Neurocomputing 334 (2019) 68–78

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Combining multilevel feature extraction and multi-loss learning for person re-identification

Weilin Zhong[a], Linfeng Jiang[a], Tao Zhang[a], Jinsheng Ji[a], Huilin Xiong[a,b,*]

[a] School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, China
[b] Institute for Sensing and Navigation, Shanghai Jiao Tong University, China

ARTICLE INFO

Article history:
Received 19 June 2018
Revised 3 November 2018
Accepted 5 January 2019
Available online 17 January 2019

Communicated by Dr Jianjun Lei

Keywords:
Multilevel feature extraction
Multi-loss learning
Recurrent comparative network

ABSTRACT

The goal of person re-identification (re-id) is to match images of the same person captured by multiple cameras with non-overlapping views. It is a challenging task due to the large spatial displacement and human pose change of person images across different views. Recently, the deep Convolutional Neural Network (CNN) has significantly improved the performance of person re-id. In this paper, we present a hybrid deep model that combines multilevel feature extraction and multi-loss learning for more robust pedestrian descriptors. The multi-loss function jointly optimizes the verification task that aims to verify if two images belong to same person, and the recognition task that aims to predict the identity of each image. Specifically, given two person images, we first apply a deep learning network, called Feature Aggregation Network (FAN), to extract their multilevel CNN features by fusing the information of different layers. For the verification task, a Recurrent Comparative Network (RCN) is presented to learn joint representation of paired CNN features. RCN determines whether two images depict the same person through focusing on discriminative regions and alternatively comparing their appearance. It is an algorithmic imitation of human decision-making process, in which a person repeatedly compares two objects before making decision about their similarity. For the recognition task, a parameter-free operation termed Global Average Pooling (GAP) is followed after each CNN feature to extract identity-related features. Extensive experiments are conducted on four datasets, including CUHK03, CUHK01, Market1501 and DukeMTMC, and the experimental results demonstrate the effectiveness of our presented method.

1. Introduction

Person re-identification is the task of recognizing specific person across camera views. Given one probe pedestrian image, person re-id aims to search the same person from a gallery of candidate images, which are captured by different cameras with non-overlapping views. With the wide deployment of camera networks in public infrastructures, person re-id has great significance in video surveillance and thus attracts increasing interest in computer vision community. Due to the low resolutions of surveillance images and the limitation of economical issue, the biological patterns including face and gait may be unreliable for identifying the same person in real world applications. Hence, existing person re-id methods rely heavily on visual features. However, the appearances of the same person across camera views possess large view variations, such as spatial misalignment, human pose change and background clutter. Fig. 1 shows some difficult examples from four datasets and those difficulties challenge the application of person re-id in realistic surveillance networks. Therefore, it is necessary to develop effective person re-id approaches.

There are two critical components in person re-id systems, namely feature extraction and metric learning. For feature representations, different visual cues are adopted to extract robust features, which include Local Maximal Occurrence representation (LOMO) [1], Symmetry Driven Accumulation of Local Features (SDALF) [2], and hierarchal Gaussian descriptor (GOG) [3]. For metric learning, different distance metrics are learned from training data to minimize intra-class distance whilst maximize inter-class distance. Representative metric learning methods include relative distance learning [4], large scale metric learning from equivalence constraint (KISSME) [5] and Pairwise Constrained Component Analysis (PCCA) [6]. Previous mainstream methods optimize two key components separately and thus their performance is limited. Recently, many researchers have utilized the popular deep Convolutional Neural Network (CNN) to solve person re-id problem [7–11]. The deep learning based methods integrate feature extraction and metric learning into one unified framework. Benefiting from the

* Corresponding author at: School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, China.
E-mail address: hlxiong@sjtu.edu.cn (H. Xiong).

https://doi.org/10.1016/j.neucom.2019.01.005
0925-2312/

Fig. 1. Typical examples of cross-view person image pairs from four datasets including CUHK03, CUHK01, Market1501, and DukeMTMC. The same person captured by different camera views presents large variations in lighting and pose, background and occlusion. Besides, another challenging situation is the defection of person detecting results.

powerful feature learning ability of deep networks, the deep methods have exhibited significantly improved performance for person re-id.

Existing deep methods solve person re-id either as binary classification problem of image pairs, or feature matching problem of different images. In the first category, the joint representation of paired images is usually learned through exhaustive search of similar patches [9,10,12,13] in several specially designed layers, which are called matching networks. At the top layer of matching network, a soft-max layer is adopted to output the matching probability of two images. These methods are effective in modelling the relationship between two images and handling spatial misalignment problem. But, they suffer from computation deficiency due to the complicated matching networks. In the second category, the image features are learned from various loss functions, which include triplet loss [11,14], pairwise contrastive loss [8,15,16] and person recognition loss [17–19]. The cosine or Euclidean distance functions are utilized for measuring the similarity between those learned features. These methods are efficient in online matching while they undervalue the metric learning, using simple cosine or Euclidean distance function for similarity computation [11,14,16]. In addition, most of the methods mentioned above make operations on the last layer global features of a backbone network such as ResNet [20], which ignores the local information from lower-level layers. For example, as displayed in Fig. 1, the color of shoes and backpacks are effective in distinguishing different persons.Those small-size differences and local details could be easily missed in the top layers after numerous spatial downsampling operations such as pooling layers.

In this paper, we present a hybrid deep model to combine multilevel feature extraction and multi-loss learning. Two commonly used person re-id models are jointly optimized using a multi-loss function, which is composed of a verification loss term and a recognition loss term. Specifically, we first build a Feature Aggregation Network (FAN) that fuses multi-layered convolutional outputs from ResNet [20] to extract multilevel CNN features for the input image pair. For the verification loss, we present a matching network called Recurrent Comparative Network (RCN) to learn joint representation. Our RCN, which is built upon the popular attention based Long Short-Term Memory (LSTM) [21,22], features two modules, namely attention module and recurrence module. The attention module aims to focus on distinct regions within feature maps, and the recurrence module aims to alternatively compare two images. In particular, we alternate the CNN feature pair to form the sequence input of the attention LSTM and the hidden states of LSTM are viewed as joint representations of image pair, which are then sent to a soft-max binary classifier outputting the matching probability. For the recognition loss, we directly pool the CNN feature maps using Global Average Pooling (GAP) and the pooled features are regarded as the identity-related features of each image. Both objectives are soft-max classification loss, in which the number of soft-max output nodes is two for verification loss and is identical to the number of person identities for recognition loss. During testing stage, the matching probability and cosine distance are combined as similarity score.

This paper has three main contributions: (1) We improve the CNN backbone proposed in [20] through fusing multilevel features, and experimental results show it can get better performance. (2) We present a Recurrent Comparative Network (RCN) to learn joint representation of image pair. Our RCN can focus on the discriminative regions and alternatively compare the appearance similarity. (3) Two types of similarity computation are combined in a multi-loss model architecture. The experiments conducted on four benchmark datasets including CUHK03 [10], CUHK01 [23], Market1501 [24], and DukeMTMC [25] demonstrate the effectiveness of the presented method.

The rest of this paper are organized as follows. In Section 2, we review the related studies. In Section 3, the architecture of our method is presented. The experimental results and model analysis are demonstrated in Section 4, followed by a conclusion drawn in Section 5.

