Applied Intelligence (2021) 51:1479–1491
https://doi.org/10.1007/s10489-020-01880-4

A divide-and-unite deep network for person re-identification

Rui Li[1] · Baopeng Zhang[1] · Zhu Teng[1] · Jianping Fan[2]

Published online: 28 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Person re-identification (person re-ID) is one of the most challenging tasks in the field of computer vision as it involves large variations in human appearances, human poses, background illuminations, camera views, etc. In recent literature, using part-level features for the person re-ID task provides fine-grained information, and has been proven to be effective. Instead of relying on additional skeleton key points or pose estimation models, this paper proposes a Divide-and-Unite Network to obtain feature embedding end-to-end. We design a deep network guided by image contents, which divides pedestrians into parts and obtains the part features with different contributions. These part features and the global feature are united to obtain the pedestrian descriptor for person re-ID. To summarize, the contributions of this work are two-fold. Firstly, a novel architecture of discriminative descriptor learning is proposed, which is based on the global feature and supplemented by part features. Secondly, a Feature Division Network is constructed to generate the part features with different contributions, where the divided parts maintain the consistency of content between different images. Extensive experiments are conducted on three widely-used benchmarks including Market1501, CUHK03, and DukeMTMC-reID. The results have demonstrated that the proposed model can achieve remarkable performance against numerous state-of-the-arts.

Keywords Person re-identification · Siamese network · Global feature · Part feature

1 Introduction

Due to the wide applications of person re-ID in video surveillance and intelligent security, it is an attractive task in the field of computer vision. Given a query, the purpose of person re-ID is to find the specific person from other camera views and these views are captured by the non-overlapping surveillance deployed at different locations. It is a very challenging problem [2] due to diverse intrinsic factors and extrinsic impacts such as appearance variations, human pose changes, illumination disparity, camera views, information loss in data collection, occlusions, etc.

With the development of deep learning, more and more researchers use the Convolutional Neural Network (CNN) for the person re-ID task, and these methods show great development potential. In the re-ID task, learning an effective feature representation is indispensable and crucial. Earlier researches extract global features on pedestrian images, and some works use different learning tasks [51] or attention mechanism [18] to get better performance. Recently, increasingly state-of-the-art methods emphasize the body part and propose various partition strategies. Many deep models encoding local information have been developed [29, 34, 38]. However, most of these methods rely on external cues such as additional skeleton key points or pose estimation models, to prevent arbitrary misalignment and background clutters. This implicitly requires large amounts of labeled training data. Besides, the latent dataset bias between pose estimation and person re-ID is hard to diminish for the semantic partition. To avoid the stumbling block caused by the dataset bias, some works [30, 45, 49] propose to learn the body part not requiring any

✉ Zhu Teng
zteng@bjtu.edu.cn

Rui Li
rui.li@bjtu.edu.cn

Baopeng Zhang
bpzhang@bjtu.edu.cn

Jianping Fan
fanj1@lenovo.com

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

[2] AI Lab, Lenovo Research, Beijing, China

additional models nor body part labeling information, and these methods show excellent performance.

There are still some problems with the above-mentioned existing methods which employ local features for re-ID. A stimulus example is that if two different pedestrians are similar in a glance, we usually need to identify the two people as different individuals through more details, such as hair and shoes. That is to say, when recognizing the identity of images, some local information (such as the hair, glasses, clothes, bags, etc.) might be more critical than other part features. Moreover, the location of the important local information varies from person to person, and the part features with the same contributions in existing methods might ignore some key information at times. Another problem is that the acquisition of part features is easily affected by the background clutter and intra-domain bias. If a uniform partition is directly employed, the part content can not be consistently maintained (see Person1 and Person2 in Fig. 1 as an example). Therefore, to extract more discriminative features for person re-ID, contributions of local clues (such as hair, bags, and so on) should not be the same for different persons. In other words, discerning clues should contribute differently in the fine recognition task of a specific identity, and the simple uniform partition should not be directly applied to the re-ID task to extract the part features of pedestrians. Motivated by this, we construct a Feature Division Network to generate the part features with different contributions, in which the divided parts preserve the consistency of content between different images through a feature transformation operation.

Unlike the existing methods which directly utilize global features, part features, or a simple combination of these features for person re-ID, we analyze the characteristics of global and local features and propose a novel architecture Divide-and-Unite Network (DUNet) to learn the discriminative descriptor. The proposed model uses CNN with Siamese structure and it is different from Siamese networks

Fig. 1 Examples of content inconsistency caused by the uniform partition. For Person1, the first part of image1 obtains the head information while the first part of image2 contains only background information

such as [43, 51],and [6] that only extract global features. In our architecture, the global feature of the pedestrian descriptor is used to learn the similarity measure between different images, and part features are utilized as the auxiliary information to enhance the discrimination of descriptors. It is worth noting that our model reduces the influence of spatial diversity on the uniform partition, and adaptively scores the importance of each partition to obtain part features with different contributions by the proposed Feature Division Network (FDNet).

To summarize, the main contributions of the paper are listed as follows.

1) We propose a novel architecture, Divide-and-Unite Network (DUNet), which extracts the global feature to discover the similarity between different images and is supplemented by part features to enhance the discrimination of descriptor.
2) We propose a Feature Division Network (FDNet) that performs feature transformation and local feature extraction. It can reduce the impacts of spatial diversity on the uniform partition and deduce the contribution of each part to achieve flexible part features.
3) Extensive experiments on three popular datasets (Market1501, CUHK03 and DukeMTMC-reID) are executed and the results show that our algorithm significantly outperforms numerous state-of-the-arts.

2 Related work

Traditional algorithms of person re-identification focus on designing feature representations [13, 15, 25, 50] and learning an effective distance metric such as LMNN [39], KISSME [11], XQDA [19] and MVLDML+ [44]. Recently, the success of deep learning in image identification inspires progressive studies on person re-ID. These methods can be divided into two groups according to their emphases. The first group [8, 21, 24, 29, 30, 34, 38, 45, 48, 49, 52] focuses on designing feature representations, where methods propose differently structured models to extract features from the original images automatically. The other group [1, 10, 22, 27, 41] pays more attention to the design of distance metrics, which aggregates the same class of samples and separates different ones as much as possible.

To learn more effective feature representations, the auxiliary information is employed at times. For example, the deep architecture proposed by C.Su et al. [29] employs the human part cues to alleviate the pose variations and learn robust features from both the global image and different local parts. A novel framework that consists of Pose guided Part Attention (PPA) and Attention-aware Feature Composition (AFC) is introduced [42] to learn a generalized

feature representation. The key points of the human body are utilized in [48] to extract the regions of interest, and then it fuses the global feature and multi-scale local features as the final pedestrian descriptor. In [38], a Global-Local-Alignment Descriptor is proposed to detect several part regions first and learn the local and global regions using a designed deep neural network. These methods rely on additional skeleton key points or pose estimation models to prevent arbitrary misalignment and background clutters. Recently, many deep learning models embedding local information without extra models have been developed [30, 45, 49] and achieve a competitive accuracy. [45] designs Part Loss Networks (PL-Net). The loss of PL-Net makes the network concentrate on the key parts and automatically detect body parts. L.Zhao et al. [49] propose a simple and effective human part-aligned representation to handle the body part misalignment problem. These methods only extract the part features of persons and neglect the contribution of each part to the content information of the pedestrian features in the re-ID task. In recent years, methods based on attention mechanisms have been widely used in computer vision [6, 7, 43], and some researchers have begun to employ this mechanism in person re-ID to obtain effective feature representations. For instance, [18] proposes an HA-CNN model to simultaneously learn hard region-level and soft pixel-level attention along with the simultaneous optimization of feature representations. This achieves satisfactory performance in the person re-ID task. Different from these methods on representation learning, we propose a divide-and-unite model for a joint feature representation, which aggregates the global feature and part features to enhance the discrimination of descriptor.

Metric learning is a widely used method for the person re-ID task whose purpose is to learn the similarity of two inputs through the deep neural network. Among the methods for metric learning, the Siamese network and triplet network are two typical structures. They usually employ contrastive loss [33], triplet loss [22, 27] and TriHard loss [10] as the loss functions to train the parameters of the network. For example, Yi et al. [46] utilize a Siamese convolutional neural network with an identification loss to train a feature representation. In recent years, to improve the accuracy of re-ID, some researchers combine the metric learning with representation learning to acquire a re-ID network. They usually utilize the identification task loss to identify the input person (IDs from datasets) or use the verification task loss to determine whether the two inputs are the same person (the IDs are the same or not). Yi et al. [46] construct a Siamese convolutional neural network and jointly learn the color feature, texture feature, and metric in the unified framework to learn a feature representation. To build a more comprehensive person re-ID algorithm, Z. Zheng et al. [51] combine the verification task and identification task and utilize the cross-entropy function as the loss of these two tasks in a Siamese network. In our work, we employ the verification task to learn the similarity of different inputs in the feature space, and the identification task to capture the discriminative pedestrian descriptor of each input.

3 The proposed method

In this section, we first describe the pipeline of the proposed method in Section 3.1, and the key components are elaborated in Sections 3.2 and 3.3.

3.1 The overall framework

To retrieve a given person in gallery images, a novel deep architecture named DUNet, is proposed. Figure 2 presents the pipeline of the proposed DUNet. In contrast to other global and part deep models for person re-ID, our global feature learns the similarity measure between different images by the constraints of identification and verification tasks, and our part features are constrained by the identification task to enhance the discrimination of the descriptor in the training process.

As shown in Fig. 2, it can be observed that the DUNet model consists of two Base Networks with shared parameters, a Dual Attention Siamese Network (DASNet), and a Feature Division Network (FDNet). In this model, the global and part features are considered simultaneously during each round of training. In the Base Network, the first two blocks of ResNet50 [9] are employed, and the rest blocks are engaged in the feature extractor of FDNet and DASNet. Given a pair of training samples, the Base Network extracts low-level features, and then these features are delivered to FDNet and DASNet respectively. On the one hand, these features are fed into the feature transformation of FDNet to learn part-level features and keep the consistency of each input before partition. Then the horizontal uniform partition is conducted to divide the input into parts with content consistency and different contributions. On the other hand, the low-level features of two inputs are delivered to DASNet to learn the global features of different inputs in the same feature space. The discriminative contents of inputs are emphasized and the less useful ones are suppressed by dual attention mechanism in DASNet. Finally, to gain the trained DUNet, we forward the global and part feature to different tasks to calculate loss and adjust parameters. In the test process, an image is fed into FDNet and one branch of DASNet to obtain the part and global features. As described in (1), the pedestrian descriptor is generated by concatenating the global and part

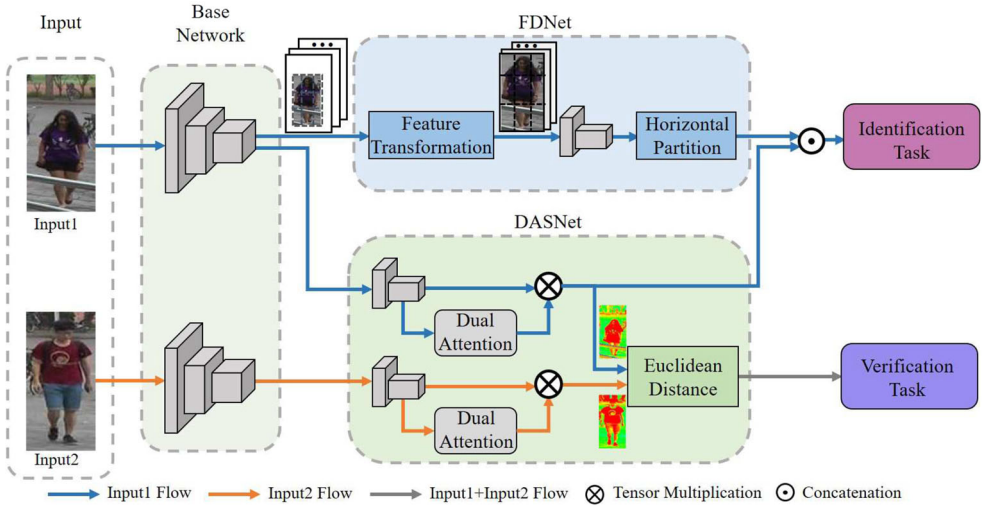


Fig. 2 The pipeline of DUNet. The DUNet consists of a Feature Division Network (FDNet) and a Dual Attention Siamese network (DASNet). The former, which contains a feature transformation branch, a deep feature representation branch, and a horizontal partition branch, is trained by the identification task. It divides an input person into parts with content consistency and extracts the part features with different contributions. The latter includes two subnetworks sharing the same structure and parameters. The part features obtained by FDNet and the global features of the same input extracted by DASNet are concatenated and sent to the identification task to determine the identity of the input image. The verification task determines whether the two input images are the same person according the global features extracted by DASNet

features, where the F_{part} and F_{global} are the part and global feature vectors. The dimensions of global and local features are respectively set to 2,048 and 512 in all experiments.

$$F_{feature} = F_{part} \odot F_{global} \tag{1}$$

3.2 Feature division network

To produce effective and discriminative part-level features, the FDNet is developed. It is composed of a feature transformation branch, a deep feature representation branch, and a horizontal partition branch as described in Fig. 3. The feature transformation branch and feature division branch are circled in the red and blue dotted boxes, respectively.

