A part-based attention network for person re-identification

Weilin Zhong[1,2] · **Linfeng Jiang**[1,2] · **Tao Zhang**[1,2] · **Jinsheng Ji**[1,2] · **Huilin Xiong**[1,2]

Abstract

Person re-identification (re-id) is the task of recognizing images of the same pedestrian captured by different cameras with non-overlapping views. Person re-id is a challenging task due to the existence of large view variations, such as spatial misalignment, background clutter and human poses change. In this paper, we handle these challenges from the following two aspects: utilizing attention mechanism to alleviate misalignment problem and exploiting the complementary effects of global-local features for more stable pedestrian descriptors. Specifically, we first present a part-based attention model consisting of a channel attention block and a spatial attention block to sequentially refine the convolutional descriptors of person body parts. The channel and spatial attention blocks weight the channels and positions of body-part feature maps to spot the informative channels and regions, respectively. Then global full-body and local body-part of the refined feature maps are pooled into global and local representations, which are jointly trained using identity classification loss. We conduct extensive experiments on four standard benchmark datasets including Market1501, CUHK03, DukeMTMC-reID, and CUHK01, and the experimental results demonstrate the effectiveness of the presented method.

Keywords Person re-identification · Part-based attention model · Global-local complementary effects

1 Introduction

Person re-identification (re-id) is the task of finding the images of the same individual captured by multiple cameras distributed at different locations. Person re-id has received much attention because of its wide applications in surveillance networks and computer vision community. It is a challenging task due to the large appearance variations in viewpoints,

✉ Huilin Xiong
hlxiong@sjtu.edu.cn

1   School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

2   Institute for Sensing and Navigation, Shanghai Jiao Tong University, Shanghai, China

illumination, human poses, and occlusion. Figure 1 displays some image samples from four public benchmark datasets. Previous mainstream methods tackle the re-id problem by first extracting hand-crafted features to represent person images and then learning distance metrics for similarity calculation of the extracted features. However, the performances of these methods are affected by the limited representation power of hand-crafted features, and the separate optimization of feature representations and distance metrics. Recently, deep learning models especially Convolutional Neural Networks (CNN) have obtained outstanding performance in various computer vision tasks [11, 46, 71], such as image recognition, object detection, and face recognition. Many researchers have explored the applications of CNN in person re-id task. Different from the hand-crafted features based systems, deep methods integrate feature extraction and metric learning into one unified framework and thus achieve large performance improvement.

Existing deep methods can be categorized into two groups, namely verification models and identification models. In the first group, verification models usually contain a Siamese CNN for feature extraction of image pairs and a distance layer outputting feature similarities. A lot of training loss functions including pairwise contrastive loss [45, 49, 61], triplet loss [29, 38, 73], quadruplet loss [3], and their variants [5] are utilized to learn a feature space, where intra-class distances are minimized and inter-class distances are

(a) Market1501

(b) CUHK03

(c) DukeMTMC-reID

(d) CUHK01

Fig. 1 Samples from four datasets including Market1501, CUHK03, DukeMTMC-reID, and CUHK01. The appearances of the same person captured by different camera views can be greatly changed with different lighting and pose, background and occlusion

maximized. These methods only utilize weak annotation information about whether two images depicting the same person and thus they do not take full advantage of the pedestrian identity labels. Moreover, verification models usually take pairwise or triplet image units as input, whose number will grow exponentially with the scale growth of dataset. This will cause convergence problem in training stage, thus easily leading to under-fitting on large-scale datasets. In the second group, the identification models that require no complex data sampling process has gained popularity among person re-id community. The identification models employ person identities as the supervisory signal and the softmax loss function is adopted for model training. But identification models need a rich amount of training samples per identity to combat over-fitting. Recently, several large-scale datasets [66, 70] are collected whilst each pedestrian has sufficient training samples. For example, there are on average 17.2 training images for each identity on Market1501 [66]. Some good training tricks [42, 43, 58, 70] are also proposed to build an effective identification model. Many studies [67, 69, 74] demonstrate that identification models yield outstanding performance on large-scale datasets requiring no special data sampling scheme. Thus, our method is built on identification model.

Person re-id is a fine-grained problem and some person identities can only be distinguished by subtle differences in body parts and small visual cues (e.g., backpacks and shoes). To this end, instead of only extracting global features, many methods exploit local information to enhance the representation ability of deep features. Typically, these methods first decompose person images into several patches, parts or stripes and then conduct part-level similarity matching for more robust feature representations. They implicitly assume the availability of well-aligned person bounding box images. However, as shown in Fig. 1, cross-view images usually undergo uncontrolled spatial misalignment due to severe human pose changes and inaccurate pedestrian detection, thus degrading the performance of part-based methods. Some other methods rely on external pose localization algorithms to remedy this problem. They directly adopt the pre-trained pose estimation models (e.g., Convolutional Pose Machines [51] and OpenPose toolkit [2]) as the part detector. But they essentially perform a two-stage scheme [62, 65], where the estimated human pose parts are first manually cropped and then they are passed through another network for part-level feature extraction. Some studies [41, 59] integrate the process of pose estimation and part feature learning into one end-to-end trainable framework. However, it remains challenging to obtain ideal semantic partition of person bodies because there are large domain bias between pose estimation datasets and person re-id datasets. An alternative solution is to fully utilize the attention mechanism that can focus on discriminative local regions. But existing attention based deep learning models for re-id usually design complicated attention algorithms (e.g., Harmonious Attention Network [26]) or depend on computationally expensive units (e.g., STN [16] in [21] and attention based LSTM [37] in [29]).

In this paper, we present a Part-based Attention Model (PAM) to alleviate misalignment problem. PAM contains a channel attention block and a spatial attention block. The two blocks aim to explore the informative features of the body-part feature maps along channel and spatial dimensions, respectively. Specifically, we uniformly slice the person body into several parts and each part is forward into PAM for channel and spatial feature refinement. PAM learns to find the informative features by assigning weights to different channels and different positions. Then the global full-body and local body-part of the refined feature maps are pooled into global and part-level feature representations, each of which is trained using identity classification loss. Our model is inspired by Squeeze-and-Excitation Network (SENet) [15] and Convolutional Block Attention Module (CBAM) [53]. SENet only considers channel interdependency while ignores spatial attention. Although both channel and

spatial attention are incorporated in CBAM, the local cues beneficial for person re-id are not exploited. We thus combine part feature learning and attention mechanism for person re-id. The contributions of this paper can be summarized as follows:

1) We present a part-based attention model for person body-part feature refinement, and the visualisation results show it can somehow alleviate the spatial misalignment problem.
2) Two level of pedestrian descriptors are simultaneously learned to leverage the complementary advantages between global and local features.
3) Extensive experiments are conducted on several datasets to validate the effectiveness of the presented method.

The remainder of this paper are organized as follows: We review the related studies in Section 2. Our method is presented in Section 3 and the experimental results are shown in Section 4. Finally, the conclusion is drawn in Section 5.

2 Related work

Person re-id plays an important role in surveillance systems and thus has drawn increasing attention in recent years. Typical person re-id systems consist of two major components, namely extracting feature representations to describe the person appearance and learning distance metrics to measure the feature similarities. For feature representations, the commonly used features contain RGB, LAB, color names [60], local binary patterns (LBP) [17, 18, 22, 57], Gabor filter feature [22], color histogram and its variants [17, 18, 52] etc. For example, Li et al. [22] combined LBP, HSV color histogram, Gabor and HoG to represent person images. Liao et al. [27] constructed a feature descriptor by maximizing the horizontal occurrence of local features. Gray et al. [10] proposed to utilize AdaBoost algorithm to select the most discriminative features. Farenzena et al. [8] exploited the symmetry and asymmetry property of body structures to extract robust features for person re-id. Kviatkovsky et al. [20] utilized the color intra-distribution signatures and proposed an illumination-invariant color descriptor. For metric learning, many machine learning algorithms are utilized to learn a mapping function from the feature space to distance space, in which the intra-class distances are minimized while the inter-class distances are maximized. For instance, Zheng et al. [68] proposed a relative distance comparison (RDC) learning model from a probabilistic prospective. Davis et al. [6] proposed information-theoretic metric learning (ITML) method based on Mahalanobis distance. Liao et al. [27] proposed the Cross-view Quadratic Discriminant Analysis (XQDA), which simultaneously learned a discriminant low-dimensional subspace and a distance metric. Other representative metric learning methods contain Local Fisher Discriminant Analysis (LFDA) [32], large scale metric learning from equivalence constraint (KISSME) [18], Large Margin Nearest Neighbor (LMNN) [14] and etc. Further, Xiong et al. [57] extended many linear models (e.g., PCCA [30] and KISSME [18]) into their kernel versions. However, the performances of these methods are limited because of the separate optimization between feature extraction and metric learning.