2. Related work

Person re-id is a valuable technique in visual surveillance and has drawn increasing attention in recent years. In general, person re-id methods can be divided into two typical stages: (1) extracting robust features to represent images; (2) learning distance metrics to measure feature similarity. For feature representations, the most useful features are usually based on color and

texture information, which include RGB, LAB, color names [26], the local binary patterns (LBP) [5,27–29], Gabor filter feature [29], color histogram and its variants [5,28,30] etc. For example, Li and Wang [29] proposed to combine four types of visual features, which included LBP, HSV color histogram, Gabor and HoG. Gray and Tao [31] used AdaBoost algorithm to select the most discriminative features. Liao et al. [1] constructed a feature descriptor by maximizing the horizontal occurrence of local features. Kviatkovsky et al. [32] presented an illumination-invariant color descriptor based on the color intra-distribution signatures. Farenzena et al. [2] extracted both symmetry and asymmetry color and texture features (SDALF) for person re-id. For metric learning, many machine learning algorithms are utilized to learn a more discriminative feature space, in which the intra-class distances are minimized while the inter-class distances are maximized. For example, Zheng et al. [4] proposed a relative distance comparison (RDC) learning method from a probabilistic prospective. Mignon and Jurie [6] proposed the Pairwise Constrained Component Analysis (PCCA) to learn distance metrics from sparse pairwise similarity constraints. Davis et al. [33] proposed information-theoretic metric learning (ITML) method based on the Mahalanobis distance. Liao et al. [1] proposed the Cross-view Quadratic Discriminant Analysis (XQDA) to learn a discriminant low-dimensional subspace and a distance metric. Xiong et al. [27] introduced kernel tricks into linear metric models. Those methods mentioned above learn distance metric based on fixed feature descriptor by handcraft, and thus the performance is limited.

Recently, deep learning has achieved outstanding performance in various computer vision tasks, and also many researchers have explored the application of deep models in person re-id task. Existing deep methods formulate person re-id either as binary classification of image pair or feature matching of different images. In the first group, many specially designed matching networks are proposed to learn joint representation of image pair. For example, Li et al. [10] proposed a patch matching layer and a max-out grouping layer in a filter pairing neural network (FPNN) to cope with the horizontal spatial displacement. Chen et al. [12] proposed to learn the joint representation by feeding the horizontally stitched image pair into a deep ranking framework. Ahmed et al. [7] proposed to compute the cross-input neighborhood differences using a neighborhood difference layer followed by a patch summary layer. Later, Wu et al. [9] extended the idea in [7] by increasing the depth of layers and using very small convolution filters. Varior et al. [34] proposed to leverage the intermediate features using a gated CNN. In the second group, the cosine or Euclidean functions are usually applied for distance computation. Yi et al. [35] proposed a siamese convolutional neural network followed by a cosine layer to calculate pairwise similarity. Ding et al. [36] utilized the triplet loss to learn a view-invariant feature space. Cheng et al. [14] proposed to learn global and local body parts features in a multi-channel CNN. Recently, many methods solve person re-id using person recognition model due to easy implementation and no need of complex sampling scheme in the training stage as in triplet model [11,14]. For example, Xiao et al. [17] trained identities classification model by mixing multiple datasets. Zheng et al. [25] proposed to increase the training data using a Generative Adversarial Network (GAN) [37]. Zhong et al. [38,39] proposed a novel camera style transfer model called Cam-GAN based on CycleGAN [40]. Later, Zhong et al. [41] applied Cam-GAN [38,39] to unsupervised person re-id. Qian et al. [42] proposed pose-normalization GAN (PN-GAN) which synthesized realistic person images conditioned on human poses. Li et al. [43] learned global-local complementary features in an identification CNN. Su et al. [44] adopted Spatial Transformer Networks (STN) [45] for accurate body-parts localization. Sun et al. [46] proposed to decorrelate the learned weight vectors using singular vector decomposition (SVD). In [18,19], the Convolutional Pose Machines (CPM) [47] was applied to explore the fine-grained pose information for more stable feature representations.

To combine the advantages of two types of person re-id strategies, our method integrates verification and recognition task in one unified framework. For the verification task, our Recurrent Comparative Network (RCN) is built upon the attention LSTM, which is widely applied to image caption [22,48] and action recognition [21]. The alternative appearance comparing process in our method is similar to the human decision-making mechanism, in which a person looks back and forth between two images when he/she is asked to verify whether they belong to the same identity. Each glimpse on the image generates a specific observation. A series of such observations between two images are then cumulatively used to make a judgement about their similarity. Similar ideas are explored in previous literatures including Attentive Recurrent Comparators (ARC) [49], Deep Co-attention based Comparators (DCCs) [50] and Comparative Attention Model (CAN) [11]. Our work departs from those methods in an improved CNN backbone to extract spatial features and a multi-loss model to learn more robust pedestrian descriptors. Moreover, our RCN differs from ARC [49] and DCCs [50] in the attention LSTM, which is first used for person re-id in CAN [11]. Compared to CAN [11], our method outputs matching probability at each comparison and the attention LSTM in our model can be seen as a matching network, which learns a joint embedding of image pairs.

Our method is also related to the models trained using multi-loss functions [8,43,51]. Triplet loss and ranking loss are combined in [8]. Verification and recognition loss are jointly optimized in [51]. Different model architectures are proposed to combine two supervisory signals. For instance, the work in [8] utilizes two different subnetworks encapsulated with several convolutional layers to obtain single-image and cross-image representations, respectively. In [51], a specially designed layer, named square layer, is used to learn joint representations of image pair, and a fully-connected layer is adopted to extract identity-related features. Different from these works, we employ RCN to learn a joint embedding of paired images and a parameter-free operation termed GAP for identity-related feature extraction. Furthermore, joint training of two tasks is a common strategy to improve the performance of video based person re-identification [52,53]. In addition, our Feature Aggregation Network (FAN) is inspired by the object detection methods that explore multi-layered information for better detection accuracy. For instance, Kong et al. [54] adopted cascaded features of different layers for object detection. Lin et al. [55] proposed a Feature Pyramid Network that used a top-down module to build lateral connections between multi-scale features. The purpose of these methods is to construct feature pyramids while our method aims to fuse multilevel features.

3. Proposed model architecture

The overall architecture of our method is shown in Fig. 2. First, we present a Feature Aggregation Network (FAN, Section 3.1) to extract multilevel CNN features for the input image pairs. Then, the paired CNN features are forward into Recurrent Comparative Network (RCN, Section 3.2) to learn joint representation of two images for verification task, and meanwhile, the GAP operation is adopted to pool the CNN maps to obtain identity-related feature of each image for recognition task. Finally, the soft-max loss is utilized to train both tasks in an end-to-end fashion (Section 3.3).

3.1. Feature aggregation network

Typically, different layers of a deep CNN model are sensitive to different patterns of the input objects. For instance, the lower-level

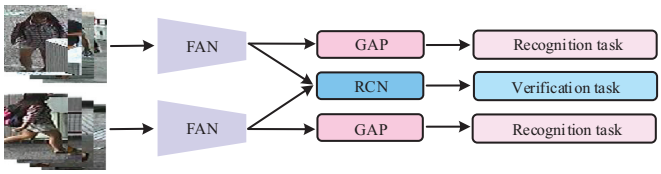


Fig. 2. Overview of our person re-identification system. The whole network architecture is trained under the joint supervision of verification and recognition loss. A Feature Aggregation Network (FAN) is first built to extract multilevel CNN features for the input two images. Then a Recurrent Comparative Network (RCN) is presented to learn joint representation of paired images for verification task, and meanwhile, the Global Average Pooling (GAP) operation is adopted to obtain identity-related features of each image for recognition task.

layers encode more local information, whereas higher-level layers capture more semantic and global information. The dominating deep methods mostly utilize the last layer outputs for person re-id. Some small visual cues that are beneficial for person re-id could be easily missed in the top layers. For example, the color of shoes and backpacks shown in Fig. 1 are effective in distinguishing different persons. We thereby fuse multi-layered features into the feature extraction module.

As shown in Fig. 3, the multi-scale feature maps extracted from different convolution blocks of ResNet-50 [20] are fused to obtain the CNN features. The features from different blocks have different resolutions and channel sizes. To fuse these multi-scale features, it is a common strategy to downsample the feature maps with pooling operation and then concatenate them along the channel axis. But, the large number of channel size in the concatenated features (e.g., concatenating the last three block features gives 512+1024+2048=3584 channels) will degrade the efficiency, especially for the recurrent comparing process in RCN. Importantly, the outputs from different levels contain various contexts and they represent different feature subspaces. Therefore, we pass the features from `Res3`, `Res4` and `Res5` to different convolution or deconvolution operations, so that the resulting feature maps are in same size. The element-wise summation is then adopted to fuse those features, generating the final CNN features with size of $12 \times 12 \times 800$. In this way, each location in the fused feature map receives multi-scale receptive fields, so that both local and global information can be encoded into the feature learning process.

3.2. Recurrent comparative network

The Recurrent Comparative Network (RCN) consists of an attention module that attends the discriminative regions and a recurrent structure that aggregates the attended information, and meanwhile, performs appearance comparing action between paired images. Our model is essentially an algorithmic imitation of human decision-making process, whose crucial characteristic is to perceive new observations at each glimpse on two images. These observations are conditioned on the previous context that has been investigated so far by the observer. A series of such observations are then accumulated to make a judgement about their similarity [49]. Similar to the work in [49], the input to our comparative network is a feature sequence, which is constructed by alternating across two images.

Fig. 4 shows the basic structure of RCN. Our RCN is built upon the attention based Long Short-Term Memory (LSTM) [21,22]. First, we revisit the LSTM model. As a variant of Recurrent Neural Networks (RNN), LSTM is a powerful tool to model the temporal dependencies and correlations over long-term time series data. The basic LSTM unit consists of a memory cell $c_t \in R^d$ and three controlling states, namely input gate $i_t \in R^d$, forget gate $f_t \in R^d$, and output gate $o_t \in R^d$, where d is vector dimensionality. The memory cell preserves the knowledge of previous step and current input while the gates control the update and flow direction of the information. Given the input $x_t \in R^D$ with dimensionality D, the LSTM unit updates the memory cell and output a hidden state $h_t \in R^d$ in the following way:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right), \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \\ h_t &= o_t \odot \phi(c_t), \end{aligned} \tag{1}$$

where σ and ϕ are element-wise sigmoid and hyperbolic tangent function, respectively. $\odot$ represents element-wise multiplication. W and b are learnable parameters, which can be trained by back-propagation through time (BPTT) algorithm [56]. We simplify the LSTM recurrence process in Eq. (1) by only considering the input x and the output hidden state h:

$$h_t = \mathrm{LSTM}(x_t, h_{t-1}). \tag{2}$$

We show how RCN integrates the appearance comparing process into LSTM using an attention module as follows. Given the input image pair $\{I_a, I_b\}$, their corresponding outputs from FAN are denoted by $\{X_a, X_b\}$, each with size of $H \times H \times C$. The recurrent structure in LSTM needs time-series input, and meanwhile, the approximation of human decision-making process usually contains repeated glimpses between two images. It is thus natural to alternate between the paired images for a finite number of times, obtaining a feature sequence of $\{X_a, X_b, X_a, X_b, \ldots, X_a, X_b\}$. Hence, at time step t, the feature cube taken to RCN can be represented as:

$$X_t = \begin{cases} X_a & \text{if } t\%2 = 0 \\ X_b & \text{else} \end{cases} \tag{3}$$

In this way, the glimpse action on the feature sequence step by step is actually a cycling process between two images, namely $X_a \to X_b$, $X_b \to X_a$, $X_a \to X_b$, etc. The tth glimpse result on X is then taken as the input of LSTM in Eq. (2), which can be expressed as:

$$x_t = f_g(X_t, P_g), \tag{4}$$

where $f_g(\cdot)$ is the glimpse function that acts on the feature cub X_t. In addition, P_g is the interval glimpse parameter.

The glimpse function can be various operations, which transfer the three-dimensional feature $X \in R^{H \times H \times C}$ into feature vector $x \in R^D$. Here, an soft attention module is adopted as the glimpse function. At each time, a location mask $l_t \in R^{H \times H}$ is generated to weight the feature map as follows:

$$\begin{aligned} x_t &= \sum l_t \odot X_t, \\ l^t &= \mathrm{softmax}(Ph_{t-1}), \end{aligned} \tag{5}$$

where Σ means summation across the spatial dimension. The feature dimensionality after summation is the channel size D, and thus $D = C$. l_t is computed by a soft-max normalized projection from previous hidden state h_{t-1}, in which the projection is parameterized by $P \in R^{D \times H^2}$. The location mask is indeed a soft-max probabilistic map, which activates higher on distinct regions.

However, for each CNN feature pair in Fig. 4, when we reverse the feature sequence $\{X_a, X_b, \ldots, X_a, X_b\}$ into $\{X_b, X_a, \ldots, X_b, X_a\}$, the matching confidence of two feature sequences will be different while they represent the same person. Here, we tackle this problem by introducing a bi-directional RCN, in which an additional backward attention LSTM takes the reversed feature sequences as input. At each time step, the bi-directional hidden states are concatenated to form the hidden state h_t, which is then sent to the binary classifier.

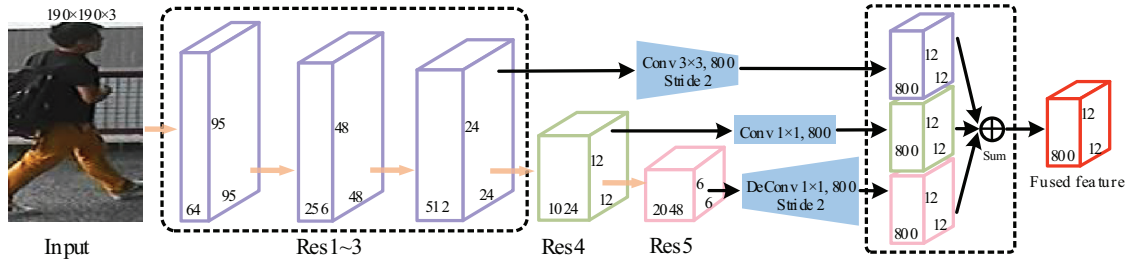


Fig. 3. The framework of the Feature Aggregation Network (FAN) based on ResNet-50 model. The outputs from `Res3`, `Res4` and `Res5` blocks are passed through different convolution or deconvolution operations to obtain feature maps with same resolution and channel size. These features are then element-wise summed to generate the final CNN features.

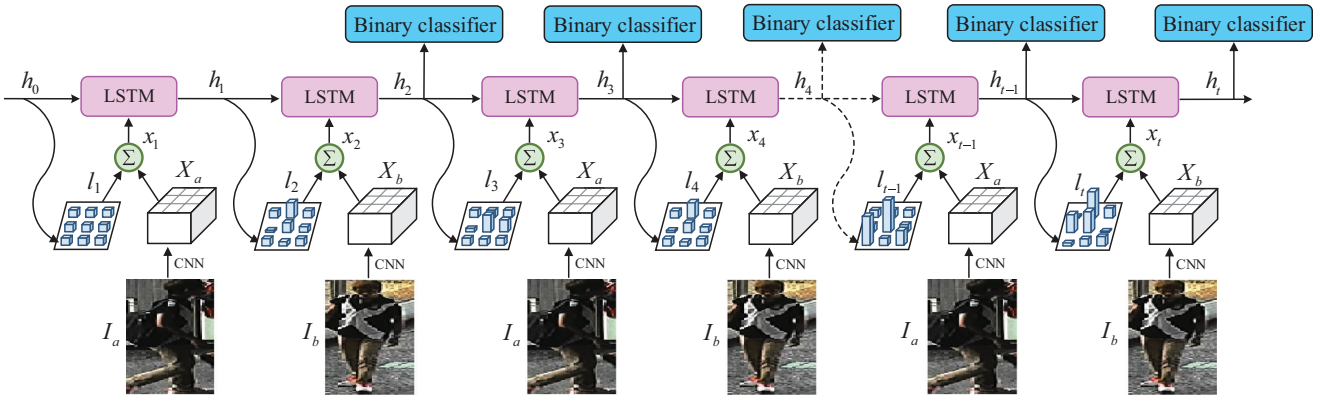


Fig. 4. The structure of the presented Recurrent Comparative Network (RCN), which works by alternatively comparing the appearance similarity of paired images. It consists of two components, namely attention module and recurrent network. The attention module weights the feature cubes while the recurrent network alternately aggregates the attended information of two images, and then, outputs the matching probability via a soft-max binary classifier at each comparison.

The attention module can easily generate similar location masks when it takes the same input, whereas the respective odd and even input images are same in RCN. Thus, a diversity regularizer is introduced to encourage the attention diversity for different regions. Since the location masks are all soft-max normalized, we flatten the location masks into vectors and use the dot product between two mask vectors to indicate their similarity. The diversity regularizer is designed as follows:

$$\mathcal{R}_{div} = \sum_{(i,j)}^{T} 1\{(i,j)\%2 = 0\}(l_i \cdot l_j)_{i \neq j} + \sum_{(m,n)}^{T} 1\{(m,n)\%2 \neq 0\}(l_m \cdot l_n)_{m \neq n}, \quad (6)$$

where $1\{\cdot\}$ is is an indicator function whose value is 1 when the expression is true, and 0 otherwise. T denotes the number of time steps.

3.3. Multi-loss learning

As shown in Fig. 2, the multi-loss framework optimizes two tasks, namely verification and recognition task. The verification task is a binary classification task, which aims to verify whether an image pair belongs to the same person. The recognition task is a multi-class classification task, which aims to predict the person identity of each image. We unify the verification and recognition task in one framework, aiming to leverage the strengthes of two tasks, and meanwhile, reduce the over-fitting. In this paper, we adopt the soft-max classification loss to train both of two tasks. Suppose we sample M images into a training batch, and generate N image pairs. Each image belongs to one of K identities.

Verification task. In RCN, corresponding to the input sequence generated by alternating paired images for G times, the recurrent network outputs a sequence of joint features as $\{h_0, h_1, h_2, h_3, \ldots, h_{2G-1}, h_{2G}\}$. In fact, the first two features h_0 and h_1 are abandoned as illustrated in Fig. 4, because only single image information instead of pairwise information is included in the first two hidden states. Then, a fully-connected layer with two output nodes is followed after the hidden states to generate the matching score $s_k = f(h_t),\ k = 0, 1$. Here, s_0 and s_1 represent the ranking score of two images belonging to the same person or not, respectively. So, the matching probability at *t-th* time step can be computed as $\hat{p}_t = \frac{exp(s_0)}{exp(s_0)+exp(s_1)}$. Finally, the verification task minimizes the following cross-entropy loss:

$$\mathcal{L}_{ver} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{t=2}^{T}\left[y^j \log \hat{p}_t^j + \left(1 - y^j\right)\log\left(1 - \hat{p}_t^j\right)\right], \quad (7)$$

where y^j is the ground truth label for jth image pair, with 0 representing the matched pair and 1 representing the unmatched pair. Besides, T is the total number of time steps.

Recognition task. For the recognition task, we generate the feature vector v of each image by directly pooling the CNN feature maps using a GAP layer. Then a fully-connected layer with K output nodes is utilized to output the identity score $z_k = f(v)$, $k = 1, \ldots, K$. So, the probability that each image belongs to the kth identity can be computed as $\hat{q}_k = \frac{exp(z_k)}{\sum_{k=1}^{K} exp(z_k)}$. Thus, the loss function of the recognition task can be defined as:

$$\mathcal{L}_{rec} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} y^i \log \hat{q}_k^i, \quad (8)$$

where y^i is the ground truth person identity.

Training phase. Combining the verification and recognition loss along with the diversity regularizer, the total loss function can be

denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{ver} + \eta\mathcal{L}_{rec} + \zeta\mathcal{R}_{div}, \tag{9}$$

where η and ζ are two balance coefficients.

Testing phase. The similarity score of the paired images $\{I_a, I_b\}$ is computed by combining the matching probability from verification task and the cosine distance between two pooled features from recognition task as follows:

$$S(a,b) = \frac{1}{T-2}\sum_{t=2}^{T}\hat{p}_t + \frac{\langle v_a, v_b\rangle}{\|v_a\|\|v_b\|}, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ means the inner product between two vectors. Note that higher similarity score means the higher probability of two images depicting the same person.

4. Experimental results

4.1. Experimental settings

Datasets. The experiments are conducted on four challenging datasets, including CUHK03 [10], CUHK01 [23], Market1501 [24] and DukeMTMC [25]. The CUKH03 dataset includes more than 13,000 images of 1467 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 Deformable Part Model (DPM) [57] detector. We evaluate our model on the bounding boxes detected by DPM, which is closer to the realistic setting. CUHK01 contains 971 persons, which are captured from two camera views. Each person has two images per view and there are in total 3884 images on this dataset. Market1501 contains 32,668 auto-detected bounding boxes of 1501 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. DukeMTMC dataset is a subset of the multi-target, multi-camera pedestrian tracking dataset [58]. We use the re-id version provided by Zheng et al. [25], which contains 34,183 images of 1401 person identities. The pedestrian bounding boxes are manually cropped. Each person is captured by at most eight different high-resolution cameras.

Evaluation protocol. We conduct the comparison experiments using single-query settings. Two widely used evaluation metrics are adopted for performance comparison, namely the cumulative matching characteristic (CMC) [59] and the mean average precision (mAP) [24]. For CUHK03 and CUHK01, we only adopt the CMC curve that returns the ranked gallery list in a descent order of the similarities to evaluate the performance. For Market1501 and DukeMTMC, we additionally report the mAP since the gallery list contains multiple matching results from different views. Each dataset is split into two subsets, namely training and testing subset with non-overlapping person identities. For CUHK03 dataset, we adopt two training/testing protocols for performance evaluation. The first one [10] randomly samples 1260 persons to form the training subset and 100 persons to constitute the testing subset. The second one is a new setting proposed in [60], which uses fixed 767 person identities for training and the rest 700 identities for testing. For CUHK01 dataset, we randomly divide the 971 persons into the training subset with 485 persons and the testing subset with 486 persons. For Market1501, we follow the standard test protocol defined by Zeeng et al. [24], which uses fixed 750 identities as training subset and the rest fixed 751 identities as testing subset. For DukeMTMC, following the evaluation protocol in [25], 1,401 identities are divided into a training subset with 702 identities and a testing subset with the rest 702 identities. Note that, for CUHK03 and CUHK01, the random split of training/testing subsets is repeated for ten times to achieve stable statistics and then the averaged matching rate is reported.

Implementation details. We implement our method based on the open source PyTorch[1] library. All the images are re-scaled into 210×210 and common data augmentation methods are applied to the resized images, including mirror flip, shifting, minor rotation, and Random Erasing (RE) [61]. Then the images are cropped into 190×190 and taken as the input of deep network. For the LSTM, we set the dimensionality of the hidden state as 400 for all the experiments. We alternate the paired images for three times, namely $T = 6$ in RCN. We train our network using stochastic gradient descent [62] with momentum of $\mu = 0.9$, weight decay of $\lambda = 5 \times 10^{-4}$. The initial learning rate is set as $\gamma = 0.005$ for CUHK03, Market1501 and DukeMTMC to accelerate convergence, while a smaller learning rate of 0.001 is set for CUHK01 because of its small data size. The learning rate decreases by a factor of 0.1 for every 8000 iterations on CUHK01, and 15,000 iterations on other datasets, respectively. In the training process, each batch contains 25 person identities, and for each identity we randomly select 4 images, and thus the batch-size is 100. The initial ratio of positive and negative pair is set as 1:1. After first decrease of learning rate, the positive-negative ratio increases to 1:2. The networks convergence stably after 30,000 iterations on CUHK01 and 60,000 iterations on other datasets.

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 LMNN [72], ITML [33], LOMO [1], KISSME [5], LMNN [63], BoW [24], GOG [3], eSDC [64], KLFDA [27], and SDALF [2]. The deep methods include FPNN [10], IDLA [7], DeepRanking [12], EmbeddingDM [16], PersonNet [9], Siamese LSTM [15], SIR&CIR [8], DGDNet [17], Gated CNN [34], End-to-end CAN [11], CNN-Embedding [51], SpindleNet [18], GAN [25], Quadruplet [65], PIE [19], DCCs [50], CPC [13], MuDeep [66], PDC [44], JLML [43], DPFL [67], IDE [60], SVDNet [46], PSE [69], OIM Loss [70], Cam-GAN [38,39], PN-GAN [42], and ACRN [71]. Note that, not all of these approaches report their performance on all of the four datasets.

Performance on CUHK03. The experimental results on two training/testing settings are shown in Tables 1 and 2, respectively. From Table 1, it can be observed that our method outperforms all the hand-crafted features based methods by a large margin. For instance, our method outperforms the best-performing hand-crafted feature GOG [3] by 22.8% in terms of Top1 matching rate. Moreover, our method performs better than all the comparing deep learning based methods. Specifically, compared to the deep methods using matching probability for similarity computation (i.e., FPNN [10], SIR & CIR [8], Gated CNN [34], CPC [13]), our method obtains at least 2.4% improvement in terms of Top1 accuracy (ours 88.3% against CPC [13] 85.9%). Our method improves the Top1 matching rate by 13.8% over Quadruplet [65] and 25.2% over End-to-end CAN [11]. The Top1 accuracy of our method is 13.0%, 7.7% and 6.3% better than MuDeep [66], JLML [43] and DPFL [67], respectively. Besides, our method outperforms PIE [19] and PDC [44] by 25.9% and 10.0%, respectively. Similar to CNN-Embedding [51] and SIR&CIR [8], verification and recognition signals are combined for model training. But, the method in [51] undervalues the pairwise relationship while our method alternatively compares the difference between the paired images.

[1] https://pytorch.org/

Table 1
Performance comparison on CUHK03 detected dataset (test=100).

Methods	Top1	Top5	Top10
SDALF [2] (ICCV10)	4.9	21.2	35.1
LMNN [63] (NIPS05)	6.3	17.5	28.2
eSDC [64] (CVPR13)	7.7	21.9	34.9
KISSME [5] (CVPR12)	11.7	33.9	48.2
LOMO [1] (CVPR15)	46.3	78.9	88.6
GOG [3] (CVPR16)	65.5	88.4	93.7
FPNN [10] (CVPR14)	19.9	50.0	64.0
EmbeddingDM [16] (ECCV16)	52.1	83.2	92.6
SIR&CIR [8] (CVPR16)	52.2	85.0	92.0
PIE [19] (arXiv17)	62.4	87.0	91.8
End-to-end CAN [11] (TIP17)	63.1	82.9	88.2
Gated CNN [34] (ECCV16)	68.1	88.1	94.6
GAN [25] (ICCV17)	73.1	92.7	96.7
Quadruplet [65] (CVPR17)	74.5	96.6	98.9
DGDNet [17] (CVPR16)	75.3	–	–
MuDeep [66] (ICCV17)	75.3	94.3	97.4
PDC [44] (ICCV17)	78.3	94.8	97.2
JLML [43] (IJCAI17)	80.6	96.9	98.7
DPFL [67] (ICCVW17)	82.0	–	–
CNN-Embedding [51] (TOMM17)	83.4	97.1	98.7
CNN-Embedding [51] + GAN [25]	84.6	97.6	98.9
CPC [13] (CVPR18)	85.9	–	98.5
Our method	88.3	98.3	99.2
Our method + RE [61]	89.2	98.7	99.3

Table 2
Performance comparison on CUHK03 detected dataset (test=700).

Methods	Top1	Top5	Top10
BoW [24] + KISSME [5]	6.4	–	–
LOMO [1] (CVPR15)	12.8	36.7	50.2
IDE(C) [60] (CVPR17)	12.8	–	–
IDE(C) [60] + XQDA [1]	21.1	–	–
IDE(R) [60] (CVPR17)	21.3	–	–
IDE(R) [60] + XQDA [1]	31.1	–	–
DPFL [67] (ICCVW17)	40.7	–	–
SVDNet [46] (ICCV17)	41.5	–	–
Our method	48.2	76.3	83.2
Our method + RE [61]	49.3	76.9	84.0

Table 3
Performance comparison on CUHK01 dataset.

