\sum_{c_k \in \mathcal{C}} c_k}{\left(\sum_{y=1,x=1}^{H,W} M_{y,x}\right)^2}, & M_{y,x} \notin \mathcal{C}. \end{cases} \tag{7}$$

When $M_{y,x} \in \mathcal{C}$, the gradient is negative and $M_{i,j}$ is supposed to be larger; when $M_{y,x} \notin \mathcal{C}$, the gradient is positive, and $M_{i,j}$ is supposed to be smaller. As a result, the majority of the activation will appear on the candidates, achieving activation (energy) concentration with less noise activated.

Activation distribution constraint. Despite the activation ratio constraint, these candidates can still be scattered over the attention map, with true positive ones on body parts and false positive ones on the background. Therefore, we define a potential loss function to reflect the concentration degree of activated points:

$$\mathcal{L}_p = \frac{1}{K} \sum_{k=1}^{K} \mathcal{P}(c_k), \tag{8}$$

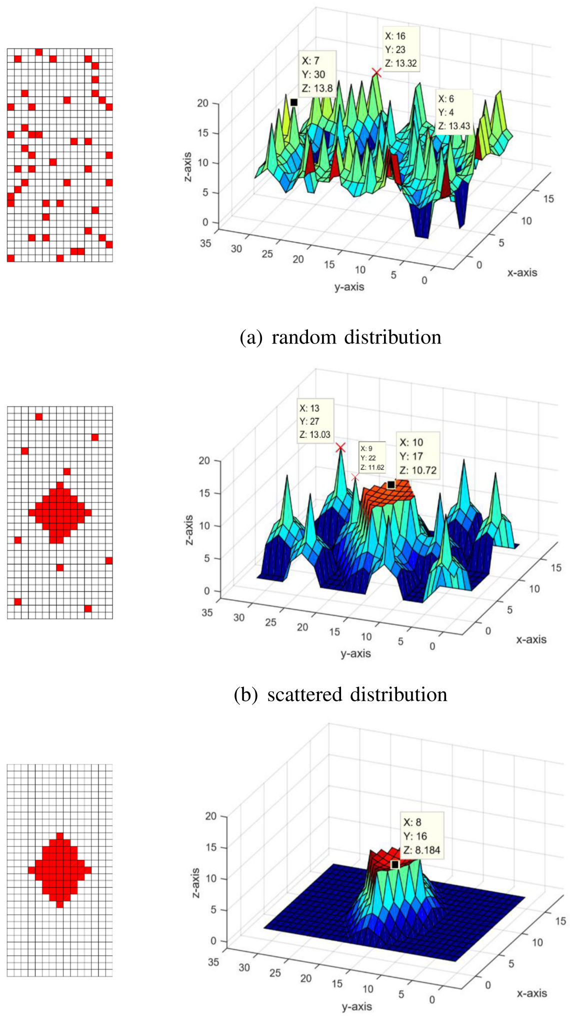
$$\mathcal{P}(c_k) = \frac{1}{K} \sum_{c_l \in \mathcal{C}} c_k c_l d_{k,l}, \tag{9}$$

where $d_{k,l}$ represents the Euclidean distance between c_k and c_l. Once c_k and c_l are clarified, the distance $d_{k,l}$ is determined. Therefore, $d_{k,l}$ is nondifferentiable with respect to c_k, and it makes no sense to calculate the gradient of $\frac{\partial \mathcal{L}_p}{\partial d_{k,l}}$. In addition, optimizing over C will minimize the values of all activated points. To address this problem, we reasonably assume that the candidate with the largest $\mathcal{P}(c_k)$ is the most likely outlier on the background. We optimize only one candidate with the largest $\mathcal{P}(c_k)$ in each iteration. Such an iterative training process can achieve a similar effect to gradually filtering out the most likely false positive candidates and alleviate the influence on true positive ones. Fig. 4(a) is a visualization of the iterative concentration process, where the outliers are gradually filtered out and candidates become concentrated. Accordingly, the activation distribution loss function is defined as:

$$\begin{aligned} \mathcal{L}_d &= \mathcal{P}(c_k^*) \\ &= \frac{1}{K} \sum_{c_l \in \mathcal{C}} c_{k^*} c_l d_{k^*,l}, \end{aligned} \tag{10}$$

$$k^* = \arg\max \mathcal{P}(c_i). \tag{11}$$

To illustrate the validity of converting Eq. (8) to Eq. (10), we exhibit the potential energy of attention maps under the condition of different distributions in Fig. 5. Images on the left side are quantified attention maps, where the red blocks represent the candidates with an activation value of 1, and the other regions correspond to a value of 0. Images on the right side are the potential energy of all candidates. Comparing the scattered distribution with the concentrated distribution, the potential energy in the scattered distribution is much higher. Furthermore, in the scattered distribution, the farther away the candidate is, the larger the potential energy will be (comparing the (X, Y, Z) of (13, 27, 13.03) and (9, 22, 11.62)). Thus, it is reasonable to gradually filter out the candidate with the largest potential energy so that the farthest one is eliminated and those remaining are concentrated with small potential energies as well as their sum. As a result, the two loss functions $\mathcal{L}_r$ and $\mathcal{L}_d$ are combined



(a) random distribution

(b) scattered distribution

(c) concentrated distribution

Fig. 5. An illustration of how the distribution affects the distance-related potential energy. The left sides are examples of feature maps with red blocks as candidates. The right sides are the potential energy distributions. (a) The random distribution has the highest potential energy for every candidate. (b) The scattered distribution has a relatively high potential energy, in which the farther the candidate is from the center, the larger the potential energy will be. (c) The concentrated distribution has the lowest potential energy as well as the lowest sum.

Algorithm 1: Iterative Concentration Process

```
Input: A training image and a corresponding label.
Output: Concentrated attention maps.
Initialize: Fine-tuned from the ImageNet pretrained model.
repeat
    Extract middle-level feature map X from a backbone network.
    for each attention module do
        Extract attention map: $M = \mathcal{F}_A(X)$;
        Select candidate points: $\mathcal{C} = \{c_k = \mathcal{T}_k(M)\}$;
        Calculate activation ratio loss $\mathcal{L}_r$ as in Eq. (6)
        Calculate potentials: $\mathcal{P}(c_k) = \frac{1}{K}\sum_{c_l \in \mathcal{C}} c_k c_l d_{k,l}$;
        Search the largest potential $\mathcal{P}(c_{k^*})$;
        Calculate activation distribution loss $\mathcal{L}_d$ as in Eq. (10)
    end for
    Calculate the loss of global and local branches;
    Calculate the total loss and gradient;
    Backward propagate the gradient and update the network;
until final iteration
```

to guide the attention modules to learn concentrated attention maps over local body parts. Algorithm 1 illustrates the iterative concentration process.