Recently, the deep learning models that jointly learn feature representations and distance metrics have dominated the person re-id community. Numerous network architectures and training loss functions are proposed to learn more robust pedestrian descriptors. One important type of CNN model is the verification model that takes image pairs or triplets as input.

For example, Li et al. [24] proposed a patch matching layer in a filter pairing neural network (FPNN) to learn the joint representations of paired images. Yi et al. [61] proposed a Siamese convolutional neural network followed by a cosine layer to calculate pairwise similarity. Ahmed et al. [1] proposed to compute the cross-input neighborhood differences in an improved Siamese CNN model. Later, Wu et al. [54] improved the work in [1] by increasing the depth of layers and using very small convolution filters. Ding et al. [7] utilized the triplet loss to learn a view-invariant feature space. Chen et al. [3] improved triplet model and designed a novel quadruplet deep network, which was trained using quadruplet loss. Hermans et al. [13] proposed a batch-hard triplet loss which selected the hardest positive and hardest negative pairs in a training batch to form the triplet unit. Cheng et al. [5] learned both global and local features in a triplet based Siamese CNN model. Varior et al. [45] proposed to model the spatial contextual information between different parts using a long short-term memory (LSTM) architecture. Meanwhile, Varior et al. [44] also proposed a gated CNN to capture effective subtle patterns. Wang et al. [49] simultaneously learned single-image and cross-image representations in a unified triplet and Siamese deep architecture. With the scale growth of person re-id datasets, another type of CNN model, namely identification model, has obtained outstanding performance. For instance, Xiao et al. [55] learned to predict the person identities from multiple datasets in a domain-guided deep network. Zheng et al. [70] proposed to enrich the diversity of training samples using a Generative Adversarial Network (GAN). Zhong et al. [77] proposed a novel camera style transfer model based on CycleGAN [78]. Lin et al. [28] constructed an attribute-person recognition (APR) network, which learned an identity embedding and simultaneously predicted the pedestrian attributes. Sun et al. [42] proposed to decorrelate the learned weight vectors of identification CNN using singular vector decomposition (SVD). Li et al. [25] proposed a multi-loss classification model to jointly learn global and local discriminative features. Many methods also construct a hybrid model that leverage the complementary advantages of two losses. For example, Wang et al. [47] trained a multi-task attentional network using both identification and verification supervisory signals. Zheng et al. [69] combined verification and identification models in a deep network. Our method is built based on identification/classification model.

One major challenge in person re-id task is the misalignment problem as shown in Fig. 1. Many methods solve this challenge by exploiting local visual similarities on predefined rigid body parts [1, 5, 24, 45, 54]. However, rigid partition of person images still can not fully capture the body structure because of pose variations and thus the performances of their methods are limited. Other methods employ external pose estimation models for accurate body part localization. For example, Zhao et al. [62] utilized the Convolutional Pose Machines (CPM) [51] to localize head-shoulder region, upper body region and lower body region. Su et al. [41] utilized Spatial Transformer Networks (STN) [16] to localize the body parts. Sarfraz et al. [35] included the confidence maps of human poses in the model training process and proposed a new unsupervised re-ranking framework. Xu et al. [59] integrated pose estimation into the feature learning stage. These methods either rely on pre-trained pose estimation models or need manual operation, which have disadvantage in time efficiency. Moreover, there exists domain bias between re-id and pose datasets, which may bring inaccurate pose localizations. Some other methods propose attention models to focus on the discriminative regions of person images. For instance, Liu et al. [29] combined attention mechanism with LSTM for informative parts localization. Li et al. [26] proposed to combine hard regional attention [16], residual attention model [48], and channel attention in an integrated deep framework. Zhao et al. [63] proposed a fully convolutional attention

model to eliminate the misalignment problem. Our work departs from those above attention models in a simple channel-spatial attention model to refine body-part features and a multi-loss function to jointly learn global and local features. Our method is mostly similar to PCB model [43]. PCB model [43] first uniformly slices the feature maps and then uses an offline block termed Refined Part Pooling (RPP) [43] to deal with spatial misalignment problem. Compared to the work in [43], our model instead integrates an attention model for feature refinement.

3 Proposed model architecture

In this work, we design an end-to-end deep model which formulates person re-id as an identity classification problem. In this section, we will present our person re-id method. First, we will describe the deep neural network utilized in our method, then we will show the details of the part-based attention model for feature refinement. Finally, we will present the multi-loss function for global and local features learning. Figure 2 illustrates the whole framework of our method.

3.1 Convolutional representations

The deep learning models especially Convolutional Neural networks (CNN) have demonstrated significant performance improvement in a series of pattern recognition applications including person re-id task. Inspired by this, we employ CNN to extract the compact appearance features for person images. The backbone network utilized in our method is ResNet50 [12]. ResNet50 is constructed by five sequential downsampling blocks. The first block is one convolutional layer and the rest are four residual blocks encapsulated with several convolutional layers with batch normalization, ReLU, and optionally pooling operations. Each block downsamples the spatial size of feature maps into its half scale. The original design of ResNet50 is shown in Fig. 3a. Given an input image with size of 384×128, the final spatial size of ResNet50 is 12×4, which may be too small to retain the spatial
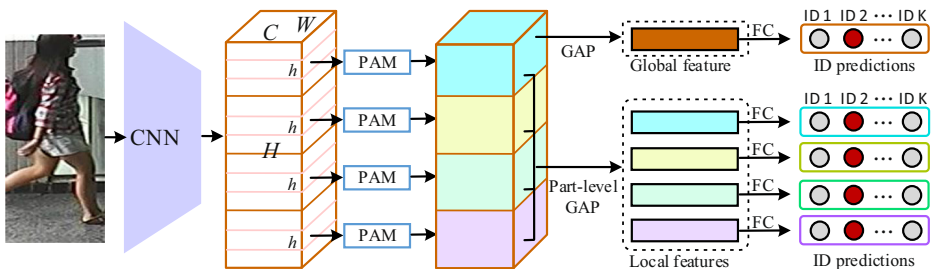

Fig. 2 Illustration of the proposed person re-identification system. We first employ a deep network to extract the base feature maps for the input image. Then, the person body is vertically sliced into several parts and each of them is passed through the presented Part-based Attention Model (PAM) for feature refinement. Two level of feature representations including global features and local features are obtained from the refined feature maps to predict person identities. GAP and FC denote global averaging pooling and fully connected layer, respectively. Here, FC layer acts as classifier. Moreover, ID is the abbreviation of identity. Take ResNet50 as example, the network output size is $2048 \times 24 \times 8$, and thus the dimensions of each pooled feature are 2048. The number of predicted IDs K is identical to the training identities on different datasets (e.g., K=750 on Market1501)

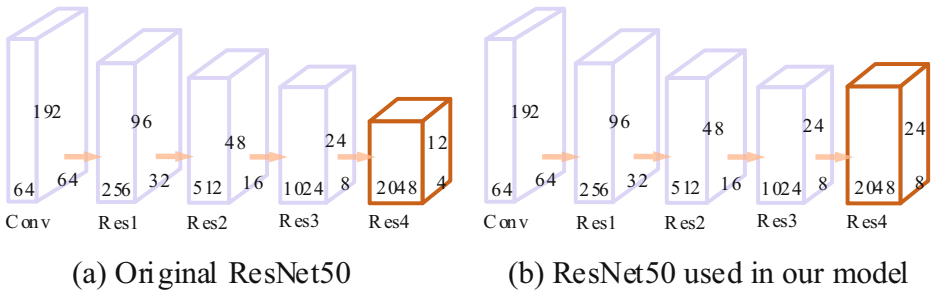(a) Original ResNet50					(b) ResNet50 used in our model

Fig. 3 Architecture of the ResNet-50 model. We use the high-level feature maps from Res4 block as the base features

regional information. We thus remove the downsampling layer in the last residual block to augment the resolution of feature maps. As illustrated in Fig. 3b, the final size of feature maps in our model is 24×8.

