Discriminative Feature Learning with Consistent Attention Regularization for Person Re-identification

Sanping Zhou [1], Fei Wang [2], Zeyi Huang [3], Jinjun Wang [1] *
1. The Institute of Artificial Intelligence and Robotic, Xi'an Jiaotong University
2. School of Computer Science and Technology, Xi'an Jiaotong University
3. Robotics Institute, Carnegie Mellon University

Abstract

Person re-identification (Re-ID) has undergone a rapid development with the blooming of deep neural network. Most methods are very easily affected by target misalignment and background clutter in the training process. In this paper, we propose a simple yet effective feedforward attention network to address the two mentioned problems, in which a novel consistent attention regularizer and an improved triplet loss are designed to learn foreground attentive features for person Re-ID. Specifically, the consistent attention regularizer aims to keep the deduced foreground masks similar from the low-level, mid-level and high-level feature maps. As a result, the network will focus on the foreground regions at the lower layers, which is benefit to learn discriminative features from the foreground regions at the higher layers. Last but not least, the improved triplet loss is introduced to enhance the feature learning capability, which can jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Experimental results on the Market1501, DukeMTMC-reID and CUHK03 datasets have shown that our method outperforms most of the state-of-the-art approaches.

1. Introduction

Person re-identification (Re-ID) is a critical technology in video surveillance, which aims to associate the same pedestrian across the non-overlapping camera views. With the blooming of convolutional neural network, the current deep feature learning based methods [5, 8, 53, 61] have significantly outperformed a variety of traditional feature learning based approaches [33, 43]. In practice, it is critical to learn a discriminative feature representation in solving the person Re-ID problem. However, the learned features are very easily degenerated by target misalignment and background clutter, because most of the deep feature learning based methods usually try to learn discriminative features from the whole input images.

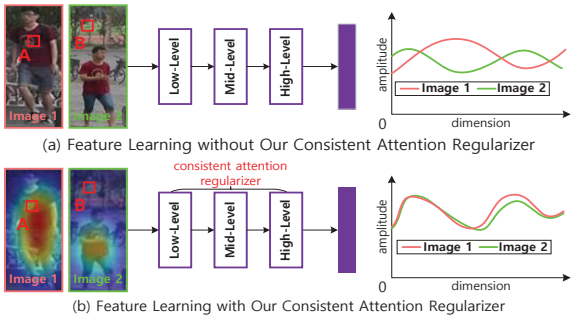

Figure 1. Motivation of our consistent attention regularizer, which aims to drive the network focus on foreground regions at the lower layers. Therefore the network will learn a discriminative feature representation to enhance the useful signals from point A and suppress the noise signals from point B, at the higher layers. From the final features learned in (a) and (b), we can find that the consistent attention regularizer is critical to associate two samples with target misalignment and background clutter.

As a data-driven approach, the deep feature learning based methods [22, 46, 50] can autonomously focus most of their attentions on the foreground regions of input images. However, the networks are very easily misguided if we haven't an explicit regularizer to drive its attention in the feature learning process [58]. To solve this problem, two mainstream approaches have been widely studied in the past few years. The first line of methods are based on the part-based networks [5, 38, 62], in which they try to learn discriminative features from the predefined body parts. The second line of methods are based on the foreground attentions [20, 29, 34, 39, 54, 59], in which person masks are used to drive the attention in a supervised manner or attention mechanisms are applied to deduce the attention in an unsupervised manner. In general, it is much easier to learn a discriminative feature representation with the annotated person masks, because it can help the network to precisely focus on the foreground regions at the lower layers.

*Jinjun Wang is the corresponding author.