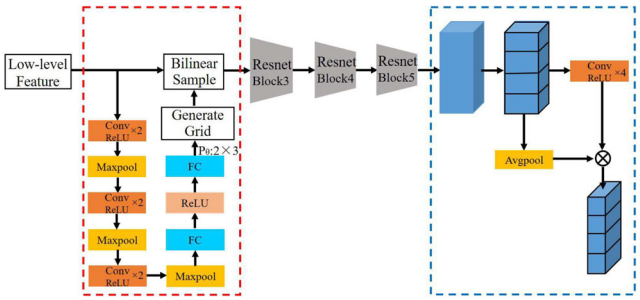


Fig. 3 The architecture of FDNet. The feature transformation branch and horizontal partition branch are circled in the red and blue dotted boxes, respectively. We extract low-level features by the Base Network and apply a feature transformation on the shallow feature. The uniform horizontal partition is conducted on the features extracted from block5 of ResNet50

Given pedestrian images with the size of $256 \times 128 \times 3$, we use the ResNet50 [9] pre-trained on ImageNet [4] as our Base Network to extract features. We remove the last fully-connected layer and break up ResNet50 into two bursts. The first two blocks are employed as the Base Network to extract low-level features, and the rest blocks are utilized in FDNet as deep feature extractor after feature transformation. The feature transformation branch receives the shallow feature maps of the pedestrian image with a size of $64 \times 32 \times 256$ as the input, which reflects the local pattern information. The feature transformation branch as shown in the red dotted box of Fig. 3 is a lightweight network that includes three conv blocks, three max pooling layers, and two fully-connected layers. We pass the feature map generated from the Base Network to the feature transformation branch to regress a 6-dimensional transformation parameter P_θ. The transformation process can be formulated as:

$$\begin{pmatrix} x^s \\ y^s \end{pmatrix} = P_\theta \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix} \tag{2}$$

where (x^s, y^s) are the source coordinates on the shallow feature maps (feature maps generated by ResNet50 block2) and (x^t, y^t) are the target coordinates on the transformed feature maps. Parameters θ_{11}, θ_{12}, θ_{21}, θ_{22} reflect the scale and rotation, while θ_{13} and θ_{23} are the translation parameters.

The output of the feature transformation branch is fed into the deep feature representation branch including block3, block4, and block5 of ResNet50. We remove the last

down-sampling block5 and obtain a feature tensor F with a size of $16 \times 8 \times 2048$. The feature division branch is used to partition the vector F into 4 horizontal stripes, each with a size of $4 \times 8 \times 2048$. Finally, the global pooling layer and the conv layer with a size of $1 \times 1 \times 2048$ are employed on the 4 horizontal stripes to achieve the part features F_{part} whose size is 4×512. Then we use four conv layers to reckon the importance of part features adaptively, which is shown in the blue dotted box of Fig. 3 and is formulated by (3). In Fig. 3, the $\otimes$ means the element-wise multiplication operation over the column vectors of the W and F_{part}.

$$W = ReLU\left(f_{w-i,b-i}\left(F_{part-i}\right)\right), \tag{3}$$

where F_{part-i} is the i_{th} part feature tensor. $f_{w-i,b-i}$ means the i_{th} convolution operation and these 4 conv layers do not share the weights. $ReLU$ represents the ReLU layer.

In this paper, we train the FDNet based on an identification task and utilize the cross-entropy loss to optimize the parameters provided by the training labels of persons. The identification loss is described in (4), where n is the number of persons in the dataset. $\widehat{y_i}$ denotes the predicted identity, and y_i denotes the ground truth ($y_i = 0$ for all i except the target class, $y_{target} = 1$).

$$L_{loss} = \sum_{i=1}^{n} -y_i \log\left(\widehat{y_i}\right) \tag{4}$$

3.3 Dual attention siamese network

To obtain a more robust global feature for similarity measure, we establish a Siamese network with dual attention, where the identification and verification tasks are utilized jointly to train the parameters. Besides, it can be observed that some information (e.g. the hair, glasses, clothes, bags, etc.) might provide more clues than others to predict the identity of input image in the person re-ID problem. Motivated by this, we develop a dual attention mechanism to seek a more optimal global feature representation. To this end, DASNet is proposed and the architecture is shown in Fig. 4. Low-level features of two inputs extracted by the Base Network are fed into the two branches of DASNet. The branches consist of several blocks from ResNet50 and the dual attention block, and they share parameters.

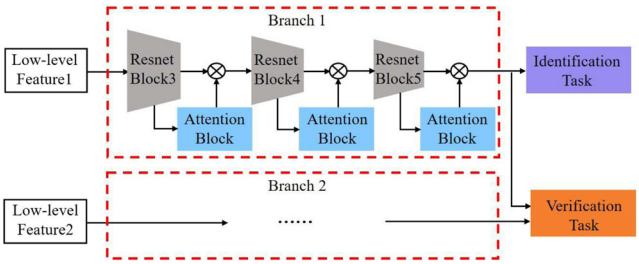


Fig. 4 The architecture of DASNet. The low-level features with a size of $64 \times 32 \times 256$ are fed into the branches of DASNet with shared parameters. The identification task predict the ID of the input and the verification task determines whether the two inputs are the same person or not

In DASNet, we construct a novel dual attention mechanism including channel attention and spatial attention, as shown in Fig. 5. The channel attention map is generated by utilizing five layers that are shown in the red dotted box and the spatial attention map is produced by five layers shown in the blue dotted box. The input of the dual attention block is a 3-D tensor $X \in \mathbb{R}^{W \times H \times C}$, where W, H, and C denote the number of pixels in the width, height, and channel dimensions, respectively. The learning of the channel attention and the spatial attention aim to produce saliency weight maps $Z_{channel} \in \mathbb{R}^{1 \times 1 \times C}$, $Z_{spatial} \in \mathbb{R}^{W \times H \times 1}$, and are formulated by (5) and (6).

$$Z_{channel} = \sigma(W_2(ReLU(W_1 AvgPool(X)))) \tag{5}$$

$$Z_{spatial} = \sigma(W_3([AvgPool(X'); MaxPool(X')])) \tag{6}$$

In (5), X is the input feature map and AvgPool denotes the average-pooling operation that aggregates spatial information of a feature map. $W_1 \in \mathbb{R}^{\frac{C}{16} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{16}}$ denote the parameters of two convolutional layers and σ indicates the sigmoid function. In (6), X' is the input feature map of spatial attention originated from $Z_{channel}$