3.2 Part-based attention model

Person re-id system has benefited a lot from manipulation of spatial information. Specially, many methods learn local feature representations from predefined rigid body parts. Albeit simple, rigid partition of person images can roughly preserve the human body structure on vertical direction, that is, head is usually at the top part, torso and leg are at the middle and bottom part. However, the person images automatically detected by offline person detectors usually contain spatial misalignment and noisy occlusion. The features extracted from rigid body parts thus can not well describe person appearances.

In this paper, we present a part-based attention model to simultaneously take advantage of body distributions and overcome the weakness of rigid spatial decomposition. Our attention model is motivated by Squeeze-and-Excitation Network (SENet) [15] that models the interdependency between different convolutional channels, and CBAM [53] that recalibrates global feature responses. Different from those methods, our attention model aims to refine body-part feature maps for more robust pedestrian representations. As shown in Fig. 2, the output features from the deep network are denoted by $X \in R^{C \times H \times W}$, where C, H, and W represent the channel, height, and width of feature maps. We first uniformly partition the convolutional output into L vertical parts. Then, each body-part feature cube is refined by PAM. As illustrated in Fig. 4, given the input part $P \in R^{C \times h \times W}$ where $h = \frac{H}{L}$, PAM sequentially generates a 1D channel attention map $M_c \in R^{C \times 1 \times 1}$ and a 2D spatial attention map $M_s \in R^{1 \times h \times W}$ to weight the channels and positions, respectively. Afterwards, the weighted feature cube is summed with the input part using element-wise operation to obtain the final refined part. The overall attention process can be denoted as:

$$
\begin{aligned}
P_c &= M_c \otimes P, \\
P_s &= M_s \otimes P_c, \\
P' &= P \oplus P_s,
\end{aligned}
\tag{1}
$$

where $\otimes$ and $\oplus$ represents element-wise multiplication and summation, respectively. P_c and P_s denote the part feature cubes weighted by channel attention map M_c and spatial attention map M_s, respectively. P' is the final refined body-part feature. Below we will describe the computation process of each attention map.
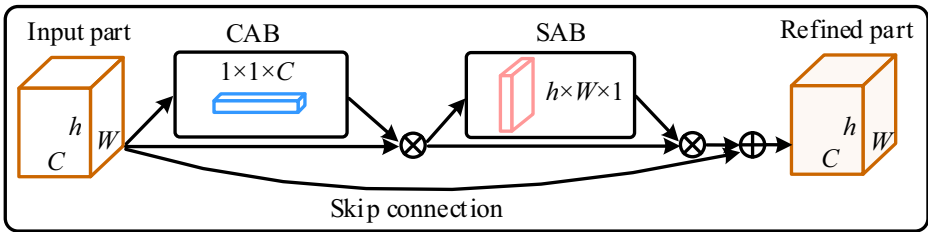
Fig. 4 Overview of the Part-based Attention Model (PAM). PAM contains a Channel Attention Block (CAB) and a Spatial Attention Block (SAB) to weight the input body-part feature cube along channel axis and spatial axis, respectively. The weighted feature cube is then element-wisely summed with the input to generate the final refined features

Channel attention block The purpose of Channel Attention Block (CAB) is to explicitly model the interdependencies between the channels of convolutional features [15]. The structure of CAB is illustrated in Fig. 5a. CAB first uses a Global Average Pooling (GAP) operation to integrate the spatial information of feature maps into a feature vector. Then the vector is forward into a multi-layer perceptron (MLP) to generate the attention map $M_c \in R^{C \times 1 \times 1}$. Specifically, we construct MLP using two fully-connected (FC) layers, whose activation outputs are in size of $R^{C/r \times 1 \times 1}$ and $R^{C \times 1 \times 1}$, respectively. Here, r is the reduction factor for the purpose of parameter reduction. In short, the channel attention can be expressed as:

$$M_c = \mathbf{W_1^{\sigma}}(\mathbf{W_0^{\text{ReLU}}}(\mathbf{GAP^s}(P))), \qquad (2)$$

where $\mathbf{GAP^s}$ means GAP operation along spatial dimension. $\mathbf{W_0} \in R^{\frac{C}{r} \times C}$ and $\mathbf{W_1} \in R^{C \times \frac{C}{r}}$ are two FC layers of MLP. Their corresponding activation functions are σ and ReLU functions, in which σ represents sigmoid function. It is worth noting that sigmoid function is used to assure that the values of attention map are in an interval of [0,1].

Spatial attention block We employ Spatial Attention Block (SAB) to automatically discover the salient regions of body-part, which is complementary to the channel attention. The structure of SAB is illustrated in Fig. 5b. SAB first aggregates the channel information into one feature map through applying GAP operation across channel axis. Then the aggregated feature map is sequentially passed through one convolutional layer to generate the spatial

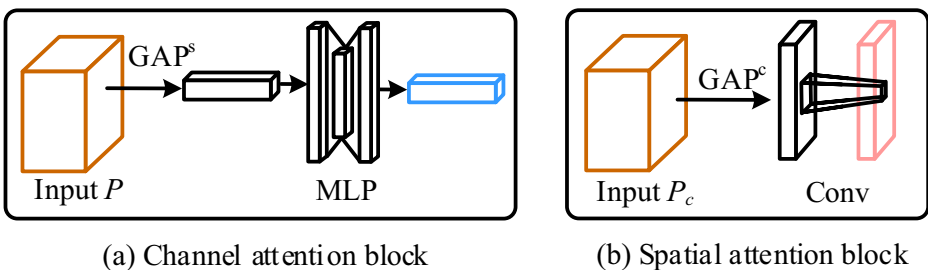(a) Channel attention block   (b) Spatial attention block

Fig. 5 Structure of Channel Attention Block (CAB) and Spatial Attention Block (SAB). GAP^s and GAP^c mean GAP operation along spatial axis and channel axis, respectively. MLP represents multi-layer perceptron

attention map $M_s \in R^{1 \times h \times W}$. The filter size and stride of convolution are 1×1 and 1, respectively. In short, the spatial attention can be denoted as:

$$M_s = \mathbf{Conv}^{\sigma}(\mathbf{GAP}^c(P_c)), \tag{3}$$

where $\mathbf{GAP}^c$ means GAP operation along channel dimension.

3.3 Multi-loss training

As shown in Fig. 2, given an input image, its corresponding output from backbone network is first decomposed into L parts and then each body-part is refined by PAM along the channel and spatial dimensions. After feature refinement, the global full-body is mapped into a set of refined parts $\{P'_1, P'_2, \ldots, P'_L\}$. To better utilize the global-local complementary cues, we extract two level of pedestrian descriptors including global features and local features. Specifically, the global features are obtained by first concatenating all the parts along vertical axis and then passing the concatenated feature maps into GAP operation. The local features are generated by pooling each refined part. The global feature and each of the part-level features are then forward into FC layer to make identity predictions.

During training phase, softmax loss is utilized to minimize the identity classification errors. In a training batch, supposing that the number of images is N and each belongs to one of K identities, the softmax loss can be written as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y^i \log \hat{p}_k^i, \tag{4}$$

where $\hat{p}_k^i$ is the probability of $i - th$ image belonging to $k - th$ identity and y^i is the ground truth identity.

The total loss function to train global and local features thus can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}^g + \sum_{i=1}^{L} \mathcal{L}_i^l \tag{5}$$

where the superscripts g and l are the abbreviations of global features and local features, respectively.