Many off-the-shelf methods [9, 57] have been widely used to generate the foreground masks for person Re-ID, however the resulting person masks are usually in poor quality due to the low resolution of input images. As a result, there is a high risk that the foreground attention will be misguided at the lower layers [34]. In order to alleviate this problem, it is better to incorporate the discriminative feature learning and foreground attention deducing in an end-to-end network, because they can benefit from each other in the training process. As shown in Figure 1, it becomes an important issue that how to deduce the foreground attentions at the lower layers, so as to learn the foreground attentive features at the higher layers and suppress the noise signals caused by target misalignment and background clutter.

In this paper, we design a simple yet effective attention network to learn a discriminative feature representation from the foreground regions for person Re-ID. Our method is inspired by the phenomenon [58] that the high-level feature maps usually contain much more semantic information than the low-level feature maps. Therefore, it will be much easier to deduce the high-quality foreground masks from the high-level feature maps rather than from the low-level feature maps. Specifically, we first design a novel feedforward attention network which can learn the foreground masks from the low-level, mid-level and high-level feature maps, respectively. Then, a novel consistent attention regularizer is designed to transmit the foreground information from the high-level to mid-level and low-level feature maps. In this manner, the high-quality foreground masks learned from the high-level feature maps can be further used to help the lower layers focus on the foreground regions. Finally, an improved triplet loss is introduced to enhance the feature learning capability, which can jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Our network is trained in an end-to-end manner, which can effectively learn discriminative features to match images of the same person in a large camera system.

The main contributions of our paper can be highlighted as follows: 1) A novel feedforward attention network is designed to learn foreground masks from the low-level, mid-level and high-level feature maps, respectively. 2) A novel consistent attention regularizer is put forward to keep the deduced foreground masks similar in the training process, which is benefit to drive the network to focus on foreground regions at the lower layers. 3) A novel triplet loss is built to supervise feature learning by jointly minimizing the intra-class distance and maximizing the inter-class distance in each triplet unit. We conduct extensive experiments on the Market1501 [56], DukeMTMC-reID [27] and CUHK03 [54] datasets, which have shown significant improvements by our method as compared with the state-of-the-art approaches.

2. Related Work

Our method aims to learn a discriminative feature representation through the consistent attention regularization, therefore we review two lines of related works in terms of deep feature learning and deep attention learning.

Deep feature learning. A robust feature representation is very critical to solve the person Re-ID problem, and the deep feature learning based methods mainly focus on learning a discriminative feature representation from input images. For this purpose, different loss functions have been developed, such as the triplet loss [8], quadruplet loss [5], center loss [47], and softmax loss [16], to guide the feature learning process. Meanwhile, a large number of well-known networks have been designed to extract features from the input images, including the ResNet [10], DenseNet [13], MobileNet [28] and ShuffleNet [23]. In addition, different part strategies [5, 17, 38, 60] have been widely used to enhance the feature representation capability of backbone networks. In recent years, the Generative Adversarial Networks (GAN) [7, 45, 58] have been extensively studied to augment the training data for person Re-ID, which is an effective way to enhance the generalization ability of leaned features. Despite learning features from the single images, another line of methods [3, 24, 49, 63] have tried to learn the temporal-spatial features from video clips. Due to the strong representation capability of deep neural network, the deep feature learning based methods have achieved the state-of-the-art performance on the benchmark datasets for person Re-ID.

Deep attention learning. The deep attention learning has been extensively studied in the computer vision community, which can effectively improve the algorithm's performance by addressing the useful information [40]. In general, the deep attention learning based methods can be divided into the supervised and unsupervised lines. In the former ones, the labeled ground truth is needed to supervise the learning process. For example, the foreground masks [15, 34, 39] have been widely used to guide the networks to focus their attentions on the body regions, so as to learn discriminative features for person Re-ID. Besides, the predefined regions [12, 55] are usually used to drive the network to learn fine features from the local regions, which have been extensively studied in solving the fine-grained image classification problem. In the later ones, the self-attention mechanisms or heuristic knowledge are usually used to guide the attention learning. For instance, several works [20, 54] have designed different attention modules to guide the networks to put their attentions on the discriminative body regions. The deep residual attention learning [41] has been successfully applied in image classification. In addition, the temporal-spatial clues [25, 35] have been widely used to supervise the attention learning in video recognition and classification.