B. Statistic-Positional-Relational Descriptor

The widely used global average pooling is not suited for local part description for two reasons. First, local parts are often learned based on middle-level features to obtain more spatial cues at the expense of semantic information. Global average pooling is more suited for translation-invariant high-level feature extraction instead of location-sensitive local representation. Second, the local part should highlight a completely different representation viewpoint to avoid becoming a multiscale ensemble of global features and should focus on detailed cues of body parts as well as their cooperative representation ability. Therefore, we devise the statistical-positional-relational (SPR) descriptor to fully describe the local parts.

Statistical Feature. Various statistics will provide more precise descriptions of the attention feature maps, reducing the chance of confusion between similar body parts. A prior work has implemented a statistical layer to extract various statistics from the input feature vector and demonstrated the effectiveness of statistical features [35]. Specifically, we use five types of statistics along the channel dimension, including the maximum ($\mathcal{F}_{\max}$), minimum ($\mathcal{F}_{\min}$), mean ($\mathcal{F}_{mean}$), median ($\mathcal{F}_{median}$) and variance ($\mathcal{F}_{var}$) of the attention feature maps. Each of them produces a C-dimensional feature vector: $\mathcal{F}_* : \mathcal{R}^{H\times W\times C} \Rightarrow \mathcal{R}^C$. The statistical features $F_S^i \in \mathcal{R}^{5C}$ are as follows:

$$\begin{aligned} F_S^i = [&\mathcal{F}_{\max}(Z^i), \mathcal{F}_{\min}(Z^i), \mathcal{F}_{mean}(Z^i), \\ &\mathcal{F}_{median}(Z^i), \mathcal{F}_{var}(Z^i)], \quad i \in \mathbb{N}. \end{aligned} \tag{12}$$

Positional Feature. As described in the introduction, the location of the same body parts in different images is unstable. Misalignment widely exists when performing local part comparison. Despite the instability, the same local part has a relatively constant position range, while different parts have different ranges of location. The positional features are used to reflect the rough locations of body parts, or, in other words, to ensure that each branch will locate the same local parts, avoiding misalignment of different parts, such as arms vs. legs or shirts vs. trousers. We use the barycenter of attention feature maps on each channel as the positional features $F_P^i \in \mathcal{R}^{2C}$:

$$F_P^i = \left[\sum_{y,x} Z_{y,x}^i * y; \sum_{y,x} Z_{y,x}^i * x\right], \quad i \in \mathbb{N}. \tag{13}$$

Relational Feature. In addition to the effectiveness of local cues of the attention mechanism, their invariance to pose for humans can enhance the distinctiveness [26]. We calculate the distance of different attention feature maps to reveal the correlation among local body parts. The relational feature of attention feature map Z^i is defined as:

$$\begin{aligned} F_R^i = [&\mathcal{D}(Z^i, Z^1), \ldots, \mathcal{D}(Z^i, Z^i - 1), \\ &\mathcal{D}(Z^i, Z^i + 1) \ldots, \mathcal{D}(Z^i, Z^N)], \end{aligned} \tag{14}$$

where $D(Z^i, Z^j) \in \mathcal{R}^C$ is the channelwise distance between Z^i and Z^j. On each channel, the distance is measured by flattening the feature map to a vector and calculating the cosine distance. The dimension of the relational feature of Z^i is thus $F_R^i \in \mathcal{R}^{(N-1)C}$.

SPR Descriptor. For each local branch, the statistical feature, the positional feature and the relational feature describe a body part from three different views. For fine-grained part representation, they are concatenated and mapped to a low-dimensional SPR descriptor via a fully connected layer. The local representation is formed by concatenating the SPR descriptors of all branches, as shown in Fig. 3.

C. *Network Training*

The concentrated attention module and the SPR descriptor are combined with a backbone network, termed the concentrated SPR network (CSPR-Net). We use ResNet-50 [17] as the backbone network for human representation. Both global and local features are embedded into a low-dimensional representation through two fully connected layers. Since global and local are two different views, we separately use two objective losses. In this work, classification loss is adopted for both the global and local views, which is denoted:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N_I} \log \frac{\exp(W_{y_i}^T x_i + b_{y_i})}{\sum_{j=1}^{N_C} \exp(W_j^T x_i + b_j)} \tag{15}$$

where i is the index of person images, x_i is the feature of the ith sample, y_i is the identity of the ith sample, W_j is the classifier for the jth identity, N_I is the number of images, and N_C is the number of identities. Together with the concentration loss, the total objective takes the following form:

$$\mathcal{L} = \mathcal{L}_{cls,global} + \mathcal{L}_{cls,local} + \lambda \mathcal{L}_r + \eta \mathcal{L}_d, \tag{16}$$

where λ and η are coefficients to control the strength of the iterative concentration process.

IV. Experiment

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the proposed method. We have tried several ways to combine the global and local features, e.g., learning a gate to select more distinctive local parts or comprehensively using the global and local neighbor relationships to get the final retrieval. However, the direct concatenation achieves the best performance. We think this is because the datasets have less occluded samples. The concatenation of all local features has made full use of all local information. Therefore, we directly concatenate the global and local features to evaluate the performance. We compare the performance of the combined features with state-of-the-art person ReID methods and investigate the contribution of each component of the proposed method, including the constrained attention module and the SPR descriptor. In addition, we evaluate the model generalization ability on other datasets.

A. *Datasets*

1) Market1501: This dataset is one of the largest benchmark datasets for person ReID. There are 32,668 detected pedestrian images in total, with 1,501 identities. These images have been captured by six cameras, including five high-resolution ones and 1 low-resolution one. For training and testing, the 1,501 identities are divided into two parts such that 751 of them are used for training and the remaining 750 are used for testing. There are 3,368 query images and 19,732 gallery images. For evaluation, we follow the same method as in [44] and report the mean average precision (mAP) and cumulative matching characteristics (CMC) in our experiment.

2) CUHK03: This dataset contains 14,097 images of 1,467 identities [45]. We follow the training/testing split proposed in [46], with 767 identities for training and 700 identities for testing. In this work, we only report the results on labeled sets.