4 Experiment results

4.1 Experimental settings

Datasets The experiments are conducted on four challenging datasets, including Market1501 [66], CUHK03 [24], DukeMTMC-reID [70], and CUHK01 [23]. Market1501 contains 32,668 auto-detected bounding boxes of 1,501 identities. It is one of the largest re-id benchmarks in the existing literatures. Images of each identity are captured by at most six cameras in front of a campus supermarket with complex environment. The Deformable Part Model (DPM) [9] detector is employed as pedestrian detector and thus human parts are not well aligned in the bounding boxes. The CUKH03 dataset includes more than 13,000 images of 1360 identities collected in a university campus. Each identity is captured from two adjoint cameras and has 4.8 images on average for each view. Two versions are provided on this dataset, namely the manually labeled version and the automatically detected version by the DPM detector. We evaluate our model on the bounding boxes detected by

DPM, which is closer to the realistic setting. DukeMTMC-reID dataset is a subset of the multi-target, multi-camera pedestrian tracking dataset [33]. We use the re-id version provided by [70], which contains 34,183 images of 1,401 person identities. The pedestrian bounding boxes are manually cropped. Each person is captured by at most eight different high-resolution cameras. CUHK01 dataset contains 971 persons, which are captured by two non-overlapping camera views. This dataset includes 3,884 images and each person identity has four images.

Evaluation protocol We conduct the comparison experiments using single-query settings. Two widely used evaluation metrics are adopted for performance comparison, namely cumulative matching characteristic (CMC) [31] and mean average precision (mAP) [66]. Each dataset is split into two subsets, namely training and testing subset with non-overlapping person identities. For Market1501, we follow the standard training/testing protocol defined by [66], which uses fixed 750 identities as training subset and the rest fixed 751 identities as testing subset. For CUHK03, we adopt the new training/testing protocol proposed in [75], which fixes 767 person identities for training and the rest 700 identities for testing. For DukeMTMC-reID, following the evaluation protocol in [70], 1,401 identities are divided into a training subset with 702 identities and a testing subset with the rest 702 identities. For CUHK01, we randomly divide the 971 persons into the training subset with 485 persons and the testing subset with 486 persons.

Implementation details We implement our method based on the open source PyTorch[1] library and we train our model on a compute note with Intel Xeon CPU (64G memory) and four Titan GPUs (48G memory in total). The backbone network is pre-trained on ImageNet [34]. In all experiments, the images are first re-scaled into 420×140 and then they are cropped into 384×128. Common data augmentation techniques are applied to the input images, including mirror flip, minor rotation, and random erasing [76]. Besides, the pixels of all images are normalized to [0,1], subtracted by mean pixel values of RGB channels and then divided by standard deviation of each channel. The GAP outputs are sequentially followed by batch normalization and dropout operation, which play a critical role in avoiding over-fitting. The dropout ratio is 50%. In the training process, each batch contains 16 person identities, and for each identity we randomly select 5 images, and thus the batch-size is 80. Adam optimizer is applied for model training and we set the initial learning rate as 3×10^{-4}. The learning rate decreases every 50 epoches by a factor of 0.1 and our model convergences stably after 180 epoches.

4.2 Comparison with state-of-the-art methods

In this section, we compare the performance of our method with recent state-of-the-art approaches, including both hand-crafted features based methods and deep learning based methods. The hand-crafted features based methods include SDALF [8], eSDC [64], BoW [66], KISSME [18] and LOMO [27]. The deep methods include PersonNet [54], End-to-end CAN [29], Siamese LSTM [45], ID-discriminative Embedding (IDE) [67, 75], Gated CNN [44], Spindle Network (SpindleNet) [62], GAN [70], Pose Invariant Embedding (PIE) [65], Deeply Learned Part-aligned Representation (DLPR) [63], CNN-Embedding [69], Pose-driven Deep Convolutional model (PDC) [41], TriNet [13], Joint

---

[1]https://pytorch.org/

Learning Multi-Loss (JLML) [25], Pose-Sensitive Embedding (PSE) [35], Cam-GAN [77], TGP [58], Harmonious Attention CNN (HA-CNN) [26], Dual Attention Matching network (DuATM) [39], Part-based Convolutional Baseline (PCB) [43], SVDNet [42], Online Instance Matching (OIM) [56], Attribute-Complementary Re-id Network (ACRN) [36], Attention-Aware Compositional Network (AACN) [59], Deep Anytime Re-ID (DaRe) [50], and Mancs [47]. Note that, some methods including DaRe [50], TriNet [13], and PSE [35] obtain better results using additional re-ranking technique [75]. For fair comparison, we only compare their results without re-ranking scheme as our method.

Performance on Market1501 The comparison results are shown in Table 1. It can be seen that deep methods, especially recent state-of-the-art models including PDC [41], JLML [25], DuATM [39], and PCB [43], perform significantly better the hand-crafted

Table 1 Performance comparison on Market1501 dataset. "*" denotes unpublished paper. "-" means no available reported results

Methods	Top1	mAP
SDALF [8] (ICCV10)	20.5	8.2
eSDC [64] (CVPR13)	33.5	13.5
BoW [66] (ICCV16)	34.4	14.1
PersonNet* [54] (ArXiv16)	37.2	18.6
End-to-end CAN [29] (TIP17)	48.2	24.4
Siamese LSTM [45] (ECCV16)	61.6	35.3
IDE [67, 75] (CVPR17)	73.9	47.8
Gated CNN [44] (ECCV16)	76.0	48.5
SpindleNet [62] (CVPR17)	76.9	-
GAN [70] (ICCV17)	78.1	56.2
PIE* [65] (ArXiv17)	78.7	53.9
DLPR [63] (ICCV17)	81.0	63.4
CNN-Embedding [69]	84.0	66.1
PDC [41] (ICCV17)	84.4	63.4
TriNet* [13] (ArXiv17)	84.9	69.1
JLML [25] (IJCAI17)	85.1	65.5
AACN [59] (CVPR18)	85.9	66.9
PSE [35] (CVPR18)	87.7	69.0
Cam-GAN [77] (CVPR18)	88.1	68.7
DaRe [50] (CVPR18)	88.5	74.2
HA-CNN [26] (CVPR18)	91.2	75.7
TGP* [58] (ArXiv18)	91.7	78.8
DuATM [39] (CVPR18)	91.4	76.6
PCB [43] (ECCV18)	92.3	77.4
PCB + RE [76]	92.8	78.7
Mancs [47] (ECCV18)	93.1	82.3
PCB + RPP [43]	93.8	81.6
PCB + RE [76] + RPP [43]	94.0	82.2
Our method	93.6	81.7

features based methods (e.g., BoW [66] and eSDC [64]), illustrating the powerful feature learning capability of deep networks. Our method achieves 93.6% Top1 accuracy and 81.7% mAP on this dataset. Our model obtains significantly better performance than PersonNet [54], Siamese LSTM [45], Gated CNN [44], and GAN [70]. The Top1 accuracy of our method is 16.7%, 14.9%, 9.2%, 7.7%, and 5.9% better than the pose-driven models SpindleNet [62], PIE [65], PDC [41], AACN [59], and PSE [35], respectively. Compared to previous attention based models End-to-end CAN [29], DLPR [63], HA-CNN [26], and DuATM [39], our method improves the Top1 accuracy by 45.4%, 12.6%, 2.4%, and 2.2%, respectively, and the mAP by 57.3%, 18.3%, 6.0%, and 5.1%, respectively. Our work is close to CNN-Embedding [69], JLML [25], Mancs [47], and PCB, which are trained using multiple losses. CNN-Embedding [69] and Mancs [47] both combine verification loss and identification loss, while our method only relies on softmax loss. JLML [25] and our method learn global-local features. Compared to JLML [25], our method needs no feature sparsity constraints. Our model shares similar structure with PCB [43], which is based on uniform body-part partition. Compared to the combination of PCB [43] and Random Erasing (RE) [76], our method obtains better re-id accuracies, with 0.8% Top1 accuracy improvement (ours 93.6% versus PCB 92.8%) and 3.0% mAP improvement (ours 81.7% versus PCB 78.7%). PCB [43] deals with the part inconsistency issue caused by rigid part partition using an offline RPP module, which is not efficient. In contrast, we use end-to-end trainable part attentions for feature refinement. Besides, the performances of our model are comparable to that of combining PCB, RE [76], and RPP.