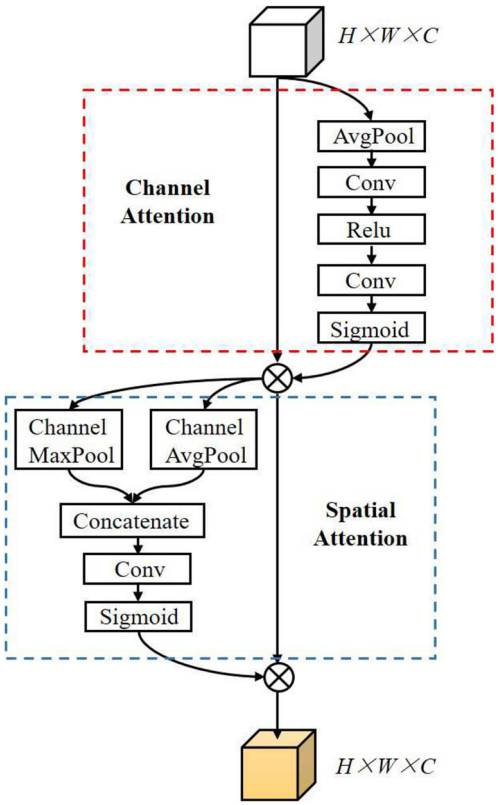


Fig. 5 The architecture of the dual attention mechanism

and X by a tensor multiplication. To compute the spatial attention, average-pooling and max-pooling operations along the channel axis are applied which are denoted by AvgPool and MaxPool, and the results are concatenated to generate an efficient feature descriptor. $W_3 \in \mathbb{R}^{7\times7}$ is the parameters of the convolutional layer. Finally, $Z_{spatial}$ and X' perform a tensor multiplication to obtain the final output of the dual attention block.

There are two branches in DASNet and the global features with dual attention are fed into two tasks. The identification task identifies the input person (IDs from the dataset) and the verification task determines whether the two inputs are the same person (the IDs are the same or not). The former task utilizes a cross-entropy loss function as described in (7) to optimize the network, where n is the number of persons, $\widehat{p_i}$ denotes the predicted ID, and p_i denotes the ground truth. $p_i = 0$ for all i except the target class $p_{target} = 1$. The verification task is a binary classification problem and the cross-entropy loss is widely used. Based on this, we further constrain the distance of the image pair to obtain a more robust global feature. We follow the strategy described in [16] and employ a novel loss for the verification task. Our loss is described in (8), where λ_1 and λ_2 are the weight coefficients of the cross-entropy loss and the new constraint. j is the class index, $j = 1$ denotes different persons, and $j = 2$ indicates the same person. y_j is the ground truth, if the image pair depicts the same person: $y_1 = 1, y_2 = 0$; otherwise $y_1 = 0, y_2 = 2$. $\widehat{y}_j$ is the predicted probability. x is the Euclidean distance of two global features whose size is $1 \times 1 \times 2048$, and c_j denotes the feature center of the j^{th} class, which is an adaptive variable. The update rule is formulated in (9), which follows the strategy in [40].

$$L_1 = \sum_{i=1}^{n} -p_i \log(\widehat{p_i}) \tag{7}$$

$$\begin{aligned} L_2 &= \lambda_1 L_{cross} + \lambda_2 L_{dis} \\ &= \lambda_1 \Big(\sum_{j=1}^{2} -y_j \log(\hat{y_j}) \Big) + \tfrac{\lambda_2}{2} \sum_{j=1}^{2} \left(y_j \left\| x - c_j \right\|_2^2 \right) \end{aligned} \tag{8}$$

$$\Delta c_k = \frac{\sum_{i=1}^{2} \delta(y_i = k) \cdot (c_k - x_i)}{1 + \sum_{i=1}^{2} \delta(y_i = k)} \tag{9}$$

Where m is the size of the mini-batch of the training stage. $\delta(condition) = 1$ if the condition is satisfied; otherwise $\delta(condition) = 0$.

To examine the effectiveness of the dual attention mechanism, we report the heatmaps achieved by the proposed DASNet and the corresponding Siamese network without the dual attention block. The results are visualized in Fig. 6 where (b) and (c) are the heatmaps of features executed by the proposed DASNet and the corresponding attention-free version. The heatmaps present the deep features and reflect the contribution of the whole image to the re-ID task, and the red area in the figure indicates the part with a large contribution. It is clear that the general Siamese network highlights information related to person re-ID, but with few useful clues and even false information for the re-ID task. For instance, vehicles, trees, and other background contents are emphasized in (c). In contrast, the proposed model can effectively eliminate the above issues. It focuses on pedestrians and suppresses the weights of the background. This demonstrates that the discriminative power of deep features extracted by DASNet model can be significantly enhanced compared with the version without the dual attention block.

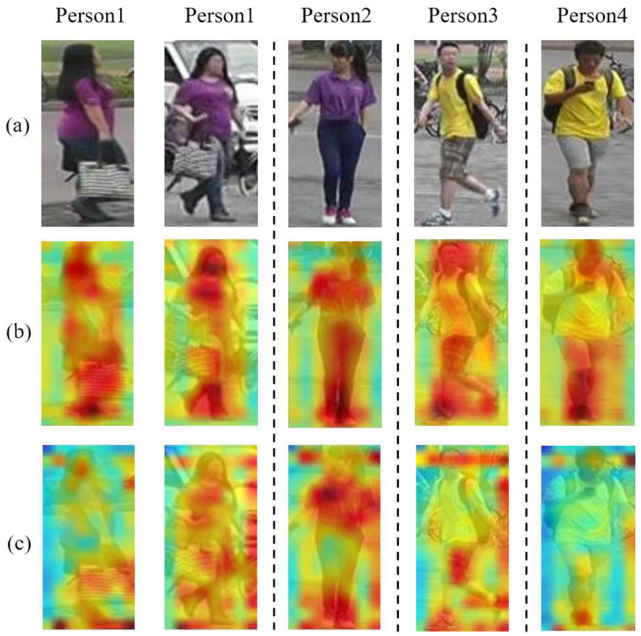


Fig. 6 The heatmaps of deep features processed by two different models. **a** The original inputs. **b** The heatmaps of features processed by DASNet. **c** The heatmaps of features processed by the corresponding Siamese network without the dual attention block. Note that features from the outputs of the last layer of ResNet50 block4 are utilized

4 Experiments

We first disclose the detailed experimental benchmarks in Section 4.1, and then the implementation details of the proposed re-ID method are described in Section 4.2. Experimental results against numerous state-of-the-arts on three benchmarks are reported in Section 4.3, and finally, we examine and analyze the effectiveness of each component of DUNet in Section 4.4.