3) DukeMTMC-ReID: This dataset is a subset of DukeMTMC for image person re-identification [47]. It provides 16,522 training images of 702 identities, 2,228 query images and 17,661 gallery images of the other 702 identities.

4) MSMT17: This dataset is a new large-scale dataset with 4,101 identities and more than 120 thousand bounding boxes [48], which have been collected from 15 cameras with both indoor and outdoor scenarios. This dataset covers a long period of time and presents complex lighting variations. A domain gap commonly exists between the training and testing sets, which makes the re-identification more challenging.

B. *Evaluation Protocol*

We follow the standard evaluation protocol, with the CMC at rank-1 and the mAP. We report the single query evaluation results on Market1501, DukeMTMC-ReID and MSMT17. The evaluation on CUHK03 is performed under the single-shot setting proposed in [46].

C. *Implementation Details*

The whole network is trained in two steps. In the first step, the backbone network is fine-tuned from the ImageNet pretrained model to obtain the global features. Then, the global and local branches are jointly learned. There are two reasons for the two-step learning. On the one hand, the objective losses for local parts and the global part are independent. They will interfere at the bottom layers, leading to unsatisfactory results. On the other hand, feature maps contain less meaningful information for the local branch at the beginning of the training procedure. The first step provides a good initialization for attention learning.

TABLE I
Dataset Description. Four Large Datasets Are Used for Supervised Learning, Including Market1501 (Market), CUHK03, DukeMTMC-reID (Duke) and MSMT17. Three Small Datasets Are Used for Cross Evaluation, Including VIPeR, CUHK01 and 3DPeS. For the 6-th Column, "H" Represents "Hand," "D" Represents "Detected" and "FR" Represents "Fast RCNN"

Dataset	Release time	Identities	Cameras	Images	Label method	Image size
Market	2015	1501	6	32217	H/D	256×128
CUHK03	2014	1467	10	13164	H/D	Varies
Duke	2017	1812	8	36441	H	Varies
MSMT17	2018	4101	15	126441	FR	Varies
VIPeR	2007	632	2	1264	H	128×48
CUHK01	2012	971	3	3884	H	160×60
3DPeS	2011	192	8	1011	H	Vary

During the training procedure, each image is resized to 384×128. For data augmentation, we apply random-sized cropping with an area ratio from 0.64 to 1.0 and an aspect ratio from $\frac{1}{3}$ to $\frac{1}{2}$ on the input image. We also use random erasing with a probability of 0.5 [49]. We use stochastic gradient descent (SGD) with a minibatch size of 32. The activation distribution constraint aims to force the candidates to concentrate on local parts of interest, while the activation ratio constraint is meant to assign large activation values to the candidates and small activation values to the noncandidates. Only when candidates are concentrated by the activation distribution constraint can the activation be assigned to local parts of interest. Therefore, the loss weights λ and η are set as 0.1 and 1.0, respectively. The number of candidates is set to between 20% and 40% of $H \times W$. In our experiment with eight branches, we use settings of two at 20%, four at 30% and two at 40%. For the first step of global branch learning, the initial learning rate is 2e-2 and shrinks to 2e-3 and 2e-4 after 50 and 100 epochs, respectively. For the second step, the learning rate of the backbone network is maintained at 2e-4, whereas that of the local branch is set at 5e-2. The learning rate shrinks to 5e-3 and 5e-4 after 50 and 100 epochs respectively. In our experiment, we train 150 epochs for each step.

D. Comparison With State-of-the-Art Methods

In this section, we compare the proposed CSPR-Net with state-of-the-art methods. The results of LOMO and BoW in Tables II, III and IV are from our implementation, as there are no results on these three datasets from the original implementation. The results in Table V are taken from [44], and the others are all taken from the original implementation. Table II shows the evaluation results on Market1501. Non-local-based methods, such as AACN [42] and DGRW [41], achieve a best CMC@1 of 92.7% and mAP of 83.0%. Local-based methods, such as MCAM [24], DuATM [25] and RPP [23], achieve a best CMC@1 of 93.8% and mAP of 81.6%. Note that the best local-based method requires an extremely high feature dimension of 12,288, while our CSPR-Net achieves the best CMC@1 of 94.2% and mAP of 84.8% with only 1,024 dimensional features. The performance further increases to 95.1% and 92.7% after reranking.

The results on CUHK03 are shown in Table III. The CMC@1 is 64.7% and 74.2% after reranking, and the mAP is 62.8% and 77.6% after reranking. The CUHK03 dataset only has two cameras for each identity. It is hard to train robust models against camera views. The results of LOMO [32], BoW [36], MLFN [40] and even some Local-based methods, such as HA-CNN [3], show limited performance. However, our method explores more accurate local cues to supplement the global features, and the results are superior to those of previous algorithms.

We further evaluate our method on two recent large-scale datasets: DukeMTMC-ReID and MSMT17. As shown in Table IV, we achieve the best CMC@1 of 83.5% and mAP of 71.9% on DukeMTMC-ReID. For MSMT17, shown in Table V, CSPR-Net significantly outperforms existing works, with a CMC@1 of 75.3% and an mAP of 50.8%. Since MSMT17 is the largest dataset with more than 120 thousands images, this result strongly demonstrates the superiority of our method. Fig. 6 exhibits the ranking results over the four different datasets, and the explicit modification further illustrates the effectiveness of the proposed method.

E. Ablation Experiments

We conduct ablation experiments to investigate the contributions of the constrained attention module and the SPR descriptor from the aspects of the local representation (L) and global+local representation (G+L) abilities, as shown in Table VI. The experiments are conducted on Market1501.

1) Investigation on the Constrained Attention Module: We argue that the effectiveness of the constrained attention module comes from both the multiscale architecture and the concentrated constraint. Since the multiscale architecture perceives most of the possible regions, the concentrated constraint is able to work as a filter to eliminate the false positive regions and retain the true positive ones. To ensure a fair comparison, we redesign a single-scale attention module (base+AM) as the baseline attention-based method, which has been widely used in previous works. We set the kernel size as 1×1, in stark contrast to the proposed architecture. Comparing the result of "base+AM" and "base+CAM," we obtain increases of approximately 2.2% in mAP and 1.3% in CMC@1 on the local representation, validating the better effectiveness of the proposed constrained attention module. However, the advantage for the global+local representation is not as significant, only 0.3% in mAP and 0.4% in CMC. The reason is that the global average pooling used for both the global and local representations cannot fully convey the extra advantage from the local parts learned by constrained attention maps. Hence, we further design another group of experiments, where we adopt the proposed SPR descriptor for fine-grained local representation. Comparing the result of "base+AM+SPR" and "base+CAM+SPR," we obtain a 3.3% increase in mAP and a 1.4% increase in CMC@1 on the local representation, while we obtain a 1.6% increase in mAP and a 0.7% increase in CMC@1 on the global+local representation. This result thoroughly verifies the previous demonstration and in turn confirms the effectiveness of the proposed constrained attention module.