Methods	Top1	Top5	Top10
SDALF [2] (ICCV10)	9.9	41.2	56.0
ITML [33] (ICML07)	15.9	35.2	45.6
LMNN [63] (NIPS05)	13.5	31.3	42.3
KISSME [5] (CVPR12)	13.5	32.0	42.9
LFDA [68] (CVPR13)	15.2	35.8	47.4
eSDC [64] (CVPR13)	19.7	32.7	40.3
KLFDA [27] (ECCV14)	26.6	50.6	62.3
GOG [3] (CVPR16)	57.8	79.1	86.2
IDLA [7] (CVPR15)	47.5	71.6	80.2
DeepRanking [12] (TIP16)	50.4	75.9	84.1
MCP-CNN [14] (CVPR16)	53.7	84.3	91.0
Quadruplet [65] (CVPR17)	62.6	83.4	89.7
DGDNet [17] (CVPR16)	66.6	–	–
Our method	70.2	85.7	92.2
Our method + RE [61]	71.5	86.5	92.5

Table 4
Performance comparison on Market1501 dataset.

Methods	Top1	Top5	mAP
SDALF [2] (ICCV10)	20.5	–	8.2
eSDC [64] (CVPR13)	33.5	–	13.5
BoW [24] (ICCV16)	34.4	–	14.1
PersonNet [9] (ArXiv16)	37.2	–	18.6
End-to-end CAN [11] (TIP17)	48.2	–	24.4
Siamese LSTM [15] (ECCV16)	61.6	–	35.3
Gated CNN [34] (ECCV16)	76.0	–	48.5
SpindleNet [18] (CVPR17)	76.9	91.5	–
GAN [25] (ICCV17)	78.1	–	56.2
PIE [19] (arXiv17)	78.7	90.3	53.9
CNN-Embedding [51] (TOMM17)	79.1	90.9	59.5
CNN-Embedding [51] + GAN [25]	84.0	93.8	66.1
PDC [44] (ICCV17)	84.4	92.7	63.4
JLML [43] (IJCAI17)	85.1	–	65.5
DCCs [50] (ArXiv2018)	86.7	95.7	69.4
PSE [69] (CVPR18)	87.7	–	69.0
Cam-GAN [38,39] (CVPR18)	88.1	–	68.7
PN-GAN [42] (ECCV18)	89.4	–	72.6
Our method	84.7	93.6	65.8
Our method + RE [61]	85.9	94.8	67.2

Therefore, our method achieves better performance than CNN-Embedding [51], with 4.9% accuracy improvement at Top1. Although the Top1 accuracy of CNN-Embedding [51] can be further improved to 84.6% by augmenting training images using GAN [25], our method still performs better than the combination of CNN-Embedding [51] and GAN [25], with 3.7% absolute improvement. Furthermore, combining our method and RE [61] can obtain a rather high Top1 accuracy of 89.2%. From Table 2, we can see that the presented method consistently outperforms all the compared methods with a large margin on the CUHK03 new protocol [60], which further demonstrates the robustness of the learned features.

Performance on CUHK01. The CUHK01 dataset is relatively a small dataset with only 485 identities for training and the deep model is prone to over-fitting on this dataset. We pre-train the deep CNN on a larger dataset CUHK03 and fine-tune the parameters on CUHK01. Table 3 shows the performance comparison between our method and state-of-the-art person re-id methods. It can be observed that our method consistently outperforms the methods using hand-crafted features. For instance, the Top1 accuracy of our method is 12.4% better than the best-performing hand-crafted feature GOG [3]. Compared to other deep methods, our method improves the Top1 matching rate by 7.6%, 16.5% and 19.8% over Quadruplet [65], MCP-CNN [14] and DeepRanking [12], respectively. It is worth noting that DGDNet [17] combines all current person re-id datasets together to form the training set, in which the number of training images is much larger than ours. Even though, our method obtains 70.2% Top1 matching rate, which is 3.6% better than DGDNet [17]. The Top1 accuracy of our method can be further improved to 71.5% by integrating RE [61].

Performance on Market1501. We compare our method with state-of-the-art methods on Market-1501 in Table 4. It can be observed that the deep methods generally performs better than the methods using hand-crafted features, illustrating the powerful feature learning ability of deep networks. Our method performs significantly better than End-to-end CAN [11], Siamese LSTM [15] and Gated CNN [34], with Top1 accuracy improvement of 36.5%, 23.1% and 8.7%, respectively. The work in SpindleNet [18], PIE [19] and PSE [69] are close to ours, which explores informative regions for robust features. But our method needs no extra pose datasets (e.g., MPII human pose dataset [73] in [18,19]) or manual operation (e.g., cropping estimated parts in [18]), which is much efficient. Furthermore, our method improves the Top1 accuracy by 6.0% over PIE [19] and 7.8% over SpindleNet [18]. The model architecture in our method is similar to CNN-Embedding [51], in which verification and recognition model are integrated to learn discriminative feature representations. It is clear that our method achieves 84.7% Top1 accuracy, which leads to 5.6% performance improvement over CNN-Embedding [51]. Many methods including DCCs [50], PSE [69], Cam-GAN [38,39], and PN-GAN [42] obtain high Top1 accuracy and mAP on this dataset, which are better than our method and its combination with RE [61]. But they essentially depend on

Table 5
Performance comparison on DukeMTMC dataset.

Methods	Top1	Top5	mAP
BoW[24] + KISSME [5]	25.1	–	12.2
LOMO [1] (CVPR15)	30.8	–	17.0
GAN [25] (ICCV17)	67.7	–	47.1
OIM Loss [70] (CVPR17)	68.1	–	–
ACRN [71] (CVPRW17)	72.6	84.8	52.0
PN-GAN [42] (ECCV18)	73.6	–	53.2
Cam-GAN [38,39] (CVPR18)	75.3	–	53.5
SVDNet [46] (ICCV17)	76.7	86.4	56.8
PSE [69] (CVPR18)	79.8	89.7	62.0
Our method	76.8	87.5	60.2
Our method + RE [61]	78.2	88.6	61.3

external resources, such as camera information [38,39] and human pose attributes [42,69].

Performance on DukeMTMC. The experimental results are shown in Table 5. It can be seen that our method outperforms most of the compared methods. For instance, the Top1 accuracy of our method is 4.2%, 3.2%, 1.5% and 0.1% better than ACRN [71], PN-GAN [42], Cam-GAN [38,39] and SVDNet [46], respectively. In terms of Top1, Top5 accuracy and mAP, PSE [69] obtains the best performance on this dataset. The Top1 matching rate of PSE [69] is 3.0% better than our method. But, the combination of RE [61] and our method achieve 78.2% Top1 accuracy, which is slightly worse than PSE [69] using auxiliary human pose cues and re-ranking scheme.

4.3. Model Analysis

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

Effectiveness of multilevel feature fusion. We improve the ResNet-50 backbone by integrating multi-layered information in a Feature Aggregation Network (FAN). As shown in Table 6, we report the experimental results on two extra backbone networks, namely AlexNet [74], VGG-19 [75]. Note that, all the fully-connected layers are removed and the input scales are same. The last three convolutional outputs from AlexNet and the last three max-pooling outputs from VGG-19 are sent to FAN, respectively. Here, we optimize the verification task without diversity regularizer for fair comparison. It can be observed that the performance of both verification and recognition loss can be significantly improved by using deeper network. Moreover, fusing three layers obtains the best performance, either for different losses or different backbone networks. This demonstrates the effectiveness of multilevel feature fusion.

Effectiveness of diversity regularizer. Since the diversity regularizer deeply interacts with verification task, here we train the verification loss individually based on FAN, namely $\eta = 0$ in Eq. (9). In Fig. 5(a), we show the performance of the diversity regularizer under different balance weight ζ from 0 to 10^6. It can be observed that, as ζ increases, the Top1 accuracy increases from 78.2% at $\zeta = 0$ to its best which is 81.4% at $\zeta = 5 \times 10^2$. Therefore, we use this value to conduct all the experiments. However, the person re-id accuracy dramatically drops when ζ increases from 5×10^3. Particularly, the Top1 accuracy decreases to 60.5% at $\zeta = 10^6$, which is 17.7% worse than that at $\zeta = 0$, namely without diversity regularizer. This reveals that a proper balance weight is critical for the diversity regularizer, otherwise it may worsen performance. In Fig. 6, we visualize the location masks learned from the comparative network. We can see that either the forward or the backward RCN can learn to focus on different regions of person images at the respective odd and even time steps. It is apparent that the body parts are more salient than the background, which illustrates that the body regions are more beneficial to re-identify different persons. The person appendixes (e.g., backpacks) are also attended, which means they can assist matching persons. Our RCN exhibits some complementary properties. For instance, the forward RCN mainly focuses on the upper person body while the backward RCN pays more attention to the lower body.

Effectiveness of multi-loss learning. We explore the interactions of verification and recognition loss on feature learning, by varying η from 0 to $+\infty$. Here, the experiments are conducted based on FAN and the diversity regularizer is added to the verification loss. The experimental results are shown in Fig. 5(b). When $\eta = 0$, only the verification task takes effect, whose accuracy is 81.4%. As η increases, the verification and recognition loss are unified to supervise the parameters optimization of deep model. The Top1 matching rate reaches its best which is 88.3% at $\eta = 10^1$. When η continues to increase, the recognition loss gradually dominates the training process. As η increases from 10^4 to infinitely large, the verification signal essentially vanishes and only recognition loss remains. Meanwhile, the accuracy decreases and finally reaches 76.7%. It is worth noting that the performance of multi-loss learning remains relatively stable across a wide range of balance coefficients (η from 10^{-1} to 10^3). We can conclude that combining two tasks can significantly improve the performance of deep features in most cases. Moreover, as shown in Table 7, multi-loss learning performs generally better than single loss on four datasets, which further demonstrates the effectiveness of multi-loss learning.

Comparison with other verification loss. In terms of verification task, there exist a variety of loss functions, which include contrastive loss [15,34] and triplet loss [11]. We perform comparison experiments by applying them into the same FAN. Regarding contrastive loss and triplet loss, similar to [11] and RCN that learn attention-aware features, the FAN outputs are forward into attention LSTM, where the time-steps are 6 and the hidden states are concatenated, and then normalized as the final representations. Online hard mining strategy [76] is adopted. From the results in Table 7, it can be observed that the proposed RCN, which learns a joint representation of image pair, outperforms triplet loss method, which performs better than the contrastive loss method. It is worth noting that recognition loss obtains better performance than contrastive loss on some large-scale datasets such as Market1501 and DukeMTMC, probably because there are a rich amount of training samples for recognition loss to combat over-fitting.

Running time analysis. In Table 8, we compare the forward time (FT) and average retrieval time (ART) per image of our method with three other existing methods, namely IDE(R) [60], CNN-Embedding [51] and IDLA [7]. For fair comparison, we re-implement their feature extraction codes using their reported

Table 6
Top1 accuracy of different layers in FAN on CUHK03 dataset (test=100). "Ver" and "Rec" denote verification and recognition loss, respectively. "L3", "L23" and "L123" represent last layer, last two layers and last three layers, respectively.

Methods	AlexNet			VGG-19			ResNet-50		
	L3	L23	L123	L3	L23	L123	L3	L23	L123
Ver	64.1	66.8	67.5	70.2	72.5	73.3	75.6	77.3	78.2
Rec	53.4	56.4	57.6	68.8	70.9	72.4	73.2	75.8	76.7

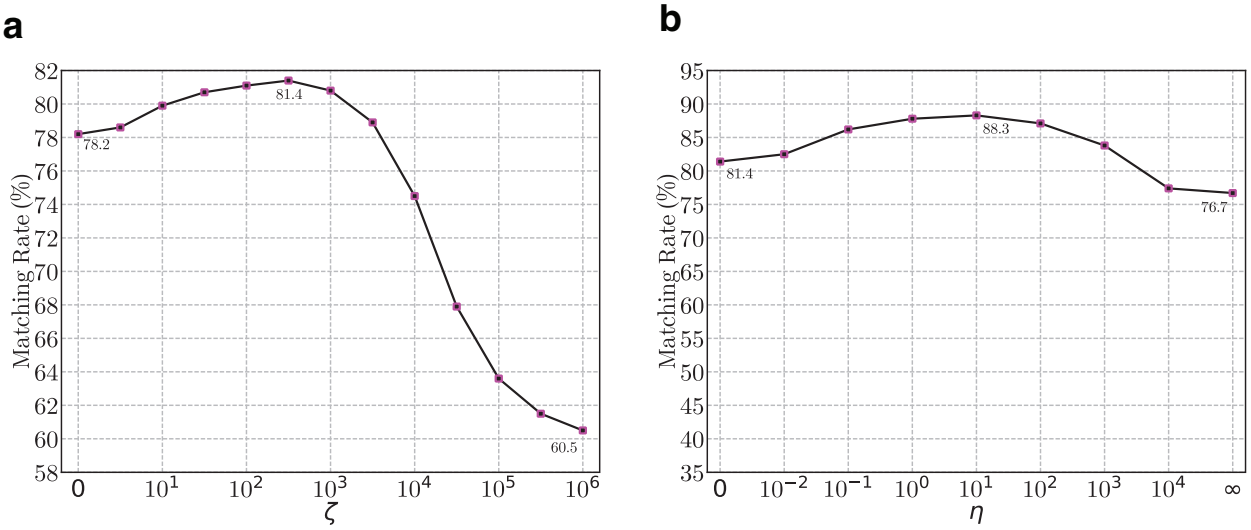


Fig. 5. Experimental results on CUHK03 dataset (test=100). (a) shows the Top1 matching rate with different ζ. (b) shows the Top1 matching rate with different η.

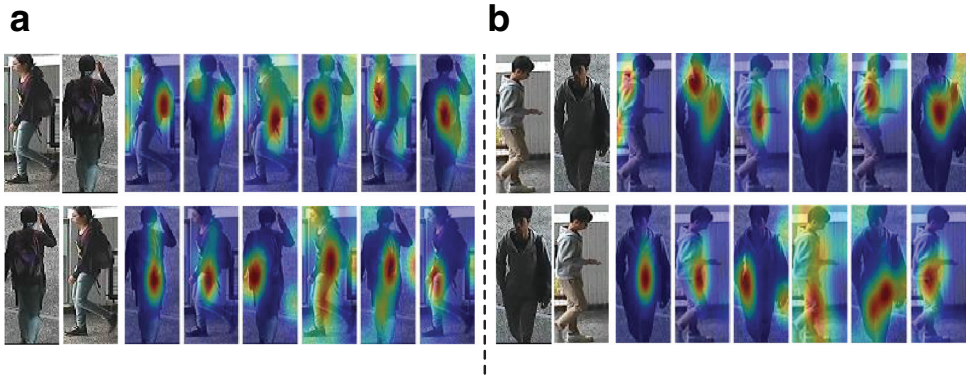


Fig. 6. Samples of the location masks learned from the bi-directional RCN. (a)(b) show two images pairs, respectively. For each image pair, the first row shows the forward location masks while the second row shows the backward location masks.