4.1 Datasets

The proposed method is evaluated on three person re-ID datasets including Market1501 [53], DukeMTMC-reID [26, 55], and CUHK03 [37], and the specifications of these datasets are reported in Table 1.

Market1501: The dataset contains 1,501 identities, 19,732 gallery images, and 12,936 training images observed under 6 camera views. On average, there are 3.6 images for each person captured from each angle. These images can be classified into two types, i.e., cropped pedestrian images and pedestrian images automatically detected by the DPM [5]. As Market1501 provides a training set and a testing set, we use images in the training set to learn the proposed model and follow the protocol [53] to report the re-ID performance.

DukeMTMC-reID: The dataset contains 1,404 identities, 2,228 queries, 17,661 gallery images, and 16,522 training images observed under 8 camera views. The pedestrian images are cropped from hand-crafted bounding boxes. DukeMTMC-reID is one of the most challenging re-ID datasets because pedestrians in the dataset wear similar clothes and might be occluded by cars or trees. In this evaluation, we follow the evaluation protocol in [55].

CUHK03: This dataset consists of 14,096 images from 1,467 identities under 6 campus camera views, with an average number of 4.8 images under each camera. Each person appears in only two views. The dataset provides two types of annotations, including manually labeled pedestrian bounding boxes and bounding boxes automatically detected by the DPM [5]. In this work, we denote the two corresponding subsets as labeled dataset and detected dataset, respectively. The new training/testing protocol proposed in [56] is adopted. The setting of re-ranking [56] employs a larger test gallery and is different from the works published earlier, such as [34] and [22]. There are 767 identities in the training set and 700 identities in the test set (The earlier setting uses 1,367 IDs for training and the other 100 IDs for testing).

Table 1 The specifications of three person re-ID benchmarks

Dataset	Market1501	DukeMTMC	CUHK03
Identities	1,501	1,404	1,467
Cameras	6	8	6
Images	32,643	36,411	14,096
Training IDs	750	702	767
Testing IDs	751	702	700
Probe images	3,368	16,522	2,100
Gallery images	19,732	17,661	6,032

4.2 Implementation details

Pretreatment: On all datasets, we resize the training images to 256 × 128. All images are subtracted by the mean image computed from the training images. The datasets are shuffled and a random order of images is employed.

Training phase: We use the ResNet50 [9] pre-trained on ImageNet [4] as our Base Network during the training phase. The initial learning rate of layers in our Base Network is set to 0.01 while the learning rate of layers in FDNet and DASNet is set to 0.1. We decay the learning rate by a factor of 0.1 every 30 epochs. The parameters of layers in FDNet and DASNet are initialized from a zero-mean Gaussian distribution with a standard deviation of 0.01. In the verification task, the parameters λ_1, λ_2 in (8) are set to 1 and 0.001, respectively. We set the maximum number of training epochs to 60 and the batch size of images to 32. In our experiments, the mini-batch SGD is adopted to update the parameters of the network.

Testing phase: Given a 256 × 128 image, we feed the image to FDNet and one branch of DASNet to obtain four 512-dim part-features and a global feature with a dimension of 2,048. The pedestrian descriptor is generated by combining the global and part features. In the evaluation stage of the test phase, the descriptor of a given query is extracted online, while the descriptors of the gallery images are obtained off-line. The features of all images are L2-normalized and we sort the Euclidean distance between the query and all the gallery features to obtain the final ranking result.

Evaluation protocol: The cumulated matching characteristic (CMC) and mean of Average Precision (mAP) are adopted as the standard metrics to evaluate the re-ID accuracy on all three datasets. CMC denotes the probability of whether one or more correctly matched images appear in top-i and we report the mean rank-i accuracy for all query images. The mAP is proposed by Zheng et al. [53] which computes the area under the Precision-Recall curve for each probe and then calculates the mean of Average Precision over all probes. At present, more and more researchers utilize both CMC and mAP as evaluation criteria [33, 34, 53].

4.3 Comparison with State-of-the-Arts

Evaluation on Market1501 The proposed method is evaluated on Market1501 against 25 existing methods as shown in Table 2. The compared methods are categorized into two groups, i.e., traditional hand-crafted methods and the deep learning methods. Our network shows a superior performance against most state-of-the-art approaches with a significant advantage. Specifically, under the single query

Table 2 Market1501 evaluation. 1st/2nd best in red/blue

Method	rank1	rank5	rank10	mAP
LOMO + XQDA [19]	43.79	–	–	22.22
BoW+Kissme [53]	44.42	63.90	72.18	20.76
WARCA [12]	45.16	68.12	76.00	–
KLFDA [14]	46.50	71.10	79.90	–
TMA [23]	47.92	–	–	22.31
LDNS [47]	55.43	–	–	29.87
HVIL [35]	78.00	–	–	–
PersonNet [20]	37.21	–	–	26.35
DGDropout [32]	59.53	–	–	31.94
Gate S-CNN [33]	65.88	–	–	39.55
LSTM S-CNN [34]	61.60	–	–	35.50
PIE [52]	79.33	90.76	94.41	55.95
Spindle [48]	76.90	91.50	94.60	–
PAR [49]	81.00	92.00	94.70	63.40
SVDnet [31]	82.30	92.30	95.20	62.10
PAN [54]	82.80	–	–	63.40
PDC [29]	84.40	92.70	94.90	63.40
JLML [17]	85.10	–	–	65.50
SAN [28]	85.90	94.90	97.00	70.10
APR [21]	87.04	95.10	96.42	66.89
PL-Net [45]	88.20	–	–	69.30
IDE+Camera+RE [57]	89.49	–	–	71.55
GLAD [38]	89.90	–	–	73.90
HA-CNN [18]	91.20	–	–	75.70
DHA-Net [36]	91.27	–	–	75.95
Ours	91.60	97.56	98.02	75.90

mode, the proposed network achieves a rank1 accuracy of 91.6% and mAP of 75.9%. Although the performance of DHA-Net [36] is slightly higher than ours by a margin of 0.05% measured in mAP, the proposed model exceeds DHA-Net by a gain of 0.33% in terms of rank1. Moreover, our performance is better than that of DHA-Net on other datasets, such as DukeMTMC-reID, which indicates that the proposed method is more robust and possesses a discriminative ability on multiple datasets.