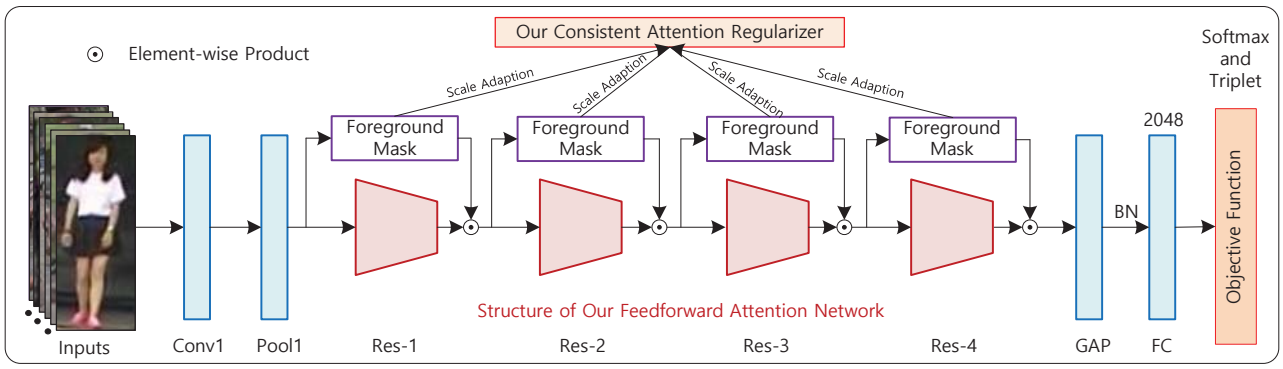

Figure 2. Illustration of our feedforward attention network, which works as follows: The foreground masks are firstly learned from the low-level, mid-level and high-level feature maps, respectively. Then, the consistent attention regularizer is applied to keep the deduced foreground masks similar, so as to drive the network focus on foreground regions at the lower layers. Finally, the improved triplet loss and softmax loss are jointly used to learn discriminative features in a multi-task learning framework.

3. Our method

Given a set of training samples $\mathbf{X} = \{\mathcal{X}_i, \mathcal{Y}_i\}_{i=1}^{N}$, in which $\mathcal{X}_i$ indicates the i^{th} input image and $\mathcal{Y}_i$ represents the corresponding label, our method tries to learn a discriminative feature representation from the foregound regions of input images. The structure of our feedforward attention network is illustrated in Figure 2, in which a novel consistent attention regularizer and an improved triplet loss are designed to learn discriminative features for person Re-ID. Without loss of generality, we choose the ResNet50 [10] as backbone. In the following paragraphs, we will explain our method in detail.

3.1. Network Structure

Our feedforward attention network aims to learn discriminative features from the foreground regions, therefore two requirements need to be satisfied in the network design. Firstly, the backbone network should be powerful enough, so as to extract discriminative features at the output layer. In our network structure, we choose the ResNet50 as our backbone, which is mainly consisted of a convolutional layer, a max pooling layer and four residual blocks. In particular, one Global Average Pooling (GAP) [21] layer and a Fully-Connected (FC) layer are used to obtain a 2048 dimensional feature vector. Besides, one Batch Normalization (BN) [14] layer is deployed between the GAP and FC layers. Secondly, an attention module should be designed to deduce the foreground masks from feature maps. For this purpose, we take heat map to represent the foreground mask and use the resulting foreground mask to filter the corresponding feature maps in the training process. As shown in Figure 3, our attention module takes the feature maps $\mathbf{T}_k$ as input and outputs the deduced foreground mask $\mathbf{H}_k$, which can be modeled as follows:

$$\mathbf{H}_k = \mathrm{M}_{\mathrm{ask}}(\mathbf{T}_k; \mathbf{\Theta}_k), \tag{1}$$

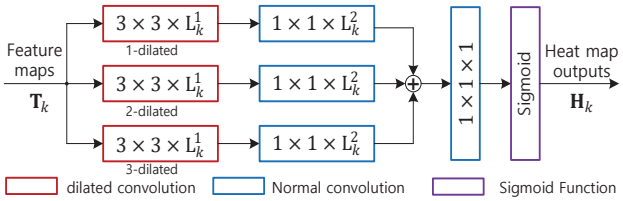

Figure 3. Illustration of our attention module. For simplicity, we suppose the input feature maps $\mathbf{T}_k$ have L_k feature maps, then we fuse them in a gradual way: $\mathrm{L}_k^1 = \frac{1}{2}\mathrm{L}_k$ and $\mathrm{L}_k^2 = \frac{1}{2}\mathrm{L}_k^1$. Besides, three dilated convolutional layers with different dilation ratios are used to deduce the foreground mask from a local to global view.

where $\mathbf{\Theta}_k$ represents the parameters of our k^{th} attention module. In our design, we have the following considerations: 1) At first, we take two convolutional layers to reduce the number of feature maps to $1/4$ of its own, so as to summarize them in a gradual way. Then, another convolutional layer with a kernel in size of 1×1 is applied to further get the heat map. At last, a sigmoid function is used to normalize the heat map in $[0, 1]$. 2) The multi-scale information has been applied to deduce the foreground masks from a local to global view. As the same in [17], three different receptive fields, namely 7, 5 and 3, have been used to extract the context information by using different dilation ratios in the dilated convolutional layers.

Once the attention module is designed, we embed it in the ResNet50 and use the resulting heat map to filter the output feature maps of each residual block as follows:

$$\mathbf{T}_k^a(x, y, c) = \mathbf{T}_k^b(x, y, c) \times \mathbf{H}_k(x, y), \tag{2}$$

where $\mathbf{H}_k(x, y)$ denotes the deduced attention response at the coordinate (x, y), $\mathbf{T}_k^a(x, y, c)$ and $\mathbf{T}_k^b(x, y, c)$ represent the output and input responses at the coordinate (x, y) from the c^{th} feature map, respectively. As shown in Figure 2, our feedforward attention network works as follows: 1) In the forward propagation, the backbone network first extracts

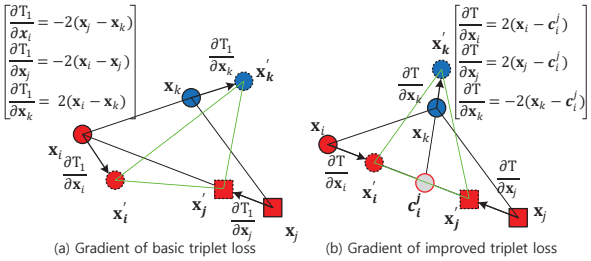

Figure 4. Differences between the two triplet losses in gradient back-propagation. In particular, our triplet loss introduces one point $\mathbf{c}_i^j$ on the line between $\mathbf{x}_i$ and $\mathbf{x}_j$ to model all the pairwise relationships in each triplet unit, so as to consistently minimize the intra-class distance in the training process.

the discriminative features from input images, then the attention module deduces the foreground mask from the corresponding feature maps, and finally the generated feature maps are further filtered by the resulting foreground masks with the element-wise product. 2) The parameters of backbone network and attention modules are jointly optimized in the backward propagation, therefore our feedforward attention network will focus most of its own attentions on the foreground regions in the next iteration.

3.2. Objective Function

The objective function is consisted of two loss terms and one regularizer, which can be formulated as follows:

$$\mathcal{L}(\mathbf{W}, \mathbf{\Theta}) = \mathcal{L}_1(\mathbf{X}; \mathbf{W}) + \alpha \mathcal{L}_2(\mathbf{X}; \mathbf{W}) + \mathcal{L}_3(\mathbf{H}; \mathbf{\Theta}), \quad (3)$$

where $\mathcal{L}_1(\cdot)$ represents the softmax loss, $\mathcal{L}_2(\cdot)$ indicates the improved triplet loss, $\mathcal{L}_3(\cdot)$ denotes the consistent attention regularizer, and α is a constant weight. In the training process, the two loss terms aim to learn a discriminative feature representation from the raw input images, and the consistent attention regularizer tries to keep these foreground masks similar, which are deduced from the low-level, mid-level and high-level feature maps, respectively.

Because of its powerful capability, the softmax loss has been widely used in training the deep neural network. Therefore, we introduce it to supervise the feature learning process, which can be formulated as follows:

$$\mathcal{L}_1(\mathbf{X}; \mathbf{W}) = \frac{1}{N} \sum\nolimits_{i=1}^{N} -\log\left(\frac{\exp(\mathbf{p}_{\mathcal{Y}_i}^{\mathrm{T}} \mathbf{x}_i)}{\sum_g \exp(\mathbf{p}_g^{\mathrm{T}} \mathbf{x}_i)}\right), \quad (4)$$

where $\mathbf{p}_g$ denotes the g^{th} column of the learned classifier, and $\mathbf{x}_i$ represents the feature vector learned by our feedforward attention network for input image $\mathcal{X}_i$.

In order to apply the improved triplet loss to learn the discriminative features from input images, we first organize the training samples into a set of triplet units, $\mathbf{S} = \{(\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k)\}$, in which $(\mathcal{X}_i, \mathcal{X}_j)$ represents a positive pair with $\mathcal{Y}_i = \mathcal{Y}_j$, and $(\mathcal{X}_i, \mathcal{X}_k)$ indicates a negative pair with $\mathcal{Y}_i \neq \mathcal{Y}_k$. In each triplet unit, we solve a ranking problem by using the improved triplet loss:

$$\mathrm{T} = [m + d(\mathbf{x}_i, \mathbf{c}_i^j) + d(\mathbf{x}_j, \mathbf{c}_i^j) - d(\mathbf{x}_k, \mathbf{c}_i^j)]_+, \quad (5)$$

where $d(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{z}_1 - \mathbf{z}_2\|_2^2$ denotes the squired distance in feature space, m represents the margin parameter, and $\mathbf{c}_i^j = \eta \mathbf{x}_i + (1-\eta)\mathbf{x}_j$ indicates one point lied on the line between $\mathbf{x}_i$ and $\mathbf{x}_j$ [1]. As a result, $\mathbf{x}_i$ and $\mathbf{x}_j$ will move towards $\mathbf{c}_i^j$, and the intra-class distance can be consistently minimized in the training process.

Discussion. To the best of our knowledge, a series of triplet losses have been designed in the past few years. The basic triplet loss [8] is defined as follows:

$$\mathrm{T}_1 = [m + d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k)]_+. \quad \text{(D1)}$$

Besides, some researchers have focused on how to improve the gradient back-propagation in their modifications. For example, the dual triplet loss [52] is defined as follows:

$$\mathrm{T}_2 = [m + d(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2}[d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_j, \mathbf{x}_k)]]_+, \quad \text{(D2)}$$

and the symmetric triplet loss [62] is defined as follows:

$$\mathrm{T}_3 = [m + d(\mathbf{x}_i, \mathbf{x}_j) - [ud(\mathbf{x}_i, \mathbf{x}_k) + vd(\mathbf{x}_j, \mathbf{x}_k)]]_+. \quad \text{(D3)}$$

Firstly, we compare the gradient back-propagation between our triplet loss and the basic one, as shown in Figure 4, and the differences come from two aspects: 1) The basic triplet loss only considers one positive pair $(\mathcal{X}_i, \mathcal{X