Performance on CUHK03 We conduct experiments on the detected version of CUHK03 dataset, which is a more realistic setting considering spatial displacement, partial occlusion, and pose changes. The comparison results are shown in Table 2. It can be seen that

Table 2 Performance comparison on CUHK03 detected dataset

Methods	Top1	mAP
BoW [66] + KISSME [18]	6.4	-
LOMO [27] (CVPR15)	12.8	11.5
IDE(C) [67, 75] (CVPR17)	12.8	-
IDE(C) [67, 75] + XQDA [27]	21.1	-
IDE(R) [67, 75] (CVPR17)	21.3	-
IDE(R) [67, 75] + XQDA [27]	31.1	-
SVDNet [42] (ICCV17)	41.5	37.3
HA-CNN [26] (CVPR18)	41.7	38.6
TGP* [58] (ArXiv18)	56.5	52.2
PCB [43] (ECCV18)	61.3	54.2
DaRe [50] (CVPR18)	61.6	58.1
PCB + RE [76]	61.8	56.4
PCB + RPP [43]	63.7	57.5
PCB + RE [76] + RPP [43]	64.0	58.2
Mancs [47] (ECCV18)	65.5	60.5
AACN [59] (CVPR18)	79.1	78.4
Our method	64.1	60.8

our method consistently outperforms all the hand-crafted features based methods by a large margin, including BoW [66] and LOMO [27]. Our model obtains 64.1% Top1 accuracy and 60.8% mAP, which performs better than most of the compared deep models. Specifically, our method outperforms SVDNet [42], HA-CNN [26], TGP [58], and DaRe [50] by 22.6%, 22.4%, 7.6%, 2.5% respectively at Top1, and 23.5%, 22.2%, 8.6%, and 2.7% respectively in mAP. The part-based models including PCB [43] and our method achieve good performance on this dataset, illustrating the effectiveness of local cues in learning robust pedestrian descriptors. Compared to the combination of PCB [43] and RE [76], our method improves the Top1 accuracy and mAP by 2.3% and 4.4%, respectively. The Top1 accuracy and mAP of combining PCB, RE [76], and RPP [43] can be improved to 64.0% and 58.2%, respectively. But our model still performs slightly better. On this dataset, although Mancs [47] and AACN [59] obtain high Top1 accuracy and mAP, our method that depends on no multi-task learning scheme [47] or complicated pose estimation algorithm [59] has advantage in model complexity.

Performance on DukeMTMC-reID The experimental results are shown in Table 3. It can be observed that our method performs better than most of the deep models. For instance, our method outperforms GAN [70], SVDNet [42], AACN [59], and DaRe [50] by 17.0%, 8.0%, 7.9%, and 5.6% respectively at Top1, and 22.3%, 12.6%, 10.1%, and 6.4% respectively in mAP. On this dataset, our method obtains 84.7% Top1 accuracy and 69.4% mAP, which are better than ACRN [36] that requires person attributes and Cam-GAN [77] that depends on camera information. Compared to PSE [35] that utilizes auxiliary human pose cues, our method improves the Top1 accuracy by 4.9%, and mAP by 7.4%. Our model performs better than PCB [43] and its combination with RE [76], as well as RPP. Besides, the performances

Table 3 Performance comparison on DukeMTMC-reID dataset

Methods	Top1	mAP
BoW[66] + KISSME [18]	25.1	12.2
LOMO [27] (CVPR15)	30.8	17.0
GAN [70] (ICCV17)	67.7	47.1
OIM [56] (CVPR17)	68.1	-
ACRN [36] (CVPRW17)	72.6	52.0
Cam-GAN [77] (CVPR18)	75.3	53.5
SVDNet [42] (ICCV17)	76.7	56.8
AACN [59] (CVPR18)	76.8	59.3
DaRe [50] (CVPR18)	79.1	63.0
DPFL [4] (ICCVW17)	79.2	60.6
PSE [35] (CVPR18)	79.8	62.0
PCB [43] (ECCV18)	81.8	66.1
PCB + RE [76]	82.9	68.3
PCB + RPP [43]	83.3	69.2
TGP* [58] (ArXiv18)	83.4	68.8
PCB + RE [76] + RPP [43]	83.8	70.5
Mancs [47] (ECCV18)	84.9	71.8
Our method	84.7	69.4

of our method are comparable to the best-performing method Mancs [47] on this dataset, which depends on complex hard examples mining scheme.

Performance on CUHK01 The experimental results are displayed in Table 4. It can be seen that our method performs much better than the hand-crafted features based models, including KISSME [18], eSDC [64], and KLFDA [57]. On this dataset, our method obtains 86.4% Top1 accuracy and 85.3% mAP, which outperforms many deep learning models. For example, our method improves the Top1 accuracy by 19.8%, 23.8%, 9.7%, and 6.5%, over DGDNet [55], Quadruplet [3], JLML [25], and SpindleNet [62], respectively. Compared to PCB [43] and its combination with RE [76], our method improves the Top1 accuracy by 3.3% and 2.6%, respectively, and the mAP by 3.5% and 2.6%, respectively. Compared to the combination of PCB [43], RE [76], and RPP [43], our model achieves slightly better performances on this dataset. Besidse, it is worth noting that our method needs no offline operation, which exhibits high efficiency in training stage.

4.3 Analysis of proposed model

We further make a comprehensive performance analysis to evaluate the effectiveness of each component of our presented method.

Effectiveness of attention model As shown in Fig. 3, the output spatial size of the last residual block is 24×8. We evaluate how the attention model contribute to the person re-id performance under different part partitions. We divide the feature map into 1, 2, 4, 6, 8, and 12 parts, whose spatial sizes are 24×8, 12×8, 6×8, 4×8, 3×8, and 2×8, respectively. Figure 6 displays the experimental results on Market1501. It can be observed that, under the same part partition, global or local features with PAM performs generally better than that without PAM, which validates the effectiveness of our attention model in boosting the re-id

Table 4 Performance comparison on CUHK01 dataset

Methods	Top1	mAP
SDALF [8] (ICCV10)	9.9	-
KISSME [18] (CVPR12)	13.5	-
eSDC [64] (CVPR13)	19.7	-
KLFDA [57] (ECCV14)	26.6	-
IDLA [1] (CVPR15)	47.5	-
MCP-CNN [5] (CVPR16)	53.7	-
DGDNet [55] (CVPR16)	66.6	-
Quadruplet [3] (CVPR17)	62.6	-
JLML [25] (IJCAI17)	76.7	-
SpindleNet [62] (CVPR17)	79.9	-
PCB [43] (ECCV18)	83.1	81.8
PCB + RE [76]	83.8	82.7
PCB + RPP [43]	84.7	84.5
PCB + RE [76] + RPP [43]	85.6	85.1
Our method	86.4	85.3

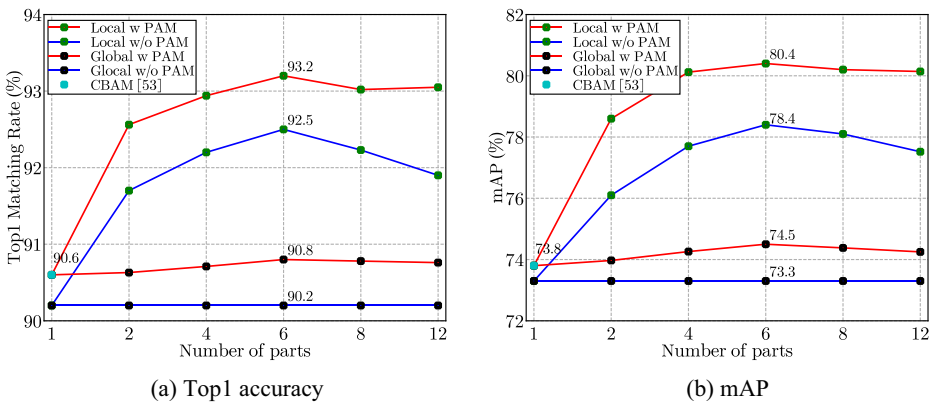(a) Top1 accuracy   (b) mAP

Fig. 6 Performance of different features under different part partitions. "w" and "w/o" represent "with" and "without", respectively

performance. Particularly, the performance of local features are largely improved with the increase of part number. When the part number $L = 1$, PAM degenerates to CBAM [53]. When the part number $L = 6$, the Top1 accuracy and mAP of local features with PAM reach their best, which are 93.2% and 80.4%, respectively. Therefore, we use this setting to conduct all the experiments. When the part number continually increases, the performance of local features with and without PAM drops with different rates. Specifically, the accuracy of local features with PAM slightly decreases while large accuracy drops can be observed when extracting local features without PAM. This implies that the performances of local features with PAM are not very sensitive to the part partitions. The local features with PAM can still obtain a relatively high Top1 accuracy even when the image is partitioned into two parts. Therefore, our method can better handle accuracy-efficiency trade-offs if applied to realistic re-id scenarios.

Performance of different attention methods In Table 5, we explore how to effectively compute the spatial attention and arrange the order of two attention modules. Note that the channel attention model is same as the SENet [15] module. It can be seen that incorporating either channel attention or spatial attention can boost the re-id performance compared to the ResNet50 baseline. The accuracies are further improved after combining two attention modules. Besides, similar performances can be observed when using different convolution kernel sizes (5×5, 3×3 or 1×1) to compute the spatial attention module. We thereby use 1×1 convolution operation in spatial attention considering the computation cost. Regarding the order of two attention modules, we evaluate three arrangements, namely sequential channel-spatial, sequential spatial-channel, and parallel use of two modules. We can observe that two sequential orders outperform parallel arrangement, possibly because channel and spatial attention models generate two different semantic embedding spaces, and simply fusing them achieves less gains. Finally, the sequential channel-spatial design is chosen in our model for its sightly better performance than the sequential spatial-channel arrangement.

Effectiveness of multi-loss training Our model is trained using multiple softmax loss. To reveal each of their ingredients contributing to the performance improvement, we report the results of baseline networks, different losses, and their combinations. The experimental results on four datasets are shown in Table 6, where the subscripts g and l represent

Table 5 Experimental results of different attention methods on Market1501 dataset

Architecture	Global feature		Local feature	
	Top1	mAP	Top1	mAP
ResNet50	90.2	73.3	92.5	79.4
ResNet50 + channel (SENet [15])	90.5	73.8	92.8	79.7
ResNet50 + spatial (5×5 conv)	90.5	73.6	92.6	79.6
ResNet50 + spatial (3×3 conv)	90.4	73.4	92.6	79.5
ResNet50 + spatial (1×1 conv)	90.4	73.5	92.7	79.5
ResNet50 + channel + spatial (5×5 conv)	90.7	74.5	93.1	80.3
ResNet50 + channel + spatial (3×3 conv)	90.7	74.3	93.3	80.3
ResNet50 + channel + spatial (1×1 conv)	90.8	74.5	93.2	80.4
ResNet50 + spatial & channel in parallel	90.7	74.0	92.9	79.8
ResNet50 + spatial + channel	90.6	74.2	93.0	80.3
ResNet50 + channel + spatial	90.8	74.5	93.2	80.4

global loss and local loss, respectively. Several important observations could be made from the results. 1) Performance improvement can be observed when augmenting the spatial size of feature maps, probably because more spatial information can be retained by using larger feature resolutions. On four datasets including Market1501, CUHK03, DukeMTMC-reID, and CUHK01, the respective accuracy improvements are 0.6%, 2.1%, 0.8%, and 1.7% at Top1, and the respective mAP improvements are 0.5%, 2.2%, 1.2%, and 0.9%. 2) It can be seen that the attention model consistently achieves better performance than baseline network. For instance, compared to $ResNet50_g$ whose spatial size is 24×8, PAM_g improves the Top1 accuracy on four datasets by 0.6%, 2.4%, 0.9%, and 2.2%, respectively, and the mAP by 1.2%, 2.3%, 1.1%, and 2.4%, respectively. 3) In general, feature embedding with local loss outperforms global loss. On four datasets, the Top1 accuracy improvements are 2.4%, 10.3%, 1.4%, and 12.1%, respectively, and the mAP improvements are 5.9%, 8.8%, 4.7%, and 13.1%, respectively. This demonstrates the benefit of incorporating local cues. 4) Combining global information and local information improves the performance over using

Table 6 Performance comparison of different losses on several datasets. The CMC Top1 accuracy (%) and mAP (%) are presented. 12×4 and 24×8 represent the spatial size of feature maps. The two subscripts g and l denote global loss and local loss, respectively

Different Losses	Market1501		CUHK03		DukeMTMC		CUHK01	
	Top1	mAP	Top1	mAP	Top1	mAP	Top1	mAP
$ResNet50_g$ (12×4)	89.6	72.8	47.7	45.1	80.5	61.2	68.8	68.3
$ResNet50_g$ (24×8)	90.2	73.3	49.8	47.3	81.3	62.4	70.5	69.2
$ResNet50+PAM_g$	90.8	74.5	52.2	49.6	82.2	63.5	72.7	71.6
$ResNet50+PAM_l$	93.2	80.4	62.5	58.4	83.6	68.2	84.8	84.7
$ResNet50+PAM_{g+l}$	93.6	81.7	64.1	60.8	84.7	69.4	86.4	85.3

Table 7 Top1 accuracy and mAP of different CNN architectures on Market1501 dataset (partition by "/")

Architecture	AlexNet	VGGNet			ResNet		
		VGG11	VGG16	VGG19	ResNet18	ResNet50	ResNet101
Layers	5	11	16	19	18	50	101
Parameters	2.45M	9.21M	14.73M	20.04M	11.17M	23.53M	42.52M
PAM_g	73.3/48.1	82.2/58.6	84.5/64.3	85.2/65.2	89.2/71.4	90.8/74.5	91.0/74.9
PAM_l	80.9/58.3	86.5/67.1	87.9/71.4	87.6/71.7	90.9/75.6	93.2/80.4	92.2/79.4
PAM_{g+l}	81.2/58.9	86.7/69.1	88.2/72.2	88.1/72.0	91.7/76.4	93.6/81.7	93.1/81.0

them individually. For instance, on Market1501, PAM_{g+l} outperforms PAM_g and PAM_l by 2.8% and 0.4%, respectively at Top1, and 7.2% and 1.3%, respectively in mAP. This shows that global and local information are complementary in nature.

Experimental results on different CNN architectures In Table 7, we conduct experiments on deep models with different parameters and layers, including two extra model architectures, namely AlexNet [19] and VGGNet [40]. Similar to ResNet50, all the FC layers and the last downsampling operation are removed. The sizes of input image are all same, which are 384×128. All the networks generate feature maps with the same resolutions, which are
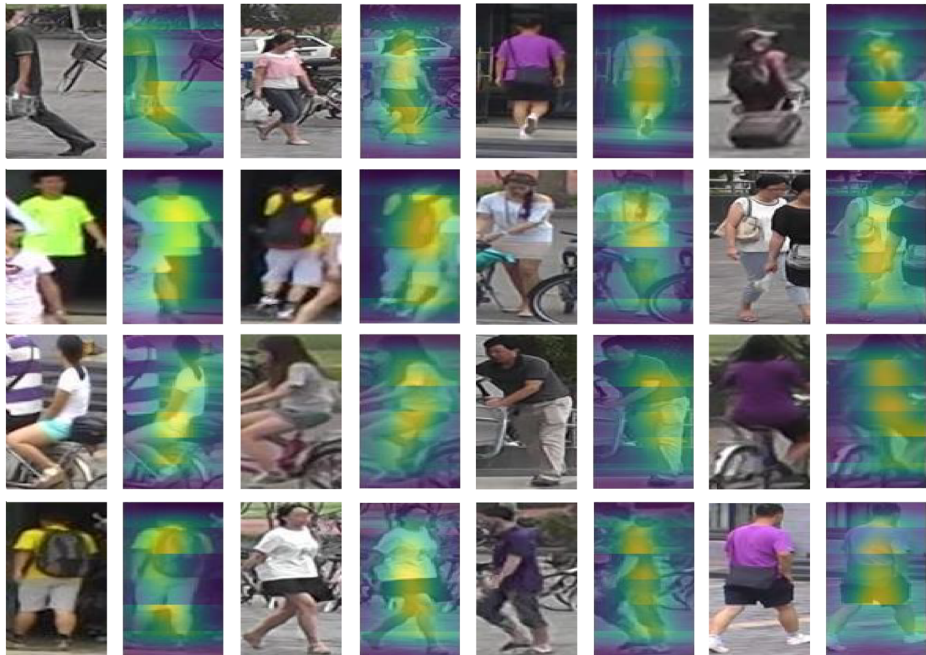

Fig. 7 Visualization of attention maps. The first three rows respectively exhibit some examples with different view variations, namely spatial misalignment, occlusion, and human pose change. The fourth row shows some examples that are well aligned

24×8. The part numbers of other models are same to that of ResNet50, which is $L = 6$. Besides, the training settings of VGGNet are same to ResNet50, while for AlexNet a slightly higher learning rate is used, which is 10^{-3}. From Table 7, we can observe that incorporating local cues and combining different losses both significantly boost the re-id performances of all CNN models, which further validates the effectiveness of the presented method in learning more robust features. Additionally, with the increases of model parameters or layers, performance improvements can be observed (for example, AlexNet versus VGG11, or VGG11 versus VGG16). But the performance improvements are not linearly correlated to model parameters or layers. The better network design also contributes to the performance improvement. For instance, compared to VGG19, ResNet18 with half parameters and nearly same layers obtains better performances. Besides, for the same model architecture, if the model parameters further increase (for example, VGG16 versus VGG19, or ResNet50 versus ResNet101), the re-id accuracies remain nearly same, possibly because of the limited dataset scale.

Qualitative results In Fig. 7, we show some attention maps produced by PAM. We can see that the model learns to assign different weights to different regions. On the attention maps, it is apparent that the human body regions are more salient than background noises, such as
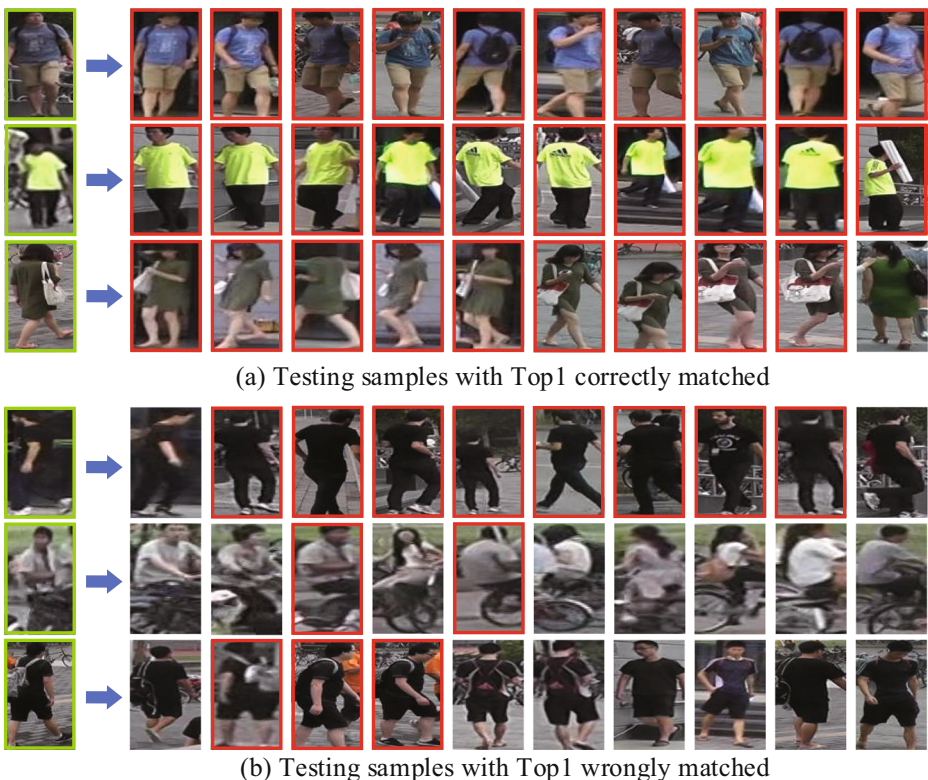

(a) Testing samples with Top1 correctly matched

(b) Testing samples with Top1 wrongly matched

Fig. 8 Samples of the retrieval results on Market1501. The images in the first column are the query images. The top 10 retrieved images are sorted according to the similarity scores from left to right. Red rectangles represent the correct matches

Table 8 Comparison of average feature extraction time and average retrieval time on Market1501 (millisecond per image)

Methods	Feature dimension	Feature extraction	Retrieval	Total
IDE(R) [75]	2048	5.8	6.33	12.13
CNN-Embedding [69]	2048	5.9	6.42	12.32
PCB [43]	12288	5.9	6.71	12.61
IDLA [1]	500	410.6	6.87	417.47
RCN [72]	4800	93.5	6.71	100.21
Our method	14336	5.9	6.93	12.83

cars and trees. The top and bottom regions are less salient than the middle body parts, probably because the faces with low resolutions and the small legs contain less discriminative information than body torsos, which include most of the clothing cues. It can be seen that the person appendixes, such as backpacks and luggage, are partially attended, which means they can assist matching persons. Our model can still focus on the body parts even under large view variations including spatial misalignment, occlusion, and human pose change. For instance, in the second row, although the target persons are occluded by other persons or objects such as bicycles, PAM pays more attention on the human body regions of target persons.

In Fig. 8, we show top 10 retrieval results from Market1501. It can be observed that our model exhibits strong robustness to pose changes, scale variations, and spatial displacement. We can see that the false matchings are mainly caused by similar looking in visual appearances. These failure cases are also very challenging from human perspective, especially the person images with extremely similar clothes and human poses in the first two rows of Fig. 8b.

Running time analysis In Table 8, we compare the average feature extraction time per image and the average retrieval time per image of our method with five other methods, including IDE(R) [75], CNN-Embedding [69], PCB [43], RCN [72], IDLA [1]. For fair comparison, we re-implement their feature extraction codes using PyTorch. The experiments are conducted on a machine with Intel Xeon CPU (64G memory) and four Titan GPUs (48G memory in total). During test phase, a batch is composed of 2,000 images. It can be seen that the identification models including IDE(R) [75], CNN-Embedding [69], PCB [43], and our method exhibit high computation efficiency compared to the verification models IDLA [1] and RCN [72]. The identification models use feature vectors that can be saved in buffer for distance calculation, and they only need to forward all the images once. But for verification models, they have to forward the same image for several times to obtain the joint feature of image pair, which is a time-consuming process.

5 Conclusion

In this paper, we present a part-based attention network with multi-loss training for the task of person re-id. Specially, the part-based attention model contains a channel attention block and a spatial attention block to refine the feature maps of person body parts along channel and spatial dimensions. The attention model is capable of alleviating the spatial misalignment problem. Besides, to fully exploit the complementary benefits of global-local cues,

two level of pedestrian descriptors including global full-body and local body-part features are extracted from the refined feature maps, each of which is trained using identification loss. We conduct extensive experiments on four public person re-id benchmarks including Market-1501, CUHK03, DukeMTMC-reID, and CUHK01. The experimental results demonstrate that our method yields higher re-id accuracy than most of state-of-the-art approaches.

References

1. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 3908–3916
2. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE conference on computer vision and pattern recognition, pp 1302–1310
3. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 403–412
4. Chen Y, Zhu X, Gong S (2018) Person re-identification by deep learning multi-scale representations. In: IEEE international conference on computer vision workshop, pp 2590–2600
5. Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: IEEE conference on computer vision and pattern recognition, pp 1335–1344
6. Davis JV, Kulis B, Jain P, Sra S, Dhillon IS (2007) Information-theoretic metric learning. In: International conference on machine learning, pp 209–216
7. Ding S, Lin L, Wang G, Chao H (2015) Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn 48(10):2993–3003
8. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: IEEE conference on computer vision and pattern recognition, pp 2360–2367
9. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
10. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision, pp 262–275
11. Guo Y, Yu L, Oerlemans A, Lao S, Song W, Lew MS (2016) Deep learning for visual understanding: a review. Neurocomputing 187(26):27–48
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
13. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
14. Hirzer M, Roth PM, Bischof H (2012) Person re-identification by efficient impostor-based metric learning. In: International conference on advanced video and signal-based surveillance, pp 203–208
15. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition, pp 7132–7141
16. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. In: Neural information processing systems, pp 2017–2025
17. Khamis S, Kuo CH, Singh VK, Shet VD, Davis LS (2014) Joint learning for attribute-consistent person re-identification. In: European conference on computer vision, pp 134–146
18. Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: IEEE conference on computer vision and pattern recognition, pp 2288–2295
19. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Neural information processing systems, pp 1097–1105
20. Kviatkovsky I, Adam A, Rivlin E (2013) Color invariants for person reidentification. IEEE Trans Pattern Anal Mach Intell 35(7):1622–34

21. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 384–393
22. Li W, Wang X (2013) Locally aligned feature transforms across views. In: IEEE conference on computer vision and pattern recognition, pp 3594–3601
23. Li W, Zhao R, Wang X (2012) Human reidentification with transferred metric learning. In: Asian conference on computer vision, pp 31–44
24. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 152–159
25. Li W, Zhu X, Gong S (2017) Person re-identification by deep joint learning of multi-loss classification. In: International joint conference on artificial intelligence, pp 2194–2200
26. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 2285–2294
27. Liao S, Hu Y, Zhu X, Li S (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE conference on computer vision and pattern recognition, pp 2197–2206
28. Lin Y, Zheng L, Zheng Z, Wu Y, Yang Y (2017) Improving person re-identification by attribute and identity learning. arXiv:1703.07220
29. Liu H, Feng J, Qi M, Jiang J, Yan S (2017) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506
30. Mignon A, Jurie F (2012) Pcca: a new approach for distance learning from sparse pairwise constraints. In: IEEE conference on computer vision and pattern recognition, pp 2666–2672
31. Moon H, Phillips PJ (2001) Computational and performance aspects of pca-based face-recognition algorithms. Perception 30(3):303–21
32. Pedagadi S, Orwell J, Velastin S, Boghossian B (2013) Local fisher discriminant analysis for pedestrian re-identification. In: IEEE conference on computer vision and pattern recognition, pp 3318–3325
33. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, pp 17–35
34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2014) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
35. Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: IEEE international conference on computer vision, pp 420–429
36. Schumann A, Stiefelhagen R (2017) Person re-identification by deep learning attribute-complementary information. In: IEEE conference on computer vision and pattern recognition workshops, pp 1435–1443
37. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv:1511.04119
38. Shi H, Yang Y, Zhu X, Liao S, Lei Z, Zheng W, Li S (2016) Embedding deep metric for person re-identification: a study against large variations. In: European conference on computer vision, pp 732–748
39. Si J, Zhang H, Li C, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 5363–5372
40. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations
41. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: IEEE international conference on computer vision, pp 3960–3969
42. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: IEEE international conference on computer vision, pp 3820–3828
43. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: European conference on computer vision, pp 501–518
44. Varior RR, Haloi M, Wang G (2016) Gated siamese convolutional neural network architecture for human re-identification. In: European conference on computer vision, pp 791–808
45. Varior RR, Shuai B, Lu J, Xu D, Wang G (2016) A siamese long short-term memory architecture for human re-identification. In: European conference on computer vision, pp 135–153
46. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018:1–13
47. Wang C, Zhang Q, Huang C, Liu W, Wang X (2018) Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: European conference on computer vision, pp 365–381
48. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: IEEE conference on computer vision and pattern recognition, pp 6450–6458

49. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 1288–1296
50. Wang Y, Wang L, You Y, Zou X, Chen V, Li S, Huang G, Hariharan B, Weinberger KQ (2018) Resource aware person re-identification across multiple resolutions. In: IEEE conference on computer vision and pattern recognition, pp 8042–8051
51. Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: IEEE conference on computer vision and pattern recognition, pp 4724–4732
52. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10(1):207–244
53. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: convolutional block attention module. In: European conference on computer vision, pp 3–19
54. Wu L, Shen C, Hengel AVD (2016) Personnet: person re-identification with deep convolutional neural networks. arXiv:1601.07255
55. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 1249–1258
56. Xiao T, Li S, Wang B, Lin L, Wang X (2017) Joint detection and identification feature learning for person search. In: IEEE conference on computer vision and pattern recognition, pp 3376–3385
57. Xiong F, Gou M, Camps O, Sznaier M (2014) Person re-identification using kernel-based metric learning methods. In: European conference on computer vision, pp 1–16
58. Xiong F, Xiao Y, Cao Z, Gong K, Fang Z, Zhou JT (2018) Towards good practices on building effective cnn baseline model for person re-identification. arXiv:1807.11042
59. Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 2119–2128
60. Yang Y, Yang J, Yan J, Liao S, Yi D, Li S (2014) Salient color names for person re-identification. In: European conference on computer vision, pp 536–551
61. Yi D, Lei Z, Liao S, Li S (2014) Deep metric learning for person re-identification. In: International conference on pattern recognition, pp 34–39
62. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017) Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: IEEE conference on computer vision and pattern recognition, pp 1077–1085
63. Zhao L, Li X, Wang J, Zhuang Y (2017) Deeply-learned part-aligned representations for person re-identification. In: IEEE international conference on computer vision, pp 3219–3228
64. Zhao R, Ouyang W, Wang X (2013) Unsupervised salience learning for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 3586–3593
65. Zheng L, Huang Y, Lu H, Yang Y (2017) Pose invariant embedding for deep person re-identification. arXiv:1701.07732
66. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2016) Scalable person re-identification: a benchmark. In: IEEE international conference on computer vision, pp 1116–1124
67. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv:1610.02984
68. Zheng WS, Gong S, Xiang T (2013) Reidentification by relative distance comparison. IEEE Trans Pattern Anal Mach Intell 35(3):653–668
69. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person re-identification. ACM Trans Multimed Comput Commun Appl 14(1)
70. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE conference on computer vision and pattern recognition, pp 3774–3782
71. Zhong W, Jiang L, Zhang T, Ji J, Xiong H (2018) A multi-part convolutional attention network for fine-grained image recognition. In: International conference on pattern recognition, pp 1857–1862
72. Zhong W, Jiang L, Zhang T, Ji J, Xiong H (2019) Combining multilevel feature extraction and multi-loss learning for person re-identification. Neurocomputing 334(21):68–78
73. Zhong W, Xiong H, Yang Z, Zhang T (2017) Bi-directional long short-term memory architecture for person re-identification with modified triplet embedding. In: IEEE international conference on image processing, pp 1562–1566
74. Zhong W, Zhang T, Jiang L, Ji J, Zhang Z, Xiong H (2019) Discriminative representation learning for person re-identification via multi-loss training. J Vis Commun Image Represent 62:267–278
75. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: IEEE conference on computer vision and pattern recognition, pp 1318–1327

76. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017) Random erasing data augmentation. arXiv:1708.04896
77. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018) Camera style adaptation for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 5157–5166
78. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision, pp 2242–2251

Publisher's note   Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Weilin Zhong received the B.S. degree from Northeastern University, China, in 2015. He is currently pursuing a Ph.D. degree in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China. His research interests are in computer vision, image processing, person re-identification, and fine-grained image recognition.

Linfeng Jiang received the B.S. degree in Computer Science from Chongqing University, China, in 2005, and the M.S. degree in Computer Science from Kunming University of Science & Technology, China, in 2011. He is currently working towards the Ph.D. degree in Department of Automation at Shanghai Jiao Tong University, China. He is interested in computer vision and probabilistic graphical theory for context modeling.

Tao Zhang received the B.S. degree in electronic information engineering from Huainan Normal University in 2011, and the M.S. degree in communication and information system from Sichuan University in 2014. Currently, he is working toward the Ph.D degree in Shanghai Jiao Tong University (SJTU), Shanghai, China. His research work focuses on PolSAR image processing and machine learning.
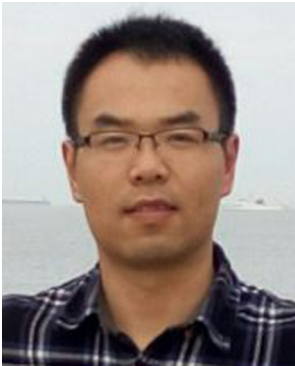
Jinsheng Ji received the B.S. degree in automation from Nanjing Agricultural University in and the M.S. degree in control science and engineering from Shanghai Jiao Tong University, China. He is currently pursuing the Ph.D. degree in Department of Automation at Shanghai Jiao Tong University, China. His research interests are computer version and machine learning.

Huilin Xiong received the B.Sc and M.Sc. degrees in Mathematics from Wuhan University, Wuhan, China, in 1986 and 1989, respectively. He received his Ph.D. degree in Pattern Recognition and Intelligent Control from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1999. He joined Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2007, and currently, he is a professor at Department of Automation of SJTU. His research interests include pattern recognition, machine learning, and bioinfomatics.