In particular, the HA-CNN model also uses an attention mechanism, which provides a baseline of re-ID models with attention. It optimizes person re-ID in misaligned images by joint learning of soft attention and hard regional attention. However, it ignores the reality that the part features with different contributions may have a better performance than just using global features. The proposed model considers the combination of part features with different contributions and the global feature to optimize person re-ID in misaligned images. The experimental results demonstrate the superiority of our model over HA-CNN, either by rank1 (91.60% vs. 91.20%) or mAP (75.90% vs. 75.70%).

Table 3 DukeMTMC-reID evaluation. 1st/2nd best in red/blue

Method	rank1	mAP
BoW+Kissme [53]	25.1	12.2
LOMO + XQDA [19]	30.8	17.0
ResNet50 [9]	65.2	45.0
ResNet50+LSRO [55]	67.7	47.1
APR [21]	70.7	51.9
JLML [17]	73.3	56.4
SVDnet-caffeNet [31]	76.6	45.8
SVDnet-ResNet50 [31]	76.7	56.8
SAN [28]	77.9	58.8
IDE+Camera+RE [57]	78.3	57.6
HA-CNN [18]	80.5	63.8
DHA-Net [36]	81.3	64.1
Ours	82.1	66.5

Evaluation on DukeMTMC-reID The proposed method is further evaluated on DukeMTMC-reID against 12 methods including the traditional hand-crafted methods of BoW+Kissme [53], LOMO+XQDA [19] and deep learning methods of ResNet50 [9], ResNet50+LSRO [55], APR [21], JLML [17], SVDnet-caffeNet [31], SVDnet-ResNet50 [31], IDE+Camera+RE [57] and HA-CNN [18] in Table 3. As observed, ours performs the best, followed by DHA-Net, which achieves 81.3% in rank1 and 64.1% in mAP. The proposed DUNet outperforms it by a margin of 0.8% in rank1 and 2.4% in mAP.

Evaluation on CUHK03 We evaluate our method on both manually labeled and auto-detected (more misalignment) person bounding boxes of the CUHK03 benchmark. Following the setting protocol in [56], we split the dataset into a training set containing 767 identities and a test set containing 700 identities. Results for the setting of manually labeled and auto-detected (more misalignment) bounding boxes are shown in Tables 4 and 5, respectively.

Table 4 CUHK03 evaluation using manually labeled bounding boxes. 1st/2nd best in red/blue

Method	rank1	mAP
BoW+XQDA [53]	7.93	7.29
LOMO + XQDA [19]	14.8	13.6
ResNet+XQDA [56]	32.0	29.6
ResNet+XQDA+re-rank [56]	38.1	40.3
PAN [54]	36.9	35.0
PAN+re-rank [54]	43.9	45.8
HA-CNN [18]	44.4	41.0
Ours	54.6	52.2

Table 5 CUHK03 evaluation using auto-detected (more misalignment) bounding boxes. 1st/2nd best in red/blue

Method	rank1	mAP
BoW+XQDA [53]	6.36	6.39
LOMO + XQDA [19]	12.8	11.5
ResNet+XQDA [56]	31.1	28.2
ResNet+XQDA+re-rank [56]	34.7	37.4
PAN [54]	36.3	34.0
MultiScale [3]	40.7	37.0
SVDnet [31]	41.5	37.3
HA-CNN [18]	41.7	38.6
PAN+re-rank [54]	41.9	43.8
Ours	51.6	49.9

The comparison methods are categorized into 2 groups including the traditional hand-crafted methods and deep learning methods, and the rank1 accuracy and mAP (%) are reported. We evaluate our method on CUHK03 using manually labeled bounding boxes against 7 existing methods (Table 4). Nine state-of-the-arts are compared on CUHK03 using auto-detected (more misalignment) bounding boxes (Table 5). It can be observed that our model achieves the best re-ID accuracy under both protocol settings in rank1 and mAP, followed by HA-CNN and PAN+re-rank. The proposed DUNet outperforms all the hand-crafted feature-based methods and the deep learning models. In particular, it exceeds the second-best model HA-CNN/PAN+re-rank by a large margin of +10.2%/+6.4% in rank1/mAP on the labeled set. The proposed model outperforms the PAN+re-rank model by a gain of +10.3% in rank1 and +6.1% in mAP on the detected set. The good performance of PAN+re-rank owes to the re-rank strategy and if the DUNet equips this strategy, it upgrades the performance as well (see more details in Section 4.4). The rank1 and mAP of HA-CNN, which has a similar performance to ours in Market1501, drop to 44.4% and 41.0%, respectively. This is a more than 10% decline compared with our DUNet, which demonstrates the robustness of the proposed model. Since the query and gallery sets in CUHK03 come from different camera views, it is more challenging than the other two datasets. However, our model still shows large improvements against HA-CNN on CUHK03 dataset, which indicates that the proposed model can extract discriminative features to process pedestrian images from different camera views.

4.4 Further analysis and discussions

We further examine the effectiveness of components in DUNet and inspect the impacts of pre-processing and post-processing steps on our model using the Market1501 benchmark.

Effectiveness of Individual Components The effectiveness of components in the proposed re-ID network is evaluated in this section and Table 6 presents the results. We establish a Siamese network based on ResNet50 as our Baseline. The network is trained by the identification and verification tasks and employs cross-entropy as the loss function. Several person re-ID results on Market1501 produced by the Baseline and our DUNet are shown in Fig. 7, where five typical queries with different pedestrians and camera views (including the front, back, and side of the body) are selected. It can be observed from the retrieval results that our model performs superiorly for the re-ID task regardless of the camera views of the queries, even with large variations in the gallery such as human appearances, light, camera views, background illuminations, and so on.

A total of five models with different components are compared including Baseline, Baseline + dual attention (dual attention integrated on ResNet50), DASNet (Instead of using cross-entropy for verification task in Baseline + dual attention model, DASNet model employs a novel loss described in [16]), DASNet + FDNet* (FDNet* is the network without the feature transformation branch), and the DUNet model. Figure 8 reports the CMC curves of these models. CMC shows the probability of whether one or more correctly matched images appear in top-i. We report the top-1 to top-50 mean accuracy for all query images. The horizontal axis of CMC is top-i, and the vertical axis is the mean accuracy (%) at top-i. The accuracy at top-i is the ratio of the number of correctly matched images in the ranked top-i retrieval results to the total number of query samples. It is clear from Table 6 and Fig. 8 that the DASNet has a significant influence to the re-ID performance, and it improves the performance by a gain of 7.0% (88.5 v.s. 81.5) and 10.0% (71.9 v.s. 61.9) by rank1 and mAP measure on Market1501. We further inspect the contributions of FDNet to the overall re-ID performance and the experimental results are shown in Table 6. As we can see that the FDNet boosts the overall performance by a margin of 3.1% (91.6 v.s. 88.5) in rank1 and 4.0% (75.9 v.s. 71.9) in mAP.

Impacts of the number of parts In this part, we analyze the impacts of the number of parts to the overall re-ID performance and it is evaluated on Market1501. Note

Table 6 Evaluation on the effectiveness of components in DUNet

Method	rank1	mAP
Baseline	81.5	61.9
Baseline + dual attention	86.4	66.7
DASNet	88.5	71.9
DASNet + FDNet*	90.1	73.1
DUNet	91.6	75.9

Fig. 7 Examples of person re-ID results on Market1501. A total of five examples are presented, and each row shows the top-10 retrieved images of Baseline and DUNet

that this number is fixed throughout all the comparative experiment evaluations. We attempt to train the proposed model with a different number of part features. When the number of parts is set to 1, it suggests that the learned feature is a global one. The results are reported in Table 7, where 4-part achieves the best performance on the dataset. The re-ID performance starts to drop with more parts, which indicates that excessive part features increase the burden of model training and consequently lead to a decline in performance.

Evaluation on DUNet with a pre-processing or a post-precessing step If a pre-processing step (such as GAN [57]) or a post-processing step (such as re-rank [56]) are employed, the re-ID performance of proposed model could be further improved. In this part, we report the performance of the proposed model with a pre-processing or a post-processing step on Market1501, and the results are displayed in Table 8.

In the pre-processing step, we use the GAN [57] as our data augmentation approach. We utilize the CycleGAN model learned in [57] to generate new samples in the style of other cameras and these new samples directly borrow the label information from the original training images. The original training images and the new samples make up our training set for the proposed re-ID model. In terms of the post-processing step, we obtain the rank list by sorting the Euclidean distance of features captured by the proposed model, and then perform re-ranking following the strategy in [56] to obtain a better re-ID result. This strategy uses a k-reciprocal encoding method and combines the original

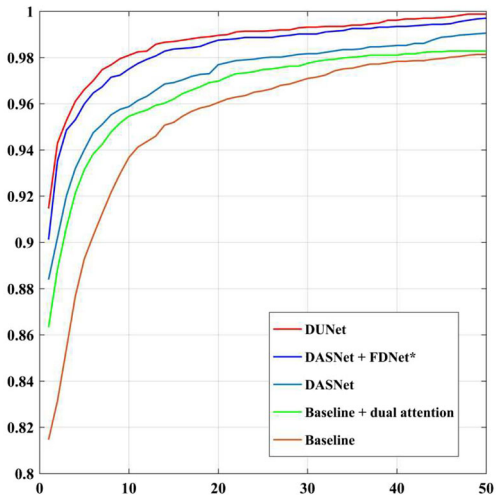


Fig. 8 The CMC curves of the our models with different components

Table 7 Impacts of the number of parts in FDM to the overall person re-ID performance

Number	rank1	mAP
1	86.3	70.4
2	89.9	73.6
4	91.6	75.9
6	91.1	74.8
8	90.2	72.7
10	89.2	71.9

Table 8 Evaluation on the proposed model with a pre-processing (GAN [57]) or post-processing step (re-rank [56])

Method	rank1	mAP
DUNet	91.6	75.9
DUNet + GAN [57]	92.5	76.1
DUNet + re-rank [56]	92.1	83.2

Euclidean distance and the Jaccard distance as the final distance. If a gallery image is similar to the query in the k-reciprocal nearest neighbors, it is more likely to be a true match. It can be observed from Table 8 that our model with a pre-processing method boosts the overall re-ID performance by a margin of 0.9%, 0.2% in rank1, and mAP, respectively. In terms of the re-rank strategy (post-precessing), mAP and rank1 measurements increase to 92.1% (+0.5%) and 83.2% (+7.3%).

5 Conclusion

In this paper, a novel architecture of discriminative representation learning, Divide-and-Unite Network (DUNet), is developed to fully exploit both global information and discriminative part information for person re-ID. Our architecture consists of the Feature Division Network (FDNet) and Dual Attention Siamese Network (DASNet). FDNet suppresses the impacts of spatial diversity on the uniform partition and extracts part features with different contributions. DASNet generates the global feature for similarity measure. Extensive evaluations are conducted on three major re-ID benchmarks to validate the performance of the proposed architecture against a wide range of state-of-the-art methods. Furthermore, the inspection of model components and pre-and post-processing steps are provided to demonstrate the design of the proposed re-ID architecture.

Acknowledgments This work was supported by the Fundamental Research Funds for the Central Universities of China (2020YJS040) and the Natural Science Foundation of China (61972027). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

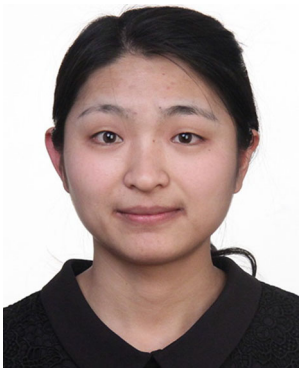
1. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: IEEE Conference on computer vision and pattern recognition(CVPR), pp 1320–1329
2. Chen W, Chen X, Zhang J, Huang K (2017) A multi-task deep network for person re-identification. In: 31St AAAI conference on artificial intelligence, pp 3988–3994
3. Chen Y, Zhu X, Gong S (2017) Person re-identification by deep learning multi-scale representations. In: IEEE International conference on computer vision workshop
4. Deng J, Dong W, Socher R, Li JL, Li K, Li FF (2009) Imagenet: a large-scale hierarchical image database. In: IEEE Conference on computer vision and pattern recognition
5. Felzenszwalb PF, Mcallester DA, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. in cvpr. In: IEEE Conference on computer vision and pattern recognition
6. Gao P, Yuan R, Wang F, Xiao L, Fujita H, Zhang Y (2020) Siamese attentional keypoint network for high performance visual tracking. Knowledge-based systems 193
7. Gao P, Zhang Q, Wang F, Xiao L, Fujita H, Zhang Y (2020) Learning reinforced attentional representation for end-to-end visual tracking. Inform Sci 517:52–67
8. Geng M, Wang Y, Xiang T, Tian Y (2016) Deep transfer learning for person reidentification. arXiv:1611.05244
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 770–778
10. Hermans A, Beyer L, Leibe B (2017) Defense of the triplet loss for person re-identification. arXiv:1703.07737
11. Hirzer M (2012) Large scale metric learning from equivalence constraints. In: IEEE Conference on computer vision and pattern recognition(CVPR), pp 2288–2295
12. Jose C, Fleuret F (2016) Scalable metric learning via weighted approximate rank component analysis. In: European conference on computer vision
13. Juengling K, Bodensteiner C, Arens M (2010) Person re-identification in multi-camera networks. In: Computer vision and pattern recognition workshops, pp 55–61
14. Karanam S, Gou M, Ziyan W, Rates-Borras A, Camps O, Radke RJ (2016) A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 1–1
15. Layne R, Hospedales TM, Gong S (2012) Person re-identification by attributes. In: BMVC
16. Li R, Zhang B, Kang D-J, Teng Z (2019) Deep attention network for person re-identification with multi-loss. Computers & Electrical Engineering 79:106455
17. Li W, Zhu X, Gong S (2017) Person re-identification by deep joint learning of multi-loss classification. In: IJCAI International joint conference on artificial intelligence, pp 2194–2200
18. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: The IEEE conference on computer vision and pattern recognition (CVPR)
19. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2197–2206
20. Lin W, Shen C, Van Den Hengel A (2016) Personnet: Person re-identification with deep convolutional neural networks. arXiv:1601.07255
21. Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. Pattern Recognition 95:151–161
22. 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):3492–506
23. Martinel N, Das A, Micheloni C, Roy-Chowdhury AK (2016) Temporal model adaptation for person re-identification. In: European conference on computer vision
24. Matsukawa T, Suzuki E (2016) Person re-identification using cnn features learned from combination of attributes. In: 23Rd

international conference on pattern recognition (ICPR), pp 2428–2433
25. Oreifej O, Mehran R, Shah M (2010) Human identity recognition in aerial images. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 709–716
26. 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
27. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 815–823
28. Shen C, Qi G-J, Jiang R, Jin Z, Yong H, Chen Y, Hua X-S (2019) Sharp Attention Network via Adaptive Sampling for Person Re-Identification. IEEE Trans Circ Syst Vid Technol 29:3016–3027
29. Chi S, Li J, Zhang S, Xing J, Gao W, Qi T (2017) Pose-driven deep convolutional model for person re-identification. In: IEEE International conference on computer vision (ICCV), pp 3980–3989, 10
30. Sun Y, Liang Z, Yi Y, Qi T, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: European conference on computer vision
31. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: IEEE International conference on computer vision
32. Tong X, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: Computer vision and pattern recognition(CVPR)
33. Varior RR, Haloi M, Wang G (2016) Gated siamese convolutional neural network architecture for human re-identification. In: Computer vision - ECCV 2016. 14th european conference., pp 791–808
34. Varior RR, Shuai B, Jiwen L, Dong X, Wang G (2016) A siamese long short-term memory architecture for human re-identification. In: Computer vision - ECCV 2016. 14th european conference, pp 135–153
35. Wang H, Gong S, Zhu X, Tao X (2016) Human-in-the-loop person re-identification. In: European conference on computer vision
36. Wang Z, Jiang J, Wu Y, Ye M, Bai X, Satoh S (2020) Learning Sparse and Identity-Preserved Hidden Attributes for Person Re-Identification. IEEE Trans Image process 29(1):2013–2025
37. Li W, Rui Z, Tong X, Wang XG (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Computer vision and pattern recognition
38. Wei L, Zhang S, Yao H, Gao W, Qi T (2019) Glad: Global-local-alignment descriptor for pedestrian retrieval. IEEE Transactions on Multimedia 21(4):986–999
39. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
40. Wen Y, Zhang K, Li Z, Yu Q (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision (ECCV)
41. Xiao Q, Luo H, Zhang C (2017) Margin sample mining loss: A deep learning based method for person re-identification. arXiv:1710.00478
42. Jing X, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. arXiv:1805.03344
43. Yang K, He Z, Zhou Z, Fan N (2020) Siamatt: Siamese attention network for visual tracking. Knowledge-based systems 203
44. Yang X, Wang M, Tao D (2018) Person re-identification with metric learning using privileged information. IEEE Trans Image Process PP(99):1–1
45. Yao H, Zhang S, Zhang Y, Li J, Qi T (2017) Deep representation learning with part loss for person re-identification. IEEE Trans Image Process PP(99):1–1
46. Yi D, Lei Z, Li SZ (2014) Deep metric learning for practical person re-identification. Computer Science, pp 34–39
47. Li Z, Xiang T, Gong S (2016) Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1239–1248
48. 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: Computer vision and pattern recognition(CVPR), pp 907–915
49. Zhao L, Xi L, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: IEEE International conference on computer vision (ICCV), pp 3239–3248
50. Zhao R, Ouyang W, Wang X (2013) Unsupervised salience learning for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3586–3593
51. Zhedong Z, Liang Z, Yi Y (2018) A discriminatively learned cnn embedding for person re-identification. Acm Transactions on Multimedia Computing Communications and Applications 14(1):13:1–13:20
52. Zheng L, Huang Y, Huchuan L, Yi Y (2019) Pose-invariant embedding for deep person re-identification. IEEE Trans Image Process 28(9):4500–4509
53. Zheng L, Shen L, Tian L, Wang S, Wang J, Qi T (2015) Scalable person re-identification: a benchmark. In: IEEE International conference on computer vision
54. Zheng Z, Zheng L, Yi Y (2017) Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology
55. Zheng Z, Zheng L, Yi Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE International conference on computer vision
56. 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
57. Zhong Z, Zheng L, Zheng Z, Li S, Yi Y (2018) Camera style adaptation for person re-identification. In: IEEE Conference on computer vision and pattern recognition

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

Rui Li received the B.S. degree and M.S. degree in computer sciense from Beijing Technology and Business University and Beijing Jiaotong University, respectively, in 2017 and 2019, Beijing, China. Now, she is working towards the Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests are both theory and applications of computer vision.

Baopeng Zhang received his Ph.D. degree in computer science from Tsinghua University in 2008. He is currently an associate professor in the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include semantic image/video classification and retrieval, statistical machine learning, large-scale semantic data management and analysis, and image privacy protection.



Zhu Teng received her B.S. degree in automation from Central South University, Changsha, China, in 2006, and the Ph.D. degree in intelligent control and automation from Pusan National University, South Korea, 2013. From 2013 to 2015, she was a postdoctoral researcher with the school of Computer and Information Technology, Beijing Jiaotong University. From 2015 to 2016, she was an assistant professor at Beijing Jiaotong University. She is now an associate professor in the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests are computer vision and machine learning.

Jianping Fan was a professor at UNC-Charlotte. He received his MS degree in theory physics from Northwestern University, Xian, China in 1994 and his PhD degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. He was a Postdoc Researcher at Fudan University, Shanghai, China, during 1997-1998. From 1998 to 1999, he was a Researcher with Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From 1999 to 2001, he was a Postdoc Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. His research interests include image/video privacy protection, automatic image/video understanding, and large-scale deep learning.