IEEE Transactions on Image Processing Volume: 28, Issue: 9 , Year 2019

Discriminative Feature Learning with Foreground Attention for Person Re-identification

Sanping Zhou, Jinjun Wang, Deyu Meng, Yudong Liang, Yihong Gong, Nanning Zheng

***Abstract*—The performance of person re-identification (Re-ID) has been seriously effected by the large cross-view appearance variations caused by mutual occlusions and background clutters. Hence learning a feature representation that can adaptively emphasize the foreground persons becomes very critical to solve the person Re-ID problem. In this paper, we propose a simple yet effective foreground attentive neural network (FANN) to learn a discriminative feature representation for person Re-ID, which can adaptively enhance the positive side of foreground and weaken the negative side of background. Specifically, a novel foreground attentive subnetwork is designed to drive the network's attention, in which a decoder network is used to reconstruct the binary mask by using a novel local regression loss function, and an encoder network is regularized by the decoder network to focus its attention on the foreground persons. The resulting feature maps of encoder network are further fed into the body part subnetwork and feature fusion subnetwork to learn discriminative features. Besides, a novel symmetric triplet loss function is introduced to supervise feature learning, in which the intra-class distance is minimized and the inter-class distance is maximized in each triplet unit, simultaneously. Training our FANN in a multi-task learning framework, a discriminative feature representation can be learned to find out the matched reference to each probe among various candidates in the gallery. Extensive experimental results on several public benchmark datasets are evaluated, which have shown clear improvements of our method over the state-of-the-art approaches.**

***Index Terms*—Person Re-identification, Convolutional Neural Network (CNN), Foreground Attentive Feature Learning.**

I. Introduction

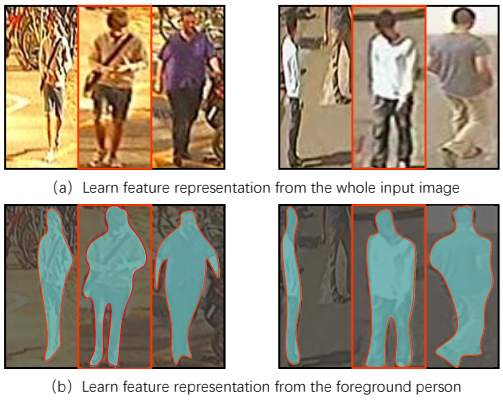
PERSON re-identification (Re-ID) is an important task for many surveillance applications such as person association [1], multi-target tracking [2] and behavior analysis [3]. Given a pedestrian image from one camera view, it tries to find out the stated person amongst a set of gallery candidates captured from the disjoint camera views. The person Re-ID problem has attracted extensive research attentions in recent years, and yet it still remains a challenging one due to the large cross-view appearance variations caused by mutual occlusions and background clutters. Therefore, the key to improve the person Re-ID performance is to learn a discriminative feature representation which is robust to the large cross-view appearance variations.

Sanping Zhou, Jinjun Wang, Yihong Gong and Nanning Zheng are with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

Deyu Meng is with School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

Yudong Liang is with School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi, China.

Corresponding author: Jinjun Wang. Email: jinjun@mail.xjtu.edu.cn.



(a) Learn feature representation from the whole input image

(b) Learn feature representation from the foreground person

Fig. 1. Motivation of our method: we aim to learn a discriminative feature representation which can address the foreground of each input image. In particular, row (a) shows that the general way learns feature from the whole image, and row (b) shows that our method learns feature only from the foreground person.

To address this problem, extensive works have been reported in the past few years, which could be roughly divided into the following two categories: 1) developing discriminative descriptors to handle the variations of person's appearance, and 2) designing distinctive distance metrics to measure the similarity between images. In the first line of works, different informative feature descriptors have been attempted by utilizing different clues, including the LBP [4], ELF [5] and LOMO [6]. In the second line of works, labeled images are used to learn effective distance metrics, including the LADF [7], LMNN [8] and ITML [9]. An evident drawback of these methods is that they consider feature extraction and metric learning as two independent steps, and therefore they cannot complement their capabilities in a joint framework.

Benefit from the strong representation capacity of deep neural network, the deep feature learning based methods [10], [11] have significantly improved the person Re-ID results on the public benchmark datasets. These methods are usually consisted of two components, *i.e.*, a neural network and an objective function. Specifically, the neural network is built to extract features from input images, and the objective function is designed to guide the training process. Representative deep neural networks include the AlexNet [12], VGGNet [13] and ResNet [14], and representative objective functions include the softmax loss function [12], triplet loss function [15] and contrastive loss function [16]. These works usually take the entire rectangular images as inputs, therefore the extracted features may easily get degenerated by the background noises. In order to solve this problem, several works [17], [18], [19] have been presented to address the foreground persons in feature

learning, which are implemented in two steps: 1) Learn two kinds of features from both the complete and masked images using a multi-path network; and 2) Concatenate the multi-path features and fuse them at the output layers. Because they use a multi-path network to extract features, heavy computations are needed at both the training and testing stages.

In this paper, we incorporate a foreground attentive neural network (FANN) and a symmetric triplet loss function into an end-to-end feature learning framework,[1] so as to learn a discriminative feature representation from the foreground images for person Re-ID. As illustrated in Fig. 1, a detected person in a rectangular image region may easily include background clutters and mutual occlusions from the other objects. If such noises can be attenuated, the extracted features will mainly come from the foreground persons, which are more discriminative and robust to the large cross-view appearance variations. This observation has motivated us to propose a novel multi-task learning framework that can jointly alleviate the side effects of backgrounds and learn the discriminative features from foregrounds. Specifically, a novel FANN is built to focus its attention on the foreground persons, in which each image is first passed through an encoder and decoder network, then the outputs of encoder network are further taken for the discriminative feature learning. The encoder network extracts features from the whole image, and the decoder network reconstructs a binary mask of each foreground person. As a result, the encoder network will gradually focus its attention on the foreground persons with the regularization of decoder network by using a novel local regression loss function. Besides, a novel symmetric triplet loss function is introduced to learn the discriminative features, in which the intra-class distance is minimized and the inter-class distance is maximized in each triplet unit, simultaneously. Training the FANN in an end-to-end manner, the foreground attentive features can be finally learned to distinguish different individuals across the disjoint camera views. Extensive experimental results on the 3DPeS [21], VIPeR [5], CUHK01 [22], CUHK03 [23], Market1501 [24] and DukeMTMC-reID [25] datasets have shown the significant improvements by our method, as compared with the state-of-the-art approaches.

The main contributions of this work can be highlighted as follows:

- We design a simple yet effective FANN to learn robust features for person Re-ID, in which the side effects of background can be naturally attenuated and the useful clues in foreground can be greatly emphasized.
- We build an effective local regression loss function to supervise the foreground mask reconstruction, in which the local information in a small neighborhood is used to smooth the isolated regions in ground truth mask.
- We introduce a novel symmetric triplet loss function to supervise the feature learning, in which the intra-class distance is minimized and the inter-class is maximized in each triplet unit, simultaneously.

[1]Note that we submitted our paper before the available of [17], [18], [19], therefore it can be viewed a co-occurring piece of work. Besides, the symmetric triplet loss function was originally proposed in [20], and this work is an extension of our conference paper.

The rest of our paper is organized as follows: In Section II, we briefly review the related works. Section III introduces our neural network and objective function, followed by a discussion of the learning algorithm in Section IV. Experimental results and ablation studies are presented in Section V. And conclusion comes in Section VI.

II. Related Work

We review three lines of related works, including the metric learning based method, the deep learning based method and the attention learning based method, which are briefly introduced in the following paragraphs.

Metric learning based method. This category of methods aim to find a mapping function from the feature space to distance space, in which distances between images of the same person are closer than those between different identities. For example, Zheng et al. [26] proposed a relative distance learning method from the probabilistic prospective. In [27], Mignon et al. learned a distance metric with the sparse pairwise similarity constraints. Pedagadi et al. [28] utilized the Local Fisher Discriminant Analysis (LFDA) to map the high dimensional features into a more discriminative low dimensional space. In [4], Xiong et al. further extended the LFDA and several other metrics by using the kernel tricks and different regularizers. Nguyen et al. [29] measured the similarity between image pairs through the cosine similarity, which was closely related to the inner product similarity. In [30], Loy et al. casted the person Re-ID problem as an image retrieval task by considering the listwise similarity. Chen et al. [31] proposed a kernel based metric learning method to explore the nonlinear relationship of samples in feature space. In [32], Hirzer et al. learned a discriminative metric by using the relaxed pairwise constraints. These methods try to learn a specific distance metric based on features extracted from the fixed feature descriptors, which could not fully discover the potential of metric learning.