Table 7
Top1 accuracy on four datasets.

Methods	CUHK03		CUHK01	Market1501	DukeMTMC
	test=100	test=700			
ContrastiveLoss (Ver)	75.3	40.7	60.5	74.6	66.4
TripletLoss (Ver)	77.2	41.4	63.8	78.2	70.5
RCN (Ver)	81.4	44.6	67.7	80.4	73.2
FAN_{GAP} (Rec)	76.7	37.5	51.6	76.2	68.7
RCN+FAN_{GAP}	88.3	48.2	70.2	84.7	76.8

Table 8
Comparison of forward time (FT) and average retrieval time (ART) on Market1501 (partition by "/", millisecond per image).

Methods	Feature	Distance	Ranking	Total
IDE(R) [60]	5.7/5.8	0.03/0.03	–/6.3	5.73/12.13
CNN-Embedding [51]	5.9/5.9	0.02/0.04	–/6.4	5.92/12.34
IDLA [7]	34.5/112.4	0.08/0.45	–/6.3	34.58/119.15
Our method	6.8/23.1	0.07/0.37	–/6.3	6.87/29.77

settings based on PyTorch and test them on large-scale dataset Market1501, which includes 3368 query images and 13,115 gallery images. The total time contains three aspects: (1) feature embedding, (2) distance calculation (matching probability) and (3) ranking (only for ART). All the experiments are conducted on a machine with Intel Xeon CPU (64G memory) and four Titan GPUs (48G memory in total). In terms of FT, it can be seen that IDE(R) [60], CNN-Embedding [51] and our method exhibit high computation efficiency compared to IDLA [7]. Regarding ART, both our method and IDLA [7] require more time than IDE(R) [60] and CNN-Embedding [51]. During retrieval phase in our implementations, a batch is composed of 4000 images. Our method and IDLA [7] have to load 2000 query images first, and then, load the gallery images with separate steps, each includes 2000 images. Therefore, the total number of batch in our method is 14. For the methods such as IDE(R) [60] and CNN-Embedding [51], they use feature vectors that can be saved in buffer for similarity computation and they only need to forward all the images once. In this case, the total number of batch is 5. This is the reason why our method needs more time in ART. The complicated increment of our method in ART will remarkably decrease or even disappear on smaller datasets (e.g., CUHK01 and CUHK03).

5. Conclusion

In this paper, we present a hybrid deep model that combines multilevel feature extraction and multi-loss learning for person re-identification. Both the verification and recognition loss are combined to supervise the parameters optimization of CNN architecture, and meanwhile, learn robust pedestrian descriptors. In particular, compared to the base CNN backbones, the Feature Aggregation Network (FAN) that fuses multilevel information can further improve the performance of re-id, either trained using verification loss or recognition loss. The presented Recurrent Comparative Network (RCN) for verification task can focus on different regions with an attention diversity regularizer, and alternatively compare the appearance similarity between paired images using a recurrent structure. We conduct extensive experiments on four benchmark datasets including CUHK03, CUHK01, Market1501, and DukeMTMC. Experimental results demonstrate that our method outperforms state-of-the-art approaches in most cases.

Acknowledgments

This study is partially supported by the National Natural Science Foundation of China under Grant 61673274.

References

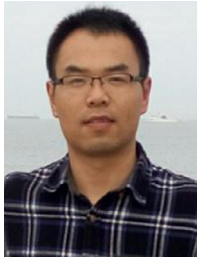
[1] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[3] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1363–1372.
[4] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 653–668.
[5] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[6] A. Mignon, F. Jurie, Pcca: a new approach for distance learning from sparse pairwise constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2666–2672.
[7] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[8] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1288–1296.
[9] L. Wu, C. Shen, A.v.d. Hengel, PersonNet: person re-identification with deep convolutional neural networks, arXiv:1601.07255 (2016).
[10] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[11] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. 26 (7) (2017) 3492–3506.
[12] S.-Z. Chen, C.-C. Guo, J.-H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.
[13] W. Yicheng, C. Zhenzhong, W. Feng, W. Gang, Person re-identification with cascaded pairwise convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1470–1478.
[14] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[15] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 135–153.
[16] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: a study against large variations, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 732–748.
[17] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[18] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 907–915.
[19] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification, arXiv:1701.07732 (2017).
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[21] S. Sharma, R. Kiros, R. Salakhutdinov, Action recognition using visual attention, arXiv:1511.04119 (2015).
[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 2048–2057.
[23] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 31–44.
[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[25] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by Gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3774–3782.
[26] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S.Z. Li, Salient color names for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 536–551.
[27] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 1–16.
[28] S. Khamis, C.H. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 134–146.
[29] W. Li, X. Wang, Locally aligned feature transforms across views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.
[30] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (1) (2009) 207–244.
[31] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 262–275.
[32] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634.
[33] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the International Conference on Machine Learning, 2007, pp. 209–216.
[34] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 791–808.
[35] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Proceedings of the International Conference on Pattern Recognition, 2014, pp. 34–39.
[36] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.
[37] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: Proceedings of the International Conference on Learning Representations, 2016.
[38] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.
[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: a novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019) 1176–1190.
[40] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.
[41] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero- and homogeneously, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 172–188.
[42] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 661–678.
[43] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: Proceedings of the International Joint Conferences on Artificial Intelligence, 2017, pp. 2194–2200.
[44] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.
[45] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Proceedings of the Neural Information Processing Systems, 2015, pp. 2017–2025.
[46] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3820–3828.

[47] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
[48] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3242–3250.
[49] S. Pranav, G. Shubham, D. Ambedkar, Attentive recurrent comparators, arXiv:1703.00767(2017).
[50] L. Wu, Y. Wang, J. Gao, D. Tao, Deep co-attention based comparators for relative representation learning in person re-identification, arXiv:1804.11027 (2018).
[51] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person re-identification, ACM Trans. Multim. Comput. Commun. Appl. 14 (1) (2017).
[52] N. Mclaughlin, J.M.D. Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1325–1334.
[53] W. Zhang, X. Yu, X. He, Learning bidirectional temporal cues for video-based person re-identification, IEEE Trans. Circuits Syst. Video Technol. (99) (2017). pp. 1–1.
[54] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, Y. Chen, Ron: Reverse connection with objectness prior networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5244–5252.
[55] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[56] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proc. IEEE 78 (10) (1990) 1550–1560.
[57] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[58] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 17–35.
[59] H. Moon, P.J. Phillips, Computational and performance aspects of PCA-based face-recognition algorithms., Perception 30 (3) (2001) 303–321.
[60] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.
[61] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, arXiv:1708.04896 (2017).
[62] L. Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 421–436.
[63] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 1473–1480.
[64] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[65] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: A deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[66] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5399–5408.
[67] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: Proceedings of the IEEE International Conference on Computer Vision Workshop, 2018, pp. 2590–2600.
[68] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.
[69] M.S. Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: Proceedings of the IEEE International Conference on Computer Vision, 2018, pp. 420–429.
[70] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3376–3385.
[71] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1435–1443.
[72] M. Hirzer, P.M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 203–208.
[73] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[74] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Neural Information Processing Systems, 2012, pp. 1097–1105.
[75] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015.
[76] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

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.