In addition to the numerical results from the experiment, we visualize the learned attention maps from both the single-scale and constrained multiscale architectures. Fig. 9 visualizes

TABLE II

EXPERIMENT RESULTS ON MARKET1501. THE FIRST TWO ROWS CORRESPOND TO METHODS BASED ON HANDCRAFTED FEATURES, WHILE THE OTHER ROWS CORRESPOND TO METHODS BASED ON DEEP LEARNING MODEL. "RK" MEANS RERANKING

Category	Method	mAP(%)	CMC@1(%)	CMC@5(%)	CMC@10(%)
Handcrafted	LOMO(CVPR2015) [33]	7.8	26.1		
	BoW(ICCV2015) [36]	14.8	35.8	52.4	60.3
Non-local-based	SVDNet(ICCV2017) [37]	62.1	82.3	92.3	95.2
	PSE(CVPR2018) [38]	69.0	87.7	94.5	96.8
	AWTL(CVPR2018) [39]	75.7	89.4	-	-
	MLFN(CVPR2018) [40]	74.3	90.0	-	-
	DGRW(CVPR2018) [41]	82.5	92.7	96.9	98.1
	AACN(CVPR2018) [42]	83.0	88.7	-	-
Local-based	Spindle-Net(CVPR2017) [18]	-	76.9	91.5	94.6
	HP-Net(ICCV2017) [2]	-	76.9	91.3	94.5
	Pose-driven(ICCV2017) [20]	63.4	84.1	92.7	94.9
	DLPAN(ICCV2017) [1]	63.4	81.0	92.0	94.7
	Aligned(Arxiv2017) [22]	72.8	89.2	96.0	-
	GLAD(TMM2018) [21]	73.9	89.9	-	-
	MCAM(CVPR2018) [24]	74.3	83.8	-	-
	HA-CNN(CVPR2018) [3]	75.7	91.2	-	-
	DuATM(CVPR2018) [25]	76.6	91.4	97.1	99.0
	DaRe(R)+RE(CVPR2018) [43]	74.2	88.5	-	-
	RPP(ECCV2018) [23]	81.6	93.8	97.5	98.5
	CSPR-Net (ours)	**84.8**	**94.2**	**98 3**	**99.0**
	CSPR-Net+RK (ours)	**92.7**	**95.1**	**99.2**	**99.7**

TABLE III

EXPERIMENTAL RESULTS ON CUHK03. THE FIRST TWO ROWS CORRESPOND TO METHODS BASED ON HANDCRAFTED FEATURES, WHILE THE OTHER ROWS CORRESPOND TO METHODS BASED ON DEEP LEARNING MODEL. "RK" MEANS RERANKING

Category	Method	mAP(%)	CMC@1(%)
Handcrafted	LOMO(CVPR2015) [32]	11.5	12.8
	BoW(ICCV2015) [36]	6.4	6.4
Non-local-based	MLFN(CVPR2018) [40]	49.2	54.7
Local-based	HA-CNN(CVPR2018) [3]	41.0	44.4
	MCAM(CVPR2018) [24]	50.2	50.1
	DaRe(R)+RE(CVPR2018) [43]	60.2	64.5
	RPP(ECCV2018) [23]	57.5	63.7
	CSPR-Net (ours)	**62.8**	**64.7**
	CSPR-Net+RK (ours)	**77.6**	**74.2**

TABLE IV

EXPERIMENTAL RESULTS ON DUKEMTMC-REID. THE FIRST TWO ROWS CORRESPOND TO METHODS BASED ON HANDCRAFTED FEATURES, WHILE THE OTHER ROWS CORRESPOND TO METHODS BASED ON DEEP LEARNING MODEL. "RK" MEANS RERANKING

Category	Method	mAP(%)	CMC@1(%)
Handcrafted	LOMO(CVPR2015) [33]	17.0	30.8
	BoW(ICCV2015) [36]	12.2	25.1
Non-local-based	SVDNet(ICCV2017) [37]	56.8	76.7
	AACN(CVPR2018) [42]	59.3	76.8
	PSE(CVPR2018) [38]	62.0	79.8
	AWTL(CVPR2018) [39]	63.4	79.8
	MLFN(CVPR2018) [40]	62.8	81.0
	DGRW(CVPR2018) [41]	66.4	80.7
Local-based	HA-CNN(CVPR2018) [3]	63.8	80.5
	DuATM(CVPR2018) [25]	64.6	81.8
	RPP(ECCV2018) [23]	69.2	83.3
	CSPR-Net (ours)	**71.9**	**83.5**
	CSPR-Net+RK (ours)	**84.7**	**87.1**

TABLE V

EXPERIMENTAL RESULTS ON MSMT17. "RK" MEANS RERANKING

Method(%)	mAP	CMC@1	CMC@5	CMC@10
GoogLeNet(CVPR2015) [30]	23.0	47.6	65.0	71.8
Pose-driven(ICCV2017) [20]	29.7	58.0	73.6	79.4
GLAD(TMM2018) [21]	34.0	61.4	76.8	81.6
CSPR-Net (ours)	**50.8**	**75.3**	**86.1**	**89.6**
CSPR-Net+RK (ours)	**67.8**	**79.9**	**86.8**	**89.4**

TABLE VI

ABLATION EXPERIMENTAL RESULTS ON MARKET1501. "L" MEANS THE LOCAL FEATURE REPRESENTATION, WHILE "G+L" MEANS THE OVERALL REPRESENTATION OBTAINED THROUGH CONCATENATION OF GLOBAL AND LOCAL FEATURES

Method	L		G+L	
	mAP(%)	CMC@1(%)	mAP(%)	CMC@1(%)
base	-	-	79.7	92.0
base+AM	74.1	89.3	81.4	92.3
base+CAM	76.3	90.6	81.7	92.7
base+AM+SPR	79.2	91.8	83.2	93.5
base+CAM+S	80.6	92.3	83.3	93.4
base+CAM+P	79.6	92.2	83.2	93.4
base+CAM+R	81.4	92.7	83.8	93.6
base+CAM+SP	81.0	92.5	83.5	93.7
base+CAM+SR	81.8	93.1	84.4	93.7
base+CAM+SPR	82.5	93.2	84.8	94.2

the local parts on attention maps under the previous two conditions. We exhibit four instances for each dataset (Market1501, CUHK03, DukeMTMC-ReID and MSMT17). Within each group of four, the first three are different viewpoints of the same pedestrian, including variations in pose, scale and camera view, while the last instance is another pedestrian. In Fig. 9(a), much noise exists, and some attention maps are substantially disorganized to the extent that several separate activated locations are included. In contrast, in Fig. 9(b), the attention maps are more concentrated with little noise. It is obvious that some attention maps have specific semantics such that some focus on legs, while others focus on the upper body or other parts. The visualization results show that the learned attention maps indeed

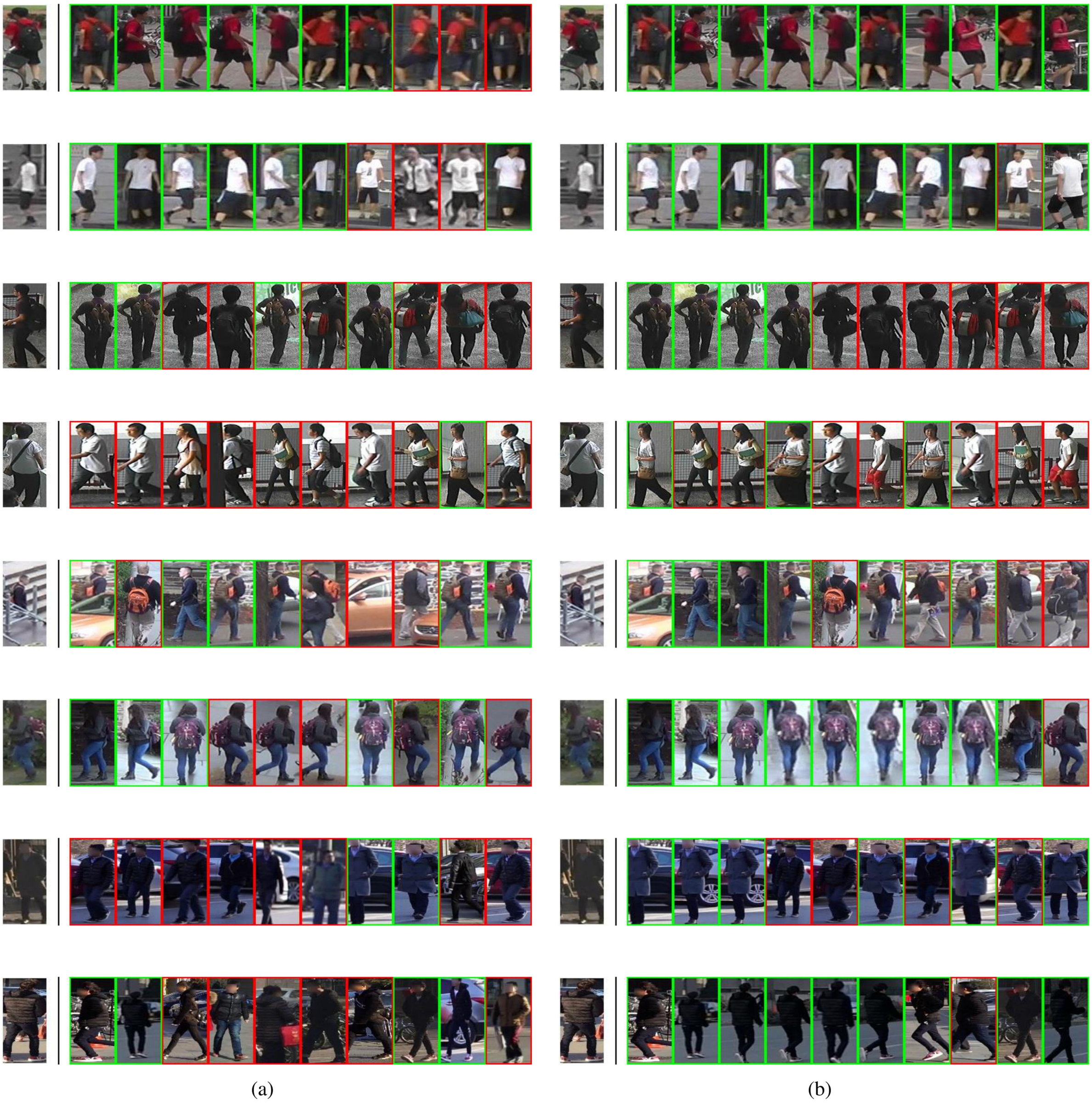
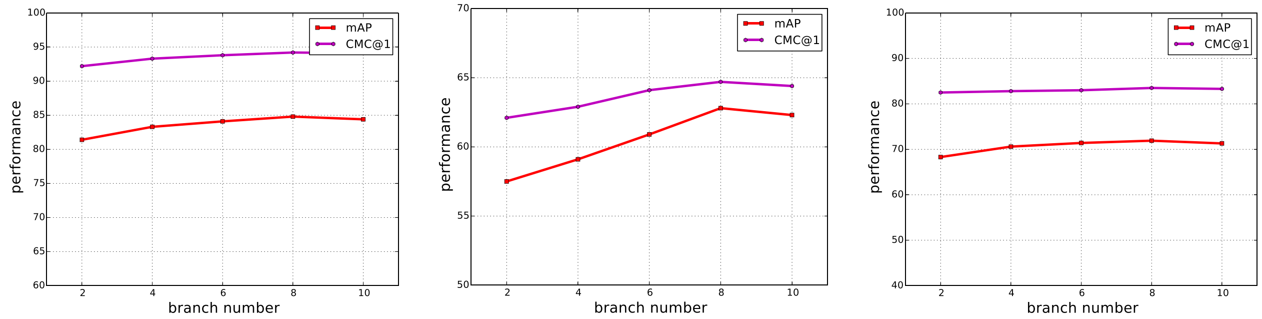


Fig. 6. Comparison of the ranking results between global features and local+global features. The green rectangles denote positive results, while the red rectangles denote negative results. In (a), global features are used to retrieve the 10 most similar images based on the query. In (b), local and global features are combined, and better retrievals are obtained. We exhibit 2 examples for each dataset (Market1501, CUHK03, DukeMTMC-reID, and MSMT17 sequentially).

have consistent semantic information, with an accurate body part localization ability.

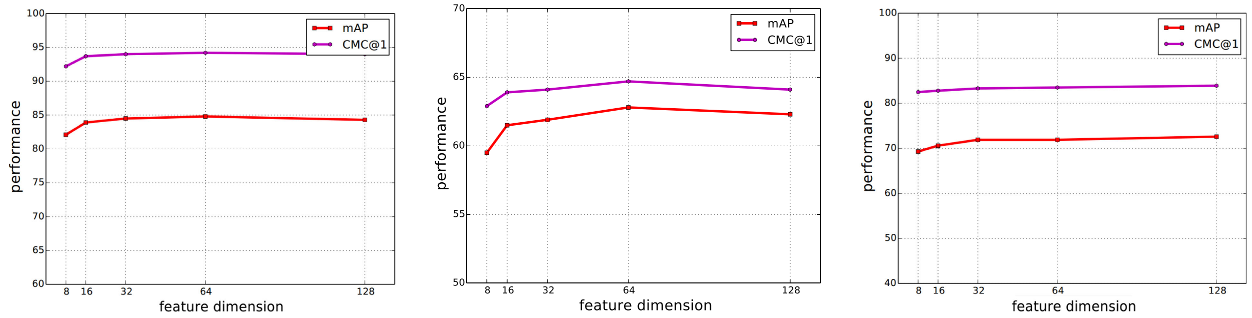
Note that we only constrain the ratio and the distribution of the activation in attention maps instead of providing explicit supervision. However, the proposed constrained attention module can still learn different and somewhat semantic local parts. This is an interesting finding that is still reasonable. First, the attention modules have the ability to learn different types of local parts, similar to how CNNs can learn distinctive kernels, due to the random initialization and stochastic gradient descent-based optimization. Second, the constraint is the necessary condition for precise local part localization. To minimize the local classification loss under the condition of concentration, the attention maps have to converge to discriminative local parts.

2) Investigation of the SPR Descriptor: In the previous section, we have demonstrated the effectiveness of the SPR descriptor from one aspect. Thus, we attempt to explore more details below. First, comparing the result of "base+AM" and "base+AM+SPR," it is apparent that the SPR descriptor contributes a substantial improvement compared with global average pooling, with approximately 5.1% mAP and 2.5% CMC@1 improvements on the local branch, as well as 1.8% mAP and 1.2% CMC@1 improvements on the global+local representation. Regarding the constrained attention module (comparing the results of "base+CAM" and "base+CAM+SPR"), 6.2% mAP and 2.6% CMC@1 improvements on the local branch can be obtained, as well as 3.1% mAP and 1.5% CMC@1 improvements on the global+local representation. The SPR descriptors



(a) Market1501 (b) CUHK03 (c) DukeMTMC

Fig. 7. Influence of the number of local branches. Each figure shows the performance curve with the change in the number of branches.



(a) Market1501 (b) CUHK03 (c) DukeMTMC

Fig. 8. Influence of the feature dimension of local branches. Each figure shows the performance curve with the change in the feature dimension.

are effective on both the single-scale and constrained multiscale attention modules, while more benefits are obtained with the constrained multiscale attention module. This result verifies that the SPR descriptor indeed captures significant local cues and is an excellent complement to the global representation.

The individual contributions of the statistical features, positional features, and relational features are also shown in Table VI. We conduct this group of experiments under the constrained attention module, as the SPR descriptor can be more effective with precisely located body parts. We directly expand the global average pooling to the statistical feature for the experiment "base+CAM+S," while we combine the positional feature and relational feature with the global average pooling for the experiments "base+CAM+P" and "base+CAM+R," respectively. Moreover, the positional feature and relational feature are subordinate descriptions, and each individual has weak distinctiveness. Comparing the results of "base+CAM+S," "base+CAM+P" and "base+CAM+R" with the result of "base+CAM," obvious improvements are obtained for all three kinds of features, especially for the relational features. The entire SPR descriptor further attains an approximately 1% improvement and achieves the best result. Among the three types of features, the statistical feature supplements the original global average pooling for a precise comparison, the positional feature ensures the consistency of learned local parts, and the relational feature measures the correlations of local parts. Moreover, the relational feature reveals something special for the local representation, which has been verified as the most valuable complement to the global part.

Note that even with the simplest attention module, our SPR descriptor has already outperformed the baseline with only local features, indicating the powerful representation ability of local parts. The superiority of base+CAM/base+CAM+SPR over base+AM/base+AM+SPR and of base+CAM+SPR/ base+AM+SPR over base+CAM/base+AM strongly demonstrates our argument that the attention mechanism should be studied from two aspects: local part discovery and local part description.

3) Investigation of the Feature Dimension: The dimension of features is an important factor that affects the performance of the local features. This dimension is determined by the number of local branches and the dimension of each branch. Therefore, we design two groups of experiments in which we vary the number of branches or the feature dimension of each local branch. For the branch number, we keep the dimension of each local branch at 64 and vary the number of branches among 2, 4, 6, 8, and 10. The results are presented in Fig. 7. It is obvious that the performance increases as the number of branches grows. The growing number of branches can capture more distinctive local parts to enrich the local cues and enhance the effectiveness of local representation. However, when the number of branches exceeds 8, the performance will slightly decrease. We surmise that this decrease is caused by the redundancy of the representation.

We also investigate the influence of the dimension of local branches. We fix the number of branches at 8 and vary the feature dimension of the local part among 8, 16, 32, 64, and 128. The results are shown in Fig. 8. For all three datasets, the performance is almost the same over the range from 16 to 128, while the dimension of 64 achieves a slightly better performance. The

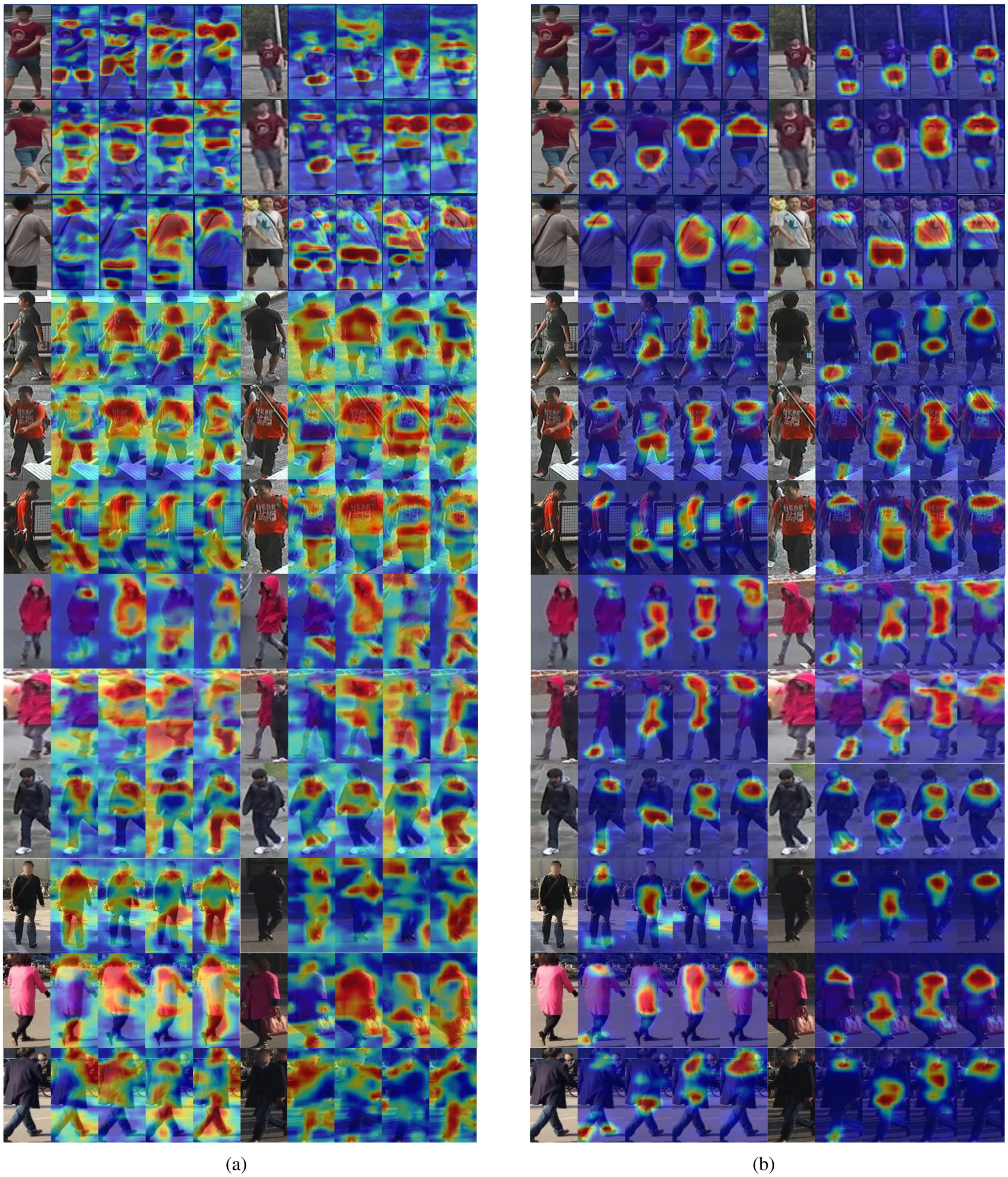


Fig. 9. Visualization results. For each dataset, we exhibit three examples and arrange them in the order of Market1501, CUHK03, DukeMTMC-reID and MSMT17. (a) and (b) Attention maps without and with the constrained attention module, respectively. In (a), the attention maps have a scattered distribution with less meaningfulness. In (b), the attention maps are more concentrated and meaningful around body parts. Moreover, in each column, local parts are well aligned among different images.

feature dimension of 8 has a relatively low performance, as an 8-dimensional feature is insufficient to load compact local cues. In real-world applications, we recommend a 32-dimensional feature of each local branch for retrieval efficiency and effectiveness.

F. Model Generalization

Unlike classical classification problems, query and gallery identities are different from those in the training stage for person ReID. Nevertheless, the domain gap between training and testing is still existed for existing datasets. In this section, we further explore the model generalization ability by training models on one dataset and evaluating them on other datasets. This experiment can reveal the real feature representation ability in real-world applications to some extent. We follow the evaluation settings in [35] and compare the CMC@1 and CMC@5 metrics. We train four models on the datasets used in this paper and evaluate them on three other datasets (Viper, CUHK01 and 3Deps).

TABLE VII

EXPERIMENTAL RESULTS FOR CROSS-DATASET EVALUATION. G MEANS THE GLOBAL FEATURE REPRESENTATION, WHILE G+L MEANS THE OVERALL REPRESENTATION OBTAINED THROUGH CONCATENATION OF GLOBAL AND LOCAL FEATURES. C-1 AND C-5 REPRESENT CMC@1 AND CMC@5, RESPECTIVELY. SRC MEANS THE SOURCE DATASETS FOR TRAINING. TAR MEANS THE TARGET DATASETS FOR TESTING

Src Tar	Market1501				CUHK03				DukeMTMC-ReID				MSMT17			
	G		G+L		G		G+L		G		G+L		G		G+L	
	C-1	C-5	C-1	C-5	C-1	C-5	C-1	C-5	C-1	C-5	C-1	C-5	C-1	C-5	C-1	C-5
Viper	17.8	40.0	21.0	43.5	18.4	38.4	20.0	40.3	17.5	36.8	20.3	41.0	17.8	32.7	20.6	40.6
CUHK01	16.6	32.5	19.5	35.6	19.4	37.0	25.1	46.1	15.8	32.7	18.5	35.9	28.2	50.2	38.8	58.5
3Deps	40.7	57.5	50.9	65.0	52.3	72.4	61.7	75.2	55.6	68.7	59.3	73.4	59.8	78.5	69.6	83.2

As shown in Table VII, the local features contribute improvements of approximately 5% in CMC@1 and 10% in CMC@5. This result again confirms the effectiveness of the method in generalization cases.

V. CONCLUSION

In this paper, we proposed exploring the attention mechanism from the two perspectives of local part discovery and fine-grained part representation for person re-identification. For local part discovery, we proposed the constrained attention module, which can learn meaningful and concentrated attention maps over body parts. For fine-grained part representation, we carefully designed a statistical-positional-relational descriptor to unlock the powerful representation ability of local parts. We conducted extensive experiments to evaluate the overall effectiveness of the proposed method, the contribution of each component, and the ability for model generalization. The visualization results demonstrated better body part learning results. The comprehensive experiments strongly verify the superiority of the proposed algorithm.

REFERENCES

[1] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3219–3228.

[2] X. Liu *et al.*, “Hydraplus-net: Attentive deep features for pedestrian analysis,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 350–359.

[3] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.

[4] L. Bazzani, M. Cristani, and V. Murino, “Symmetry-driven accumulation of local features for human characterization and re-identification,” *Comput. Vis. Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[5] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences,” in *Proc. 2nd ACM/IEEE Int. Conf. Distrib. Smart Cameras*, 2008, pp. 1–6.

[6] N. O’Hare and A. F. Smeaton, “Context-aware person identification in personal photo collections,” *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 220–228, Feb. 2009.

[7] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, “Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts,” *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 701–714, Jun. 2007.

[8] R. R. Varior, M. Haloi, and G. Wang, “Gated siamese convolutional neural network architecture for human re-identification,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.

[9] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, *arXiv:1610.02984*.

[10] M. Ye *et al.*, “Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing,” *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

[11] S. Zhou *et al.*, “Large margin learning in set-to-set similarity comparison for person reidentification,” *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 593–604, Mar. 2018.

[12] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui, “Enhancing person re-identification in a self-trained subspace,” *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 13, no. 3, pp. 1–23, 2017.

[13] J. Yu, D. Tao, J. Li, and J. Cheng, “Semantic preserving distance metric learning and applications,” *Inf. Sci.*, vol. 281, pp. 674–686, 2014.

[14] L. Wu, C. Shen, and A. V. D. Hengel, “Personnet: Person re-identification with deep convolutional neural networks,” 2016, *arXiv:1601.07255*.

[15] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[16] H. Yao *et al.*, “Deep representation learning with part loss for person re-identification,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] H. Zhao *et al.*, “Spindle net: Person re-identification with human body region guided feature decomposition and fusion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1077–1085.

[19] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 384–393.

[20] C. Su *et al.*, “Pose-driven deep convolutional model for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3960–3969.

[21] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “Glad: Global-local-alignment descriptor for scalable person re-identification,” *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.

[22] X. Zhang *et al.*, “Alignedreid: Surpassing human-level performance in person re-identification,” 2017, *arXiv:1711.08184*.

[23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[24] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Mask-guided contrastive attention model for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[25] J. Si *et al.*, “Dual attention matching network for context-aware feature sequence based person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5363–5372.

[26] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNNs for fine-grained visual recognition,” 2015, *arXiv:1504.07889*.

[27] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, “Semantic segmentation with second-order pooling,” in *Proc. Eur. Conf. Comput. Vis*, 2012, pp. 430–443.

[28] J. Yu, C. Hong, Y. Rui, and D. Tao, “Multitask autoencoder model for recovering human poses,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 5060–5068, Jun. 2018.

[29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Int. Conf. Learn. Representations*, 2015.

[30] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[31] M. Jaderberg *et al.*, “Spatial transformer networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[32] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[33] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2197–2206.

[34] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1268–1277.

[35] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 990–998.

[36] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.

[37] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3800–3808.

[38] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 420–429.

[39] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6036–6046.

[40] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2109–2118.

[41] Y. Shen *et al.*, "Deep group-shuffling random walk for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2265–2274.

[42] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2119–2128.

[43] Y. Wang *et al.*, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8042–8051.

[44] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

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

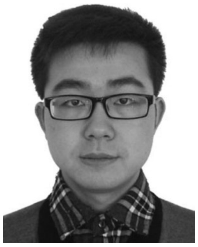
[46] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1318–1327.

[47] 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.*, 2016, pp. 17–35.

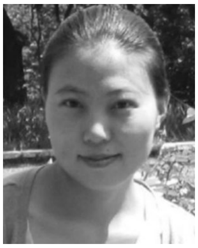
[48] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 79–88.

[49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*.

Chaoqun Wan received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2016. He is currently working toward the Ph.D. degree with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. His research interests include computer vision and machine learning.



Yue Wu received the B.E. degree in electronic engineering and Ph.D. degree in information and communication engineering from the University of Science and Technology of China, Hefei, China, in 2012 and 2017, respectively. He is a Senior Algorithm Engineer with Alibaba Group, Hangzhou, China. His research interests include multimedia, computer vision, machine learning, and data mining.



Xinmei Tian (M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. She is an Associate Professor with CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application Systems, University of Science and Technology of China. Her current research interests include multimedia information retrieval and machine learning. She was the recipient of the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation Award in 2013.

Jianqiang Huang is the Director with Alibaba DAMO Academy. His research interests focus on the visual intelligence in the city brain project of Alibaba. He was the recipient of the second prize of National Science and Technology Progress Award in 2010.

Xian-Sheng Hua (M'05–SM'14–F'16) received the B.S. and Ph.D. degrees in applied mathematics from Peking University, Beijing, China, 1996 and 2001, respectively. He is now a Distinguished Engineer/VP of Alibaba Group, leading a team working on large-scale visual intelligence on the cloud. He joined Microsoft Research Asia, Beijing, China, in 2001, as a Researcher. He was a Principal Research and Development Lead in Multimedia Search for the Microsoft search engine, Bing, Redmond, WA, USA, from 2011 to 2013. He was a Senior Researcher with Microsoft Research Redmond, Redmond, WA, USA, from 2013 to 2015. He became a Researcher and Senior Director of the Alibaba Group, Hangzhou, China, in April of 2015, leading the Visual Computing Team in Search Division, Alibaba Cloud and then DAMO Academy. He has authored or coauthored more than 200 research papers and has filed more than 90 patents. His research interests include big multimedia data search, advertising, understanding, and mining, as well as pattern recognition and machine learning. He is an ACM Distinguished Scientist. He is currently an Associate Editor for the IEEE Transactions on Multimedia and *ACM Transactions on Intelligent Systems and Technology*. He was a Program Co-Chair for the IEEE ICME 2013, ACM Multimedia 2012, and the IEEE ICME 2012. He was one of the recipients of the 2008 MIT Technology Review TR35 Young Innovator Award for his outstanding contributions on video search. He was the recipient of the Best Paper Awards at ACM Multimedia 2007, and Best Paper Award of the IEEE Transactions on CSVT in 2014. He will be serving as a General Co-Chair of ACM Multimedia 2020.