Deep learning based method. This category of methods usually incorporate feature extraction and metric learning into a joint framework, in which a neural network is used to extract features and a distance metric is used to compute losses and back-propagate gradients. For example, Ahmed et al. [10] proposed a novel deep neural network which took the pairwise images as inputs, and output a similarity value indicating whether two input images were the same person or not. In [33], Xiao et al. applied a domain guided dropout algorithm to learn the general features for person Re-ID. Ding et al. [11] introduced a triplet neural network to learn the relative similarity in solving the person Re-ID problem. In [34], Wang et al. proposed an unified triplet and siamese deep architecture, which could jointly extract the single-image and cross-image feature representation. Zhou et al. [35] applied a recurrent neural network to jointly learn the spatial and temporal features from video sequence. In [36], Shen et al. designed a novel group-shuffling random walk network for fully utilizing the affinity information between gallery images in both the training and testing stages. Xiao et al. [37] proposed an unified framework which can jointly

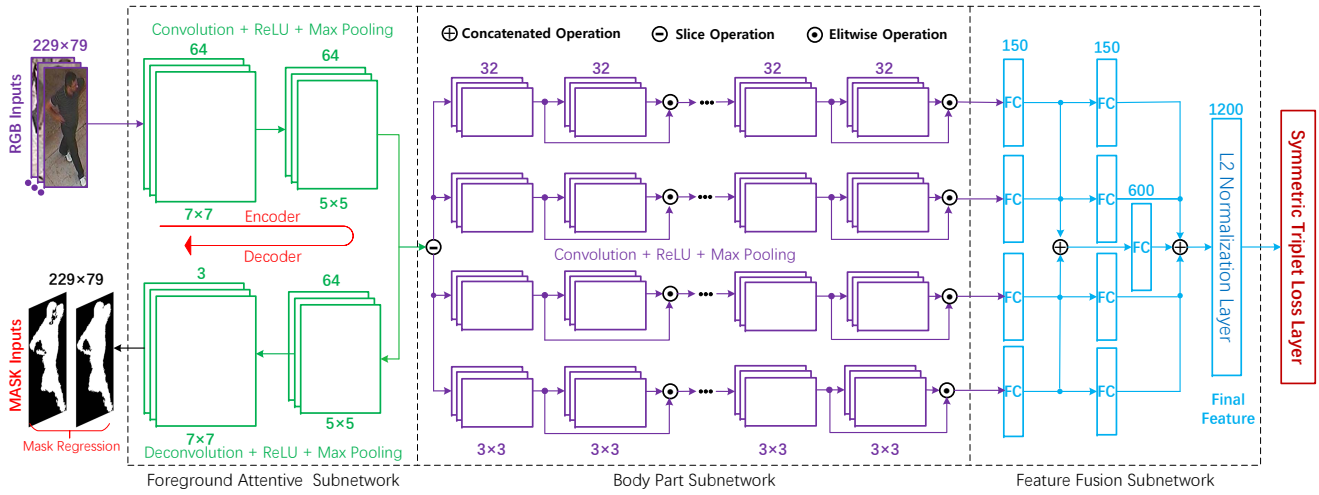


Fig. 2. Illustration of our deep neural network, which is consisted of the foreground attentive subnetwork, the body part subnetwork and the feature fusion subnetwork. Specifically, the foreground attentive subnetwork aims to focus the attention on foreground by passing each input image through an encoder and decoder network. Then, the encoded feature maps are averagely sliced and discriminately learned in the following body part subnetwork. Afterwards, the resulting feature maps are fused in the feature fusion subnetwork. Finally, the final feature vectors are normalized to a unit sphere space and learned by following the symmetric triplet loss layer.

handle the pedestrian detection and person Re-ID in a single network. One major limitation of these methods is that they take the whole image as input, which isn't able to extensively address the foreground persons. Therefore, the learned features will be easily effected by the background noises.

Attention learning based method. This category of methods aim at learning a discriminative feature representation from input images by using different attention mechanisms [38], which can be roughly divided into two categories, *i.e.*, the supervised and unsupervised approaches. In the former ones, the ground truth is needed to supervise the attention learning. For example, Kalayeh et al. [39] took the human parsing results to guide the feature learning for person Re-ID. In [18], Song et al. designed a mask-guided network to drive the network's attention on the foregrounds of input images, which was effective to learn features from the discriminative body regions. Meanwhile, Tian et al. [19] also studied how to alleviate the side effect of background in feature learning. In the latter ones, the attention learning process is usually driven by a specific task or regularizer, which is less effective because no labeled information is available. For example, Wang et al. [40] proposed a residual attention network which embedded an attention mechanism in the network for image classification. In [41], Zhao et al. proposed a part-aligned representation learning method to aggregate the similarities between the corresponding regions of person images. Li et al. [42] designed a harmonious attention network to jointly learn the soft pixel attention and hard regional attention for person Re-ID. In [43], Gheissari et al. introduced a novel spatial-temporal segmentation algorithm to generate the salient regions for person Re-ID. The supervised methods are usually more expensive but effective than the unsupervised ones. In order to pursue higher accuracy, we presented a novel supervised attention learning method to learn discriminative features from the foreground persons, in which the regression task is designed to regularize the feature learning by gradually reconstructing the foreground masks in the training process. Therefore, the complexity of feature extraction network will not be increased in the testing phase, as compared with the existing attention learning based methods in person Re-ID.

III. Multi-task Framework for Foreground Attentive Feature Learning

A. Foreground Attentive Neural Network

The goal of our FANN is to learn a discriminative feature representation from the foregrounds of input images. The proposed network is shown in Fig. 2, which is consisted of the foreground attentive subnetwork, the body part subnetwork and the feature fusion subnetwork. The details are explained in the following paragraphs.

Foreground attentive subnetwork. The foreground attentive subnetwork aims to focus its attention on the foregrounds of input images, so as to alleviate the side effects of backgrounds. Our adopted paradigm is to pass each input image through an encoder and decoder network, in which the encoder network extracts features from the RGB images and the decoder network reconstructs the binary mask of foreground person. The encoder network will naturally focus its attention on the foreground, since the decoder network can gradually reconstruct the binary foreground mask in the learning process. Specifically, the input images are first resized to 229×79 and passed through two 64 learned filters in size of 7×7 and 5×5 with strides 3 and 2, respectively. Then, the resulting feature maps are passed through a rectified linear unit (ReLU) and followed by a max pooling kernel in size of 3×3 with stride 1. These layers constitute the encoder network, and the outputs of encoder network are further fed into the decoder network and the body part subnetwork, simultaneously. The decoder network is consisted of two deconvolutional layers, which are with 64 and 3 learned filters in size of 5×5 and 7×7 with strides 2 and 3, respectively. In addition, a rectified linear unit (ReLU) is put between the two layers. The output of decoder network is used to reconstruct the binary mask of

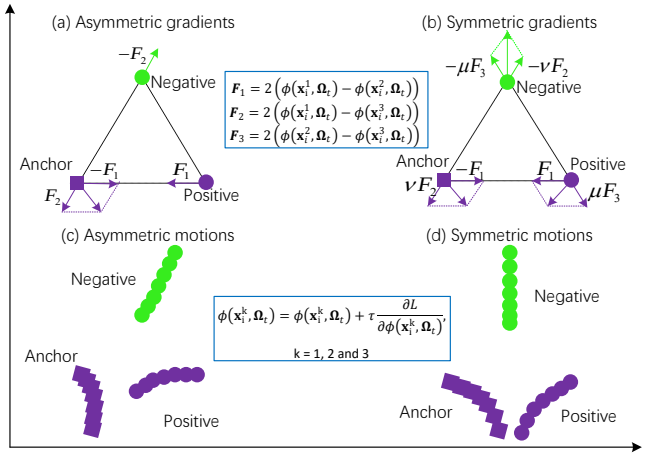


Fig. 3. Illustration of gradient back-propagations and motion trajectories driven by two different triplet loss functions. Specifically, (a) shows the gradients of asymmetric triplet loss function; (b) shows the gradients of symmetric triplet loss function, (c) shows the motion trajectory driven by asymmetric triplet loss function, and (d) shows the motion trajectory driven by symmetric triplet loss function. In the optimization process, we adaptively update u and v, so as to jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit.

foreground person, so as to adaptively drive the attention of encoder network on the foregrounds of input images.

Body part subnetwork. The body part subnetwork aims at learning a discriminative feature representation from different body parts, which is inspired by the idea that different body parts have different weights in representing one person [10]. The resulting feature maps of encoder network are first averagely sliced into four equal parts across the height channel, and then the sliced feature maps are fed into the body part subnetwork for feature learning. The body part subnetwork is consisted of four sets of residual blocks [14], in which these convolutional layers do not share parameters, so as to discriminatively learn feature representations from different body parts. In each residual block, we pass each set of sliced feature maps through two small convolutional layers, in which both of them have 32 learned filters in size of 3×3 with stride 1. The outputs of first small convolutional layer are summarized with the outputs of second small convolutional layer by using the eltwise operation. Then, a rectified linear unit (ReLU) is followed after them. Finally, the resulting feature maps are passed through a max pooling kernel in size of 3×3 with stride 1. In order to enhance the feature representation capacity, we add several residual blocks after the first one, and all of them are in the same shape. Notice that the actual number should be determined by the scale of training dataset.

Feature fusion subnetwork. The feature fusion subnetwork aims to fuse the learned features and normalize them to a unit sphere space. It is consisted of four teams of fully connected layers and a L2 normalization layer. Specifically, the local feature maps of each body part are first discriminately learned by following two small fully connected layers in each team. The dimensions of these small fully connected layers are 150. Then, a rectified linear unit (ReLU) is added between them. Afterwards, the discriminatively learned features of the first four small fully connected layers are concatenated to be fused by following a large fully connected layer, whose dimension is 600. Finally, the resulting feature vectors are further concatenated with the outputs of second four fully connected layers, so as to generate the final 1200 dimensional feature vectors for representation. In addition, a L2 normalization layer is used to regularize the magnitude of each feature vector to be unit. Therefore, the similarity comparison measured in the Euclidian distance is equivalent to that by using the cosine distance.

B. Multi-Task Objective Function

Let $\mathbf{Y} = \{\mathbf{X}_i, \mathbf{M}_i\}_{i=1}^{N}$ be the input training data, in which $\mathbf{X}_i$ denotes the RGB image, $\mathbf{M}_i$ represents the mask of foreground, and N is the number of training samples. Specifically, $\mathbf{X}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \mathbf{x}_i^3\}$ indicates the i^{th} triplet unit, in which $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ are two images with the same identity, $\mathbf{x}_i^1$ and $\mathbf{x}_i^3$ are two mismatched images with different identities. Besides, $\mathbf{M}_i = \{\mathbf{m}_i^1, \mathbf{m}_i^2, \mathbf{m}_i^3\}$ represents the corresponding foreground mask of $\mathbf{X}_i$. The goal of our FANN is to learn filter weights and biases that can jointly minimize the ranking error and the reconstruction error at the output layers, respectively. A recursive function for an M-layer deep model can be defined as follows:

$$\begin{gathered} \mathbf{Y}_i^l = \phi(\mathbf{W}^l * \mathbf{Y}_i^{l-1} + \mathbf{b}^l) \\ i = 1, \cdots, N; l = 1, \cdots, M; \mathbf{Y}_i^0 = \mathbf{Y}_i. \end{gathered}, \tag{1}$$

where $\mathbf{W}^l$ denotes the filter weights of the l^{th} layer, $\mathbf{b}^l$ refers to the corresponding biases, $*$ denotes the convolution operation, $\phi(\cdot)$ is an element-wise non-linear activation function such as ReLU, and $\mathbf{Y}_i^l$ represents the feature maps generated at layer l for $\mathbf{Y}_i$. For simplicity, we consider the deep parameters as a whole $\mathbf{\Omega} = \{\mathbf{W}, \mathbf{b}\}$, in which $\mathbf{W} = \{\mathbf{W}^1, \cdots, \mathbf{W}^M\}$ and $\mathbf{b} = \{\mathbf{b}^1, \cdots, \mathbf{b}^M\}$.

In order to train our FANN in an end-to-end manner, we apply a multi-task objective function to supervise the learning process, which is defined as follows:

$$\begin{aligned} &\min_{\mathbf{\Omega}, \mathbf{u}, \mathbf{v}} \mathrm{E}(\mathbf{\Omega}, \mathbf{u}, \mathbf{v}) = \\ &\sum_{i=1}^{N} \mathrm{L}_1(u_i, v_i, \phi(\mathbf{X}_i, \mathbf{\Omega}_t)) + \zeta \mathrm{L}_2(\phi(\mathbf{X}_i, \mathbf{\Omega}_r), \mathbf{M}_i) + \eta \mathrm{R}(\mathbf{\Omega}), \end{aligned} \tag{2}$$

where $\mathrm{L}_1(\cdot)$ denotes the symmetric triplet loss term, $\mathrm{L}_2(\cdot)$ represents the local regression term, $\mathrm{R}(\cdot)$ indicates the parameter regularization term, and ζ, η are two fixed weight parameters. Specifically, $\mathbf{u} = [u_1, \ldots, u_N]$ and $\mathbf{v} = [v_1, \ldots, v_N]$ are two adaptive weights which control the symmetric gradient back-propagation. Besides, $\mathbf{\Omega} = [\mathbf{\Omega}_t, \mathbf{\Omega}_r]$, in which $\mathbf{\Omega}_t$ is the parameters of deep ranking network and $\mathbf{\Omega}_r$ is the parameters of deep regression network.

Symmetric triplet loss term. The goal of our symmetric triplet loss function is to jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit, so as to learn a discriminative feature representation to correctly match images of each individual captured from the disjoint camera views. Its superiority against the asymmetric triplet loss function [15] is that the deduced gradients to the positive samples are symmetric, as shown in Fig. 3, which is

very essential to consistently minimize the intra-class distance in the training process [2]. The hinge loss of our symmetric triplet loss function is formulated as follows:

$$\mathrm{L}_1 = \max\{\mathrm{M} + \mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^2) - [\mathrm{u}_i \mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3) + \mathrm{v}_i \mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)], 0\}\,, \tag{3}$$

where M is the margin between the positive pair and negative pair, and $\mathrm{d}(\cdot)$ denotes the pairwise distance measured in the unit spherical space, which is defined as follows:

$$\mathrm{d}(\mathbf{x}_i^j, \mathbf{x}_i^k) = \|\phi(\mathbf{x}_i^j, \mathbf{\Omega}_t) - \phi(\mathbf{x}_i^k, \mathbf{\Omega}_t)\|_2^2. \tag{4}$$

In practice, we need to normalize $\|\phi(\mathbf{x}_i^j, \mathbf{\Omega}_t)\|_2^2 = 1$ and $\|\phi(\mathbf{x}_i^k, \mathbf{\Omega}_t)\|_2^2 = 1$, therefore the distance measured in the Euclidean space is equivalent to that measured in the unit spherical space. The smaller the distance $\mathrm{d}(\mathbf{x}_i^j, \mathbf{x}_i^k)$ is, the more similar the two input images $\mathbf{x}_i^j$ and $\mathbf{x}_i^k$ are, and vice versa. Notice that the improved triplet loss function [44] is also lack in ability to deduce the symmetric gradients to positive pairs, because it can't keep $\mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3) \approx \mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)$ in the training process. In the optimization section, we will explain the underling reason in detail.

Local regression loss term. The goal of our local regression loss function is to minimize the reconstruction error at the output of decoder network. As a result, the encoder network will be regularized by the decoder network in reconstructing the binary masks, and the attention of encoder network can be gradually focused on the foreground persons. We measure the reconstruction error of each pixel in a local neighborhood, which is formulated as follows:

$$\mathrm{L}_2 = \sum_{k=1}^{3} \|K_\sigma * (\phi(\mathbf{x}_i^k, \mathbf{\Omega}_r) - \mathbf{m}_i^k)\|_F^2, \tag{5}$$

where K_σ represents a truncated Gaussian kernel with the standard deviation of σ, which is formulated as follows:

$$K_\sigma(x-y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} \exp(-\dfrac{|x-y|^2}{2\sigma^2}), if\ |x-y| \le \rho, \\ 0 \qquad\qquad\qquad , else. \end{cases}, \tag{6}$$

where ρ indicates the radius of local neighborhood O_x which is centered at the point of x. By considering the reconstruction problem in a local neighborhood, the final performance is more robust to the poor mask annotation. As shown in Fig. 4, some pixels in the foreground are wrongly labeled as background, and the reconstruction accuracy will be seriously effected if we reconstruct the foreground mask by only measuring the point to point difference. In our method, we measure the point to set difference by jointly considering the neighborhood information, therefore the foreground mask will be properly reconstructed if most of the pixels in a local neighborhood can be rightly annotated.

[2]The reason of why our symmetric triplet loss function outperforms the asymmetric one is that it can accelerate the motion of positive samples in the vertical direction, as shown in Fig. 3. Therefore, the intra-class distance can be consistently minimized in the training process.

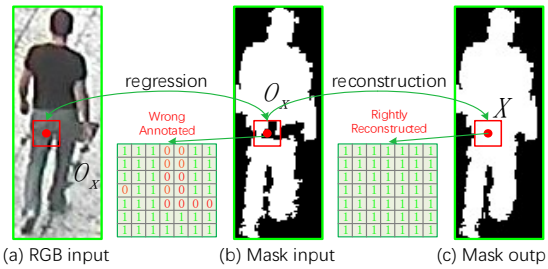


Fig. 4. Illustration of the binary mask reconstruction in a local neighborhood. In practice, some wrongly annotated foreground pixels can be properly rectified by considering the reconstruction in a local neighborhood.

Parameter regularization term. The goal of our parameter regularizer is to smooth the parameters of the entire neural network, which is formulated as follows:

$$\mathrm{R} = \sum_{l=1}^{M} \|\mathbf{W}^l\|_F^2 + \|\mathbf{b}^l\|_2^2\,, \tag{7}$$

where $\|\cdot\|_F^2$ indicates the Frobenius norm, and $\|\cdot\|_2^2$ denotes the Euclidian norm.

IV. Optimization

We apply the momentum method to optimize the direction control weights and the stochastic gradient descent algorithm to optimize the deep parameters, which are introduced in the following paragraphs.

The weight parameters u_i and v_i can be adaptively updated in the training process by using the momentum method, so as to jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. In order to simplify this problem, we define $u_i = \alpha_i + \beta_i$ and $v_i = \alpha_i - \beta_i$, and therefore the two parameters can be optimized by only updating β_i in each iteration. The partial derivative of our symmetric triplet loss function with respect to β_i can be formulated as follows:

$$t = \begin{cases} \frac{\partial \mathrm{T}(\mathbf{x}_i^1, \mathbf{x}_i^2, \mathbf{x}_i^3)}{\partial \beta_i}, & if\ \mathrm{T} > 0\,, \\ 0 \qquad\qquad , & else. \end{cases}, \tag{8}$$

where $\mathrm{T} = \mathrm{M} + \mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^2) - [u_i \mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3) + v_i \mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)]$, and $\frac{\partial \mathrm{T}}{\partial \beta_i}$ is formulated as follows:

$$\frac{\partial \mathrm{T}}{\partial \beta_i} = \|\phi(\mathbf{x}_i^2, \mathbf{\Omega}_t) - \phi(\mathbf{x}_i^3, \mathbf{\Omega}_t)\|_2^2 - \|\phi(\mathbf{x}_i^1, \mathbf{\Omega}_t) - \phi(\mathbf{x}_i^3, \mathbf{\Omega}_t)\|_2^2. \tag{9}$$

Then, β_i can be optimized as follows:

$$\beta_i \leftarrow \beta_i - \gamma \cdot t, \tag{10}$$

where γ is the weight updating rate. It can be clearly seen that when $\|\phi(\mathbf{x}_i^1, \mathbf{\Omega}_t) - \phi(\mathbf{x}_i^3, \mathbf{\Omega}_t)\|_2^2 > \|\phi(\mathbf{x}_i^2, \mathbf{\Omega}_t) - \phi(\mathbf{x}_i^3, \mathbf{\Omega}_t)\|_2^2$, namely $t < 0$, then u_i will be decreased while v_i will be increased; and vice verse. As a result, the strength of back-propagation to samples in each triplet unit will be adaptively tuned, in which the anchor and the positive will be clustered, and the negative one will be far away from the hyper-line expanded by the anchor and the positive. Without this property, the improved triplet loss function [44] can not consistently minimize the intra-class distance in the training process.

In order to apply the stochastic gradient descent algorithm to optimize the deep parameters, we compute the partial derivative of our objective function as follows:

$$\frac{\partial \mathrm{E}}{\partial \mathbf{\Omega}}=\sum_{i=1}^{N}\ell_1(u_i,v_i,\phi(\mathbf{X}_i,\mathbf{\Omega}_t))+\zeta\ell_2(\phi(\mathbf{X}_i,\mathbf{\Omega}_r),\mathbf{M}_i)+\eta\sum_{l=1}^{M}\mathbf{\Omega}^l, \tag{11}$$

where the first term represents the gradient of symmetric triplet loss function, the second term denotes the gradient of local regression loss function, and the third term indicates the gradient of parameter regularizer.

By the definition of $\mathrm{T}(\mathbf{x}_i^1,\mathbf{x}_i^2,\mathbf{x}_i^3)$ in Eq. (8), the gradient of our symmetric triplet loss term can be computed as follows:

$$\ell_1=\begin{cases}\frac{\partial \mathrm{T}(\mathbf{x}_i^1,\mathbf{x}_i^2,\mathbf{x}_i^3)}{\partial \mathbf{\Omega}_t}, & if\ \mathrm{T}>0,\\ 0 & , else.\end{cases}, \tag{12}$$

where $\frac{\partial \mathrm{T}}{\partial \mathbf{\Omega}_t}$ is formulated as follows:

$$\begin{aligned}\frac{\partial \mathrm{T}}{\partial \mathbf{\Omega}_t}&=2(\phi(\mathbf{x}_i^1,\mathbf{\Omega}_t)-\phi(\mathbf{x}_i^2,\mathbf{\Omega}_t))'\frac{\partial\phi(\mathbf{x}_i^1,\mathbf{\Omega}_t)-\partial\phi(\mathbf{x}_i^2,\mathbf{\Omega}_t)}{\partial\mathbf{\Omega}_t}\\ &-2u_i(\phi(\mathbf{x}_i^1,\mathbf{\Omega}_t)-\phi(\mathbf{x}_i^3,\mathbf{\Omega}_t))'\frac{\partial\phi(\mathbf{x}_i^1,\mathbf{\Omega}_t)-\partial\phi(\mathbf{x}_i^3,\mathbf{\Omega}_t)}{\partial\mathbf{\Omega}_t}.\\ &-2v_i(\phi(\mathbf{x}_i^2,\mathbf{\Omega}_t)-\phi(\mathbf{x}_i^3,\mathbf{\Omega}_t))'\frac{\partial\phi(\mathbf{x}_i^2,\mathbf{\Omega}_t)-\partial\phi(\mathbf{x}_i^3,\mathbf{\Omega}_t)}{\partial\mathbf{\Omega}_t}\end{aligned} \tag{13}$$

According to the definition of our local regression loss term in Eq. (5), the gradient can be computed as follows:

$$\ell_2=\sum_{k=1}^{3}2K_\sigma*(K_\sigma*(\phi(\mathbf{x}_i^k,\mathbf{\Omega}_r)-\mathbf{m}_i^k))\frac{\partial\phi(\mathbf{x}_i^k,\mathbf{\Omega}_r)}{\partial\mathbf{\Omega}_r}. \tag{14}$$

It is clear that the gradients of samples can be easily calculated given the values of $\phi(\mathbf{x}_i^k,\mathbf{\Omega}_t)$, $\partial\phi(\mathbf{x}_i^k,\mathbf{\Omega}_t)/\partial\mathbf{\Omega}_t$ and $\phi(\mathbf{x}_i^k,\mathbf{\Omega}_r)$, $\partial\phi(\mathbf{x}_i^k,\mathbf{\Omega}_r)/\partial\mathbf{\Omega}_r$ in each mini-batch, which can be easily obtained by running the forward and backward propagation in the training process. As the algorithm needs to back-propagate the gradients to learn a foreground attentive feature representation, we call it the foreground attentive gradient descent algorithm. Algorithm 1 shows the overall process of our implementation regime.

Algorithm 1 Foreground Attentive Gradient Descent.

Input:

Training data $\mathbf{Y}$, learning rate τ, maximum iterative number H, weight parameters ζ,η, kernel parameters σ,ρ, margin parameter $\mathcal{M}$, initial weights of u_i and v_i and updating rate γ.

Output:

The network parameters $\mathbf{\Omega}=[\mathbf{\Omega}_t,\mathbf{\Omega}_r]$.

repeat

1. Extract the features of $\phi(\mathbf{x}_i^k,\mathbf{\Omega}_t)$ and $\phi(\mathbf{x}_i^k,\mathbf{\Omega}_r)$ in each triplet unit by the forward propagation.

repeat

a) Compute the gradient of $\frac{\partial \mathrm{T}}{\partial\beta_i}$ according to Eq. (9);

b) Update weights u_i and v_i according to Eq. (10);

c) Compute the gradients of ℓ_1 and ℓ_2 according to Eq. (12) and Eq. (14);

d) Update the gradients of $\frac{\partial \mathrm{E}}{\partial\mathbf{\Omega}}$ according to Eq. (11);

until Traverse all the triplet inputs $\{\mathbf{y}_i^1,\mathbf{y}_i^2,\mathbf{y}_i^3\}$ in each min-batch;

2. Update $\mathbf{\Omega}^{(h+1)}=\mathbf{\Omega}^{(h)}-\tau_h\frac{\partial \mathrm{E}}{\partial\mathbf{\Omega}^{(h)}}$ and $h\leftarrow h+1$.

until $h>\mathrm{H}$

V. Experiments

A. Datasets and Settings

Benchmark datasets. We evaluate our method on six datasets, including the 3DPeS [21], VIPeR [5], CUHK01 [22], CUHK03 [23], Market1501 [24] and DukeMTMCre-ID [25], which are briefly introduced in the following paragraphs. [3] The 3DPeS dataset contains 1,011 images of 192 persons captured from 8 outdoor cameras with different viewpoints, and each person has 2 to 26 images. The VIPeR dataset contains 632 person images captured by two cameras in an outdoor environment, and each person has only one image in each camera view. The CUHK01 dataset contains 971 persons captured from two camera views in a campus environment, and there are two images for each person under every camera view. The CUHK03 dataset contains 14,097 images from 1,467 persons, which is captured from six cameras in a campus environment and each person only has two camera views. The Market1501 dataset contains 32,668 images of 1,501 persons in a campus environment, in which each person is captured by six cameras at most, and two cameras at least. The DukeMTMC-reID dataset is consisted of 1,812 identities captured from 8 different cameras, in which 16,522 samples from 702 identities are used for training, 2,228 samples of another 702 identities are used as queries, and the remaining 17,661 samples are used for the gallery set.

Parameter settings. The parameters are taken as follows: The weights are initialized from two zero-mean Gaussian distributions with the standard deviations of 0.01 to 0.001, and the bias terms are set as 0. The learning rate $\omega=0.01$, and decayed by 0.1 at every 10,000 iterations, the margin parameter $\mathrm{M}=0.1$, the kernel parameters $\rho=3,\sigma=0.01$, the weight parameters $\zeta=0.02$ and $\eta=0.05$, the initial adaptive weights $u=0.6$ and $v=0.4$, and the weight updating rate $\gamma=0.01$. If not specified, we use the same parameters in all the experiments.

Evaluation protocol. Our experiments use the Cumulative Matching Characteristic (CMC) curve to measure the performance, which is an estimation of finding the corrected top n match. For the 3DPeS and VIPeR datasets, we follow the single-shot protocol in [11], in which 96 persons from the 3DPeS dataset and 316 persons in the VIPeR dataset are randomly chosen to train the deep neural network, and the remaining identities are used to evaluate the performance. For the CUHK01 and CUHK03 datasets, we follow two data partition protocols to split the datasets into the training sets and testing sets. Specifically, 100/486 persons of the CUHK01 dataset and 100/700 persons of the CUHK03 dataset are used to evaluate the performance, and the remainings are

[3]The 3DPeS dataset provides the foreground masks, and the foreground masks of images in other datasets are obtained by using the algorithm [45] in link http://www.robots.ox.ac.uk/~szheng/CRFasRNN.html.

TABLE I
THE MATCHING RATES(%) COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CUHK01 AND CUHK03 DATASETS, IN WHICH '-' MEANS THEY DO NOT REPORT THE CORRESPONDING RESULT.

Methods	Year	CUHK01 (p=100)			CUHK01 (p=486)			CUHK03 (p=100)			CUHK03 (p=700)		
		Top 1	Top 5	Top10	Top 1	Top 5	Top10	Top 1	Top 5	Top10	Top 1	Top 5	mAP
kLFDA [4]	2014	42.7	69.0	79.6	32.7	59.0	69.6	48.2	59.3	66.4	-	-	-
LOMO+XQDA [46]	2015	77.6	94.1	97.5	63.2	83.9	90.0	52.0	82.2	92.1	14.8	-	13.6
IDLA [10]	2015	65.0	89.5	93.0	47.5	71.5	80.0	54.7	86.5	94.0	-	-	-
ITML [9]	2017	17.1	42.3	55.1	16.0	35.2	45.6	5.5	18.9	30.0	-	-	-
SVDNet [47]	2017	-	-	-	-	-	-	-	95.2	97.2	40.9	-	37.8
PAN [48]	2017	-	-	-	-	-	-	-	-	-	36.9	56.9	35.0
Quadruplet [49]	2017	79.0	96.0	97.0	62.6	83.0	88.8	74.5	96.6	99.0	-	-	-
DPFL [50]	2017	-	-	-	-	-	-	-	-	-	43.0	-	40.5
MLFN [51]	2018	-	-	-	-	-	-	82.8	-	-	54.7	-	49.2
PRGP [19]	2018	-	-	-	80.7	95.0	97.5	91.7	98.2	98.7	-	-	-
MLS [52]	2018	88.2	98.2	99.4	-	-	-	87.5	97.9	99.5	-	-	-
HA-CAN [42]	2018	-	-	-	-	-	-	-	-	-	44.4	-	41.0
DGRW [36]	2018	-	-	-	-	-	-	**94.9**	98.7	99.3	-	-	-
MGCAN [18]	2018	-	-	-	-	-	-	-	-	-	50.1	-	50.2
BraidNet [53]	2018	93.0	-	99.9	-	-	-	88.2	-	98.7	-	-	-
AACN [54]	2018	88.1	96.7	98.2	-	-	-	91.4	98.9	99.5	-	-	-
PN-GAN [55]	2018	-	-	-	67.7	86.6	91.8	79.8	96.2	98.6	-	-	-
DaRe [56]	2018	-	-	-	-	-	-	-	-	-	66.1	-	66.7
Our FANN	2018	**98.1**	**99.8**	**100**	**81.2**	**95.3**	**99.1**	92.3	**99.2**	**100**	**70.2**	**86.1**	**70.4**

TABLE II
THE MATCHING RATES(%) COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE 3DPeS DATASET, IN WHICH '-' MEANS THEY DO NOT REPORT THE CORRESPONDING RESULT.

Methods	Year	Top 1	Top 5	Top10	Top15	Top20
KISSME [57]	2012	22.9	48.7	62.2	72.4	78.1
LF [28]	2013	33.4	45.5	69.9	76.5	81.0
kLFDA [4]	2014	54.0	77.7	85.9	90.0	92.4
MFA [4]	2014	41.8	65.5	75.7	-	85.2
ME [58]	2015	53.3	76.8	86.0	89.4	92.8
SCSP [59]	2016	57.3	78.9	85.0	89.5	91.5
JSTL [33]	2016	56.0	-	-	-	-
WARCA [60]	2016	51.9	75.6	-	-	-
Spindle [61]	2017	62.1	83.4	90.5	-	95.7
P2S [20]	2017	71.2	90.5	95.2	96.9	97.6
SPL [62]	2018	72.2	90.7	95.3	96.8	97.5
PRGP [19]	2018	64.1	87.4	90.4	-	93.7
Our FANN	2018	**78.9**	**92.3**	**95.7**	**98.1**	**99.4**

TABLE III
THE MATCHING RATES(%) COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE VIPeR DATASET, IN WHICH '-' MEANS THEY DO NOT REPORT THE CORRESPONDING RESULT.

Methods	Year	Top 1	Top 5	Top10	Top15	Top20
RPLM [32]	2012	27.3	55.3	69.0	77.1	82.7
sLDFV [63]	2012	26.5	56.4	70.9	-	84.6
kBiCov [64]	2015	31.1	58.3	70.7	-	82.4
Triplet [11]	2015	40.5	60.8	70.4	78.4	84.4
LNDS [65]	2016	51.2	82.1	90.5	-	-
Quadruplet [49]	2017	49.1	73.1	81.9	-	-
Spindle [61]	2017	53.8	74.1	83.2	-	92.1
SSM [66]	2017	53.7	-	-	-	96.1
PDC [67]	2017	51.3	74.0	84.2	-	91.5
SPL [62]	2018	56.3	83.0	92.0	93.8	95.9
PRGP [19]	2018	50.6	70.3	79.1	-	88.0
MLS [52]	2018	50.0	73.1	84.4	-	-
Our FANN	2018	**58.4**	**83.7**	**92.2**	**93.9**	**96.4**

used to train the deep neural network. For the Market1501 and DukeMTMC-reID datasets, we used the provided data partition methods to prepare the training and testing samples. Besides, the mean Average Precision (mAP) is also used to evaluate the performance on the CUHK03, Market1501 and DukeMTMC-reID datasets. To obtain a statistical result, we repeated the testing 10 times to report the average result.

B. Comparison Results

Firstly, we will compare our method with the state-of-the-art approaches on the six public benchmark datasets, respectively. Secondly, the performances of attention learning based methods will be solely evaluated, so as to compare how much they can improve the final results. For clarity, we highlight the best results in bold.

Comparisons with the state-of-the-arts. The detailed results are shown in Table II to Table V, from which we can see that our FANN has achieved the competitive results on nearly all of the six public benchmark datasets. Specifically, our FANN outperforms the previous best performed SPL [62] method by 6.7% on the 3DPeS dataset in the Top 1 accuracy. Besides, our FANN also outperforms the previous best performed SPL [62] method by 2.1% on the VIPeR dataset in the Top 1 accuracy. For the CUHK01 and CUHK03 datasets, our FANN outperforms the previous best performed BraidNet [53] by 5.1%, while lags behind the previous best performed DGRW [36] method by 2.6% in the Top 1 accuracy, when 100 identities are randomly chosen to evaluate the performance, respectively. When 486 identities from the CUHK01 dataset are used to evaluate the performance, our FANN outperforms the previous best performed PRGP [19] method by 0.5% in the Top 1 accuracy. In addition, our FANN outperforms the previous best performed DaRe [56] by 4.1% and 3.7% in terms of the Top 1 accuracy and mAP, when 700 identities are used to evaluate the performance on the CUHK03 dataset, respectively. The same conclusion can be get on the Market1501 and DukeMTMC-reID datasets using the single-query evaluation, in which our FANN outperforms the previous best performed GCSL [68] and PCB [69] methods by 0.6%, 0.3% and 0.9%, 0.7% in the Top 1 accuracy and mAP on the Market1501 and DukeMTMC-reID datasets, respectively.

TABLE IV
THE MATCHING RATES(%) IMPROVED BY EACH OF OUR CONTRIBUTIONS ON THE SIX BENCHMARK DATASETS, RESPECTIVELY.

Methods	3DPeS		VIPeR		CUHK01 (100)		CUHK01 (486)		CUHK03 (100)		CUHK03 (700)		Market.		Duke.	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	mAP	Top 1	mAP	Top 1	mAP
Baseline1	67.3	89.2	47.3	73.1	79.8	90.4	64.1	86.1	73.9	92.1	52.6	51.8	67.6	45.4	64.4	43.1
Baseline2	65.4	87.1	43.2	72.2	75.8	86.2	58.5	82.2	68.9	89.1	48.4	47.9	62.2	39.6	60.1	39.6
S	72.5	90.7	50.9	80.8	92.1	95.9	72.3	89.6	81.4	95.2	62.1	62.3	78.4	54.1	72.1	60.1
L	73.1	90.9	51.2	81.1	92.7	96.8	74.4	91.2	83.6	96.4	65.0	63.7	84.6	64.7	77.6	64.2
F	72.1	89.7	50.1	80.2	90.1	95.3	74.1	90.8	82.4	96.1	61.9	60.7	77.9	60.9	71.4	58.4
S + F	75.2	91.8	55.1	81.9	94.2	98.1	76.5	92.5	88.1	97.1	66.7	65.9	86.1	67.6	77.1	63.6
S + L	**78.9**	**92.3**	**58.4**	**83.7**	**98.1**	**99.8**	**81.2**	**95.3**	**92.3**	**99.2**	**70.2**	**69.5**	**94.4**	**82.5**	**85.2**	**70.2**

TABLE V
THE MATCHING RATES(%) COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE MARKET1501 AND DUKEMTMC-REID DATASETS, IN WHICH '-' MEANS THEY DO NOT REPORT THE CORRESPONDING RESULT.

Methods	Year	Market.		Duke.	
		Top 1	mAP	Top 1	mAP
LDNS [65]	2016	61.0	35.6	-	-
S2S [70]	2017	65.3	40.0	-	-
DSPL [62]	2017	72.9	46.7	-	-
JLML [71]	2017	83.9	64.4	-	-
PDC [67]	2017	84.1	63.4	-	-
SVDNet [47]	2017	82.3	62.1	76.7	56.8
SSM [66]	2017	82.2	68.8	-	-
DPFL [50]	2017	88.6	72.6	79.2	60.6
PRGP [19]	2017	81.2	-	-	-
MLFN [51]	2018	90.0	74.3	81.0	62.8
DGRW [36]	2018	92.7	82.5	80.7	66.4
DuATM [72]	2018	91.4	76.6	81.8	64.6
BraidNet [53]	2018	83.7	69.5	76.4	59.5
AACN [54]	2018	85.9	66.9	76.8	59.3
SGGNN [73]	2018	92.3	82.8	81.8	68.2
PN-GAN [55]	2018	89.4	72.6	73.6	53.2
GCSL [68]	2018	93.5	81.6	84.9	69.5
PCB [69]	2018	93.8	81.6	83.3	69.2
Our Method (FANN)	2018	**94.4**	**82.5**	**85.2**	**70.2**

Comparisons of attention learning. The attention learning based methods can usually improve the discriminative ability of learned features in solving the person Re-ID problem, because they can further address the foreground persons in the training process. As discussed above, the supervised methods often outperform the unsupervised ones in the final accuracy. In Fig 5, we compare our method with the other four attention learning based methods on the Market1501 dataset, in which the HSP [39] and MGCAM [18] are the supervised methods, while the DLPA[41] and HA-CAN [42] are the unsupervised methods. From the results, we can see that the worse results are obtained by the DLPA, and the best performances are achieved by our FANN. Besides, we also notice that the HA-CAN significantly outperforms the MGCAM in the Top 1 accuracy, which indicates that it is possible to learn the attentive features in an unsupervised manner. In the future study, we will strive to design an attention mechanism in network, so as to improve the feature representation capability without using the expensive foreground annotations.

C. Ablation Study

Firstly, we will evaluate how much each of our contributions improves the final person Re-ID results. Secondly, the effectiveness of our FANN in background suppression will

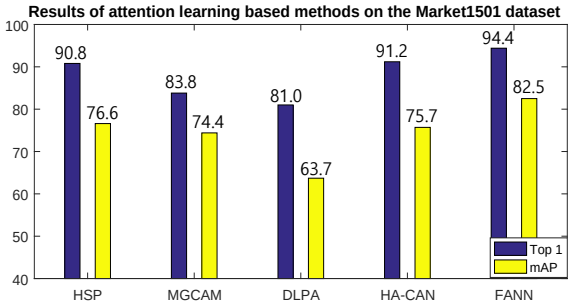


Fig. 5. Comparison between our method and the other attention learning based approaches on the Market1501 dataset, in which the HSP [39], MGCAM [18] and our FANN belong to the supervised approaches, and the DLPA[41] and HA-CAN [42] are the unsupervised methods.

be illustrated, including the visualization of learned feature maps and the robustness of our FANN to different ground-truth masks. Then, we will show the robustness of our method to different parameter settings and study how to set the number of residual blocks in the body part subnetwork. Finally, some ranking examples will be presented and discussed.

Improvements by each contribution. In order to show how much each contribution improves the final results, we carefully design seven different experiments on each dataset, as shown in Table IV. In particular, **Baseline 1** denotes the performances that we get rid of the decoder network and takes the asymmetric triplet loss function to train the remaining network, and **Baseline 2** indicates the results that we use the masked foreground images to replace the inputs in **Baseline 1**. Besides, **S** means the performances that we get rid of the decoder network and takes our symmetric triplet loss function to train the remaining network, **L** represents the results that we use the asymmetric triplet loss function and local regression loss function to train the whole network, and **F** indicates the performances that we use the asymmetric triplet loss function and Euclidean loss function to train the whole network. What's more, **S + F** denotes the results that we use our symmetric triplet loss function and Euclidean loss function to train the whole network, and **S + L** represents the results obtained by jointly using the symmetric triplet loss function and local regression loss function, which is actually equivalent to our FANN method.

From the results we can see that the **S + L** significantly outperforms the other six situations on all the six benchmark datasets, which can well explain the effectiveness of our symmetric triplet loss function, the local regression loss function and the neural network in learning the discriminative

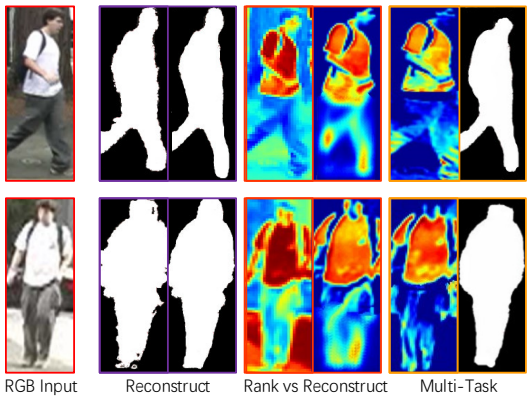


Fig. 6. Illustration of the learned feature maps and masks by our neural network. From left to right: the RGB inputs, the ground truth masks by using the off-the-shelf segmentation method, the foreground masks obtained in the reconstruction task, the feature maps learned in the ranking and reconstruction tasks, the feature maps and foreground masks obtained in the multi-task.

features from the foregrounds of input images. For simplicity, we take the results on the VIPeR dataset to explain the detailed improvements: 1) Compare the performances in **Baseline 1** and **Baseline 2**, we can find that directly feeding the masked images to train the neural network can not improve the person Re-ID results, because the strong edge responses brought by the masks are harmful to learn the discriminative features. 2) Compare the results between **Baseline 1** and **S**, between **F** and **S + F**, between **L** and **S + L**, we can find that our symmetric triplet loss function can improve the Top 1 accuracy by 5.2%, 3.1%, 5.8% in the three cases, respectively. 3) For the improvements of our local regression loss function, we compare the results between **Baseline 1** and **L**, between **F** and **L**, between **S + F** and **S + L**, and the results show that our local regression loss function can improve the Top 1 accuracy by 5.8%, 1.0%, 3.7% in the three cases, respectively. 4) Finally, we evaluate the improvements bought by our neural network by comparing the results between **Baseline 1** and **F**, between **Baseline 1** and **L**, between **S** and **S + F**, which show that our FANN improves the Top 1 accuracy by 4.8%, 5.8%, 2.7% in the three cases, respectively. What's more, the same conclusions can be found if we evaluate the performances on the other five datasets.

Effectiveness in background suppression. In this paragraph, we will explain the internal reason of how our FANN focuses its attention on the foregrounds of input images. In our FANN, we apply the encoder-decoder mechanism to drive the attention, in which we reconstruct the mask of foreground at the output of decoder network, and the encoder network will be naturally regularized by the decoder network in the training process. As a result, the encoder network will pay more attention on the foregrounds of input images, which is effective to suppress noises in the background. Then, the resulting feature maps of encoder network will be fed into the subsequent networks for discriminative feature learning. Incorporating the mask reconstruction and feature learning into a multi-task learning framework, a discriminative feature representation can be learned to further improve the final person Re-ID results.

Some representative feature maps of two input images are shown in Fig. 6, in which the two images represent the same person under two disjoint camera views. Specifically, the second column shows the ground truth masks obtained by the segmentation method, and the third column represents the binary masks reconstructed by only running the reconstruction task. The results indicate that our local regression loss function is effective in reconstructing the binary mask. The fourth and fifth columns illustrate the feature maps of encoder network in the ranking and reconstruction tasks, which indicate that the reconstruction task can focus more attentions on the foreground than the ranking task. The sixth and seventh columns represent the feature maps of encoder network and the reconstructed masks by using the multi-task objective function, which illustrate that running the two tasks in an end-to-end manner is more beneficial to learn the foreground attentive features for person Re-ID.

TABLE VI
THE MATCHING RATE (%) ON OUR METHOD BY USING DIFFERENT GROUND TRUTH MASKS ON THE SIX BENCHMARK DATASETS.

Datasets	Baseline		Mask 1		Mask 2	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
3DPeS	72.5	90.7	76.3	92.0	**78.9**	**92.3**
VIPeR	50.9	80.8	56.4	82.1	**58.4**	**83.7**
CUHK01	92.1	95.9	96.6	99.2	**98.1**	**99.8**
CUHK03	81.4	95.2	91.3	98.7	**92.3**	**99.2**
Market1501	78.4	54.1	92.1	94.9	**94.4**	**95.3**
DukeMTMC	72.1	60.1	83.2	90.8	**85.2**	**91.6**

In addition, our FANN is robust to the quality of generated masks in a certain extent. In order to support this point of view, we take two methods, *i.e.*, one instance segmentation based method [45] and one saliency detection based method [74], to generate the ground truth masks. Some generated masks on the Market1501 dataset are shown in Fig. 8, in which we can see that masks generated by the instance segmentation based method are in higher quality than that by the saliency detection based method. Using the two kinds of masks as ground truth, we evaluate our method on the six benchmark datasets. The results are shown in Table VI, in which the 'Baseline' denotes the result without using the mask information, 'Mask 1' indicates the results by using the masks generated by the saliency detection based method, and 'Mask 2' represents the results by using the masks generated by the instance segmentation based method. From the results, we can conclude that: 1) The two kinds of masks can help our FANN improves the person Re-ID results; 2) The poorer masks will lead to a bit worser results, however we also notice that the differences are not very large.

Robustness to parameter setting. There are several hyper parameters in our method, and they are sensitive to the final person Re-ID results in different extents. In the following paragraphs, we will first evaluate our method with varying parameters: the margin parameter M, the kernel parameters ρ and σ, the weight parameter ζ, and the initial adaptive weights u and v. Specifically, we change one parameter and keep the others fixed in each experiment, so as to illustrate the sensitiveness of method to each parameter. Then, we further study how to set the number of residual blocks in the

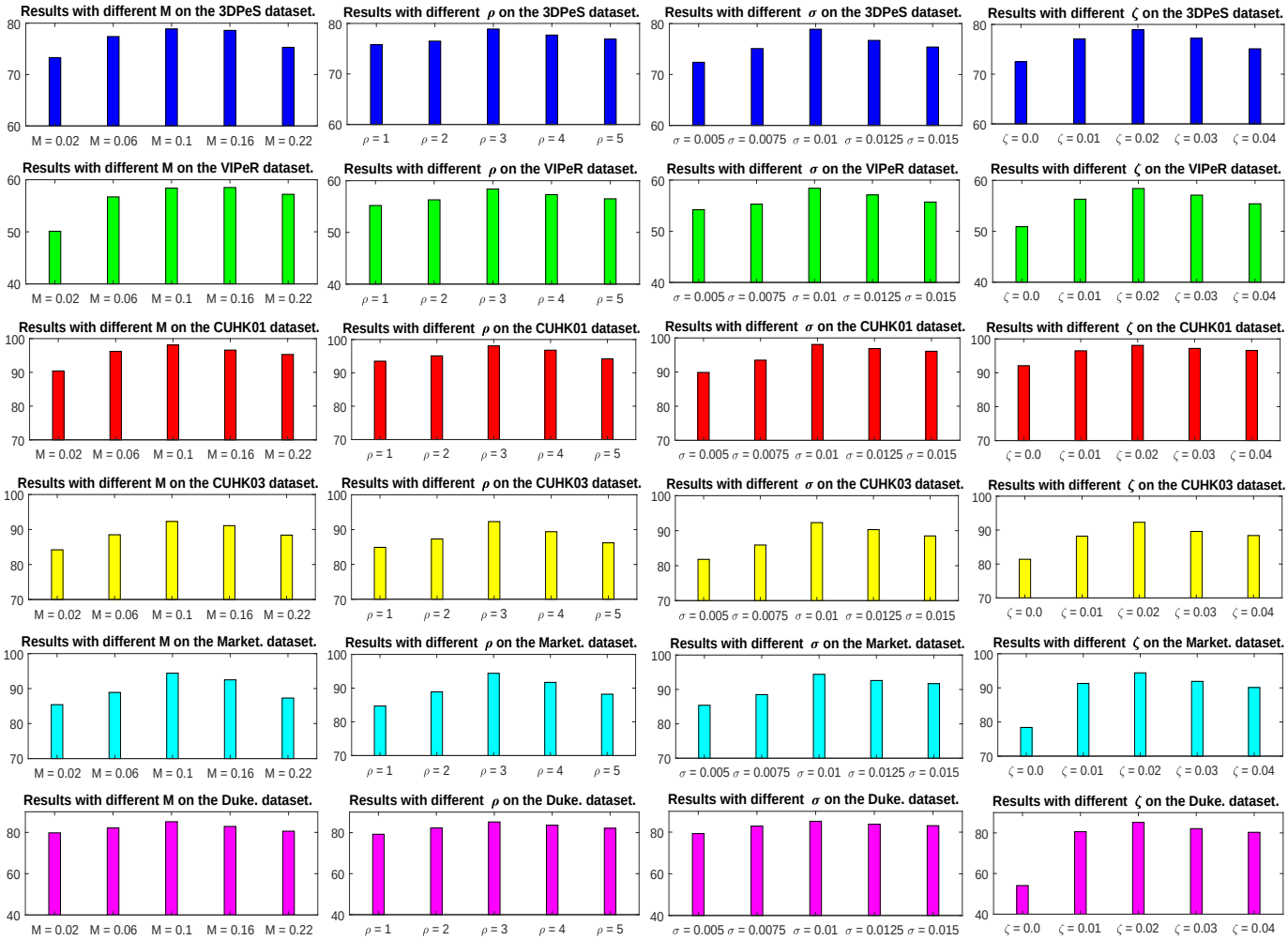


Fig. 7. Comparison of different parameter settings to the final person Re-ID performances on the six benchmark datasets. Specifically, the first to sixth rows show the detailed results on the 3DpeS, VIPeR, CUHK01, CUHK03, Market1501 and DukeMTMC-reID datasets, respectively.

experiments on each dataset.

Firstly, we evaluate the influence of margin parameter M, kernel parameters ρ and σ, and weight parameter ζ to the final person Re-ID results. The detailed results are shown in Fig. 7, in which we have reported the Top 1 accuracy with different parameter settings. From the results, we can conclude three conclusions: 1) For the margin parameter, small M will lead to small discriminative power between the positive and negative pairs, and large M will make the model pay more attention to the hard training samples. Both parameter settings are not beneficial to keep a better generation ability of learned model on the testing data. What's more, there are a relative large sliding interval of M to keep an approximate performance on the testing data, when an optimal margin parameter is chosen in the experiments. 2) For the kernel parameters, small ρ will make the reconstruction task sensitive to the isolated regions in the ground truth mask, and large ρ will cause the edge blur when reconstructs the foreground mask. Small σ will lead the reconstruction task sensitive to the difference between the reconstructed mask and ground truth mask, and large σ will make the algorithm pay less attention to the foregrounds of input images. We argue that too small or too large ρ and σ are not beneficial to the generation ability of learned model, however our algorithm allows a large variation of the two parameters around the optimal values. For the weight parameter, small ζ will also make our algorithm pay less attention to the foregrounds of input images, and large ζ will make our algorithm pay too much attention to the foregrounds of input images. Therefore, an suitable weight should be chosen to keep the person Re-ID performance.

The main difference between our symmetric triplet loss function and the asymmetric triplet loss function is that it introduces another negative distance between two samples from the same camera view to regularize the gradient back-propagation, as shown in Fig. 3. As a result, the intra-class distance can be minimized and the inter-class distance can be maximized in each triplet unit, simultaneously. In our formulation, two negative distances are weighted to represent the inter-class distance, and the weight parameters can be adaptively updated in the feature learning process, so as to deduce the symmetric gradient back-propagation. Because we apply the L2 normalization after the resulting feature vectors, the final person Re-ID performance is very stable with different initializations to the weight parameters u and v. In Table VII, we give some detailed analysis results on the six benchmark datasets. From the results, we can conclude

Fig. 8. Illustration of the generated masks by different methods, in which the first row shows the RGB images, the second row shows the masks generated by the saliency detection based method, and the third row shows the masks generated by the instance segmentation based method.

TABLE VII
THE MATCHING RATE (%) ON SIX BENCHMARK DATASETS IN TERM OF $\mathbf{u}$ AND $\mathbf{v}$ WITH USING THE L2 NORMALIZATION.

Datasets	u = 1.0,v = 0.0		u = 0.6,v = 0.4		u = 0.4,v = 0.6	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
3DPeS	74.2	91.2	**78.9**	**92.3**	77.2	92.1
VIPeR	53.2	81.3	**58.4**	**83.7**	57.3	82.4
CUHK01	92.1	96.8	98.1	**99.8**	**98.3**	99.5
CUHK03	80.9	94.9	**92.3**	99.2	92.3	**99.3**
Market.	88.9	92.1	**94.4**	**95.3**	92.1	94.8
Duke.	78.9	87.2	**85.2**	**91.6**	83.6	90.3

that the Top 1 accuracy of our method on the six benchmark datasets are robust to different weight initializations. There are two underlying reasons: 1) The weight updating algorithm in Eq. (8) to Eq. (10) can measure the difference between $\mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3)$ and $\mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)$, so as to keep $\mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3) \approx \mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)$ in the optimization. Therefore, the weights can be adaptively updated to keep the symmetric gradient back-propagation. 2) For the i^{th} triplet input, the L2 normalization is applied to keep $\|\phi(\mathbf{x}_i^j, \mathbf{\Omega}_t)\|_2^2 = 1, j = 1, 2, 3$ at the output layer, therefore the difference between $\mathrm{d}(\mathbf{x}_i^1, \mathbf{x}_i^3)$ and $\mathrm{d}(\mathbf{x}_i^2, \mathbf{x}_i^3)$ is bounded in $[0, 2]$. As a result, the L2 normalization will make our algorithm more robust to the numerical stability. For comparison, we evaluate the performances of our method without using L2 normalization, as shown in Table VIII, on the six benchmark datasets. From the results, we can see that: 1) The best results of two cases are similar on the six benchmark datasets, which indicates that both the unit sphere space and the Euclidean space are suitable for similarity comparison.[4] 2) Because the distances in unit sphere space are bounded, the performances on the six benchmark datasets are more stable even with different initializations.

In order to obtain better performances on both the small and large datasets, we use different numbers of residual blocks in the body part subnetwork. Specifically, we use

[4]For fair comparison, we set the $\mathrm{M} = 1.0$ when conduct experiments without using L2 normalization on the six benchmark datasets.

TABLE VIII
THE MATCHING RATE (%) ON SIX BENCHMARK DATASETS IN TERM OF $\mathbf{u}$ AND $\mathbf{v}$ WITHOUT USING THE L2 NORMALIZATION.

Datasets	u = 1.0,v = 0.0		u = 0.6,v = 0.4		u = 0.4,v = 0.6	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
3DPeS	73.8	90.8	**78.1**	**93.1**	74.6	92.3
VIPeR	53.6	81.4	**57.8**	**83.8**	55.1	81.9
CUHK01	91.7	96.5	**97.2**	**99.3**	93.6	98.9
CUHK03	79.3	94.5	**91.9**	**98.7**	85.8	96.9
Market.	86.5	91.1	**92.8**	**95.1**	88.2	92.2
Duke.	77.1	86.5	**84.1**	**90.6**	78.9	88.5

TABLE IX
THE MATCHING RATE (%) OF USING DIFFERENT NUMBERS OF RESIDUAL BLOCKS ON THE SIX BENCHMARK DATASETS.

Datasets	num = 1	num = 2	num = 3	num = 4	num = 5
3DPeS	77.1	**78.9**	76.2	74.1	69.1
VIPeR	56.2	**58.4**	54.2	49.1	43.6
CUHK01	94.9	97.2	**98.1**	96.5	91.0
CUHK03	82.3	87.1	89.4	**92.3**	91.4
Market.	85.1	90.3	91.5	**94.4**	93.9
Duke.	73.9	78.1	82.8	**85.2**	84.3

two residual blocks on the 3DPeS and VIPeR datasets, three residual blocks on the CUHK01 dataset, and four residual blocks on the CUHK03, Market1501 and DukeMTMC-reID datasets, respectively. The main reason is that training a deep neural network usually needs a large amount of labeled data. Therefore, the under-fitting problem will be caused if we use a small dataset to train a complex network. We think this is the right reason why the deep neural networks, such as the VGGNet [13] and ResNet [14], usually can't work well on the small datasets. Besides, the heavy deep neural networks usually have stronger representation capacity than the light ones. As a result, we choose different numbers of residual blocks to adapt the scale of different datasets. In order to support our understanding, we evaluate the performance of our FANN with different numbers of residual blocks on each dataset. The results are shown in Table IX, in which we can find that the shallower networks can obtain better results than the deeper ones on the 3DPeS, VIPeR and CUHK01 datasets, while the deeper networks can obtain better results than the shallower ones on the CUHK03, Market1501 and DukeMTMC-reID datasets.

Some examples of ranking. Finally, we illustrate some real ranking examples of our method on the six benchmark datasets, as shown in Fig. 9, including both the successful cases and failure cases. Specifically, the images in green boxes are the probes, which are used to find out the matched references from the gallery. The image in red box is a true match to the corresponding probe, in which the smaller order indicates the better performance. In the successful cases of our method, all the matched candidates are found out in the first place from various candidates in the gallery. The results indicate that our method can learn a discriminative feature representation to overcome the large cross-view appearance variations. However, we also notice that there are a fraction of failure cases in our method, in which the matched references can not be ranked firstly from the very similar candidates. In

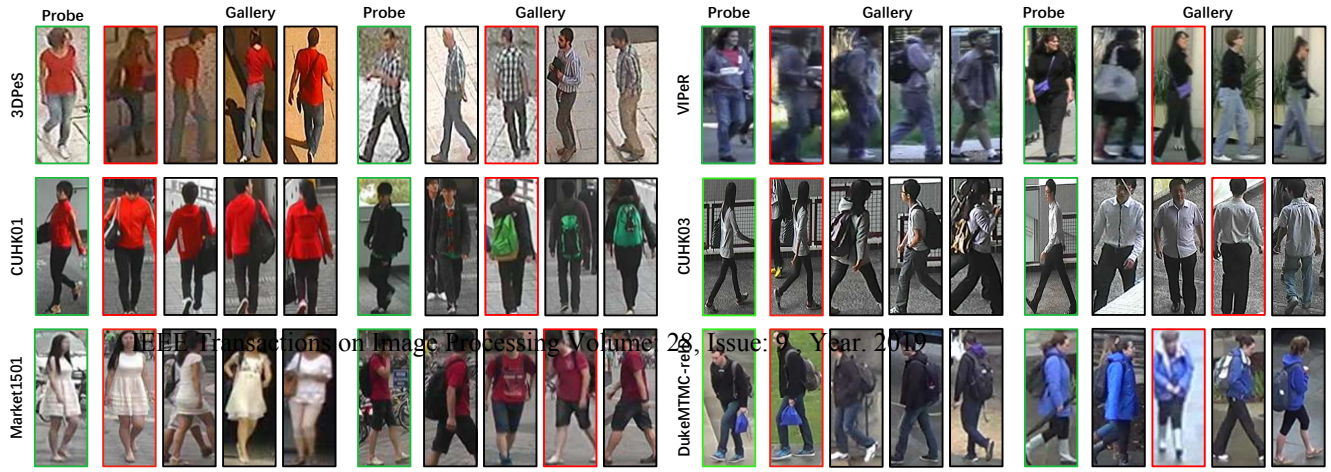


Fig. 9. Illustration of some ranking examples on the six benchmark datasets, in which the probe images are denoted by green bounding boxes and the matched references are indicated by using the red bounding boxes. Notice that both the successful cases and failure cases are given for better comparison.

the future study, we will study how to enrich the diversity of similar training samples, so as to reduce the failure cases in solving person Re-ID problem.

VI. Conclusion

In this paper, we propose a simple yet effective deep neural network to learn a discriminative feature representation from the foreground of each input image for person Re-ID. Firstly, a FANN is constructed to jointly enhance the positive influences of foregrounds and weaken the side effects of backgrounds, in which an encoder and decoder network is built to guide the whole network to directly learn a discriminative feature representation from the foreground persons. Secondly, a novel local regression loss function is designed to deal with the isolated regions in the ground truth masks by considering the local information in a neighborhood. Thirdly, a symmetric triplet loss function is introduced to supervise the feature learning process, which can jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Extensive experiments on the 3DPeS, VIPeR, CUHK01, CUHK03, Market1501 and DukeMTMC-reID datasets are conducted, and the results have shown that our method can significantly outperform the state-of-the-art approaches.

Acknowledgment

This work is jointly supported by the National Key Research and Development Program of China under Grant No. 2017YFA0700805, and the National Science Foundation of China under Grant No. 61473219.

References

[1] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 8, pp. 1114–1127, 2008.

[2] S. Zhang, J. Wang, Z. Wang, Y. Gong, and Y. Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognition*, vol. 48, no. 2, pp. 580–590, 2015.

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 34, no. 3, pp. 334–352, 2004.

[4] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[5] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[6] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[7] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[8] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[10] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[11] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 1735–1742.

[17] M. M. Kalayeh, E. Basaran, M. Gkmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] S. Zhou, J. Wang, J. Wang, G. Yihong, and Z. Nanning, "Point to set similarity based deep feature learning for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2017.

[21] D. Baltieri, R. Vezzani, and R. Cucchiara, "Sarc3d: a new 3d body model for people tracking and re-identification," in *Proceedings of the 16th International Conference on Image Analysis and Processing*, Ravenna, Italy, Sep. 2011, pp. 197–206.

[22] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3594–3601.

[23] W. Li, R. Zhao, T. Xiao, and 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.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[25] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[26] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 3, pp. 653–668, 2013.

[27] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2666–2672.

[28] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3318–3325.

[29] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Computer Vision–ACCV 2010*. Springer, 2011, pp. 709–720.

[30] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 3567–3571.

[31] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573.

[32] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 780–793.

[33] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 1249–1258.

[34] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 1288–1296.

[35] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, July 2017.

[36] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2265–2274.

[37] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 3376–3385.

[38] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[39] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, July 2017.

[41] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[42] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[43] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *null*. IEEE, 2006, pp. 1528–1535.

[44] S. Zhang, Y. Gong, and J. Wang, "Deep metric learning with improved triplet loss for face clustering in videos," in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 497–508.

[45] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[46] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[47] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

[49] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[50] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2590–2600.

[51] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," *arXiv preprint arXiv:1803.09132*, 2018.

[52] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," *arXiv preprint arXiv:1803.11353*, 2018.

[53] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1470–1478.

[54] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," *arXiv preprint arXiv:1805.03344*, 2018.

[55] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[56] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8042–8051.

[57] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[58] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[59] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

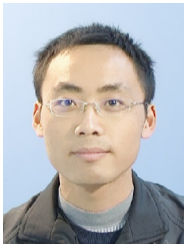
[60] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *European Conference on Computer Vision*. Springer, 2016, pp. 875–890.

[61] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[62] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, and N. Zheng, "Deep self-paced learning for person re-identification," *Pattern Recognition*, 2017.

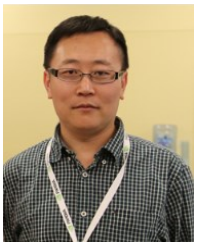
[63] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision*. Springer, 2012, pp. 413–422.

[64] ——, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.
[65] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239–1248.
[66] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[67] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*. IEEE, 2017, pp. 3980–3989.
[68] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8649–8658.
[69] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *The European Conference on Computer Vision (ECCV)*, September 2018.
[70] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set-to-set similarity comparison for person reidentification," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 593–604, 2018.
[71] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.
[72] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," *arXiv preprint arXiv:1803.09937*, 2018.
[73] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *The European Conference on Computer Vision (ECCV)*, September 2018.
[74] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection." IJCAI, 2018.



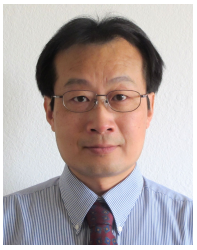
Sanping Zhou received the M.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree in Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. His research interests include machine learning, deep learning and computer vision, with a focus on medical image segmentation, person re-identification, image retrieval, image classification and visual tracking.

Jinjun Wang received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, China, in 2000 and 2003, respectively. He received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. From 2006 to 2009, he was with NEC Laboratories America, Inc., as a Research Scientist, and Epson Research and Development, Inc., as a Senior Research Scientist, from 2010 to 2013. He is currently a Professor with Xi'an Jiaotong University. His research interests include pattern classification, image/video enhancement and editing, content-based image/video annotation and retrieval, semantic event detection, etc.



Deyu Meng Deyu Meng received the B.S., M.S., and Ph.D. degrees from Xian Jiaotong University, Xian, China, in 2001, 2004, and 2008, respectively. From 2012 to 2014, he took his two-year sabbatical leave in Carnegie Mellon University. He is currently a Professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xian Jiaotong University. His current research interests include self-paced learning, noise modeling, and tensor sparsity.

Yudong Liang received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2010 and 2017, respectively. He is currently an assistant Professor with School of Computer and Information Technology, Shanxi University. His research interests include Machine Learning and Computer Vision, with a focus on image super-resolution, image quality assessment and deep learning.



Yihong Gong received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from The University of Tokyo, Japan, in 1987, 1989, and 1992, respectively. In 1992, he joined Nanyang Technological University, Singapore, as an Assistant Professor with the School of Electrical and Electronic Engineering. From 1996 to 1998, he was a Project Scientist with the Robotics Institute, Carnegie Mellon University, USA. Since 1999, he has been with the Silicon Valley branch, NEC Labs America, as a Group Leader, the Department Head, and the Branch Manager. In 2012, he joined Xi'an Jiaotong University, China, as a Distinguished Professor. His research interests include image and video analysis, multimedia database systems, and machine learning.

Nanning Zheng (SM93-F06) graduated from the Department of Electrical Engineering, Xian Jiaotong University, Xian, China, in 1975, and received the M.S. degree in information and control engineering from Xian Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He jointed Xian Jiaotong University in 1975, and he is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems. Dr. Zheng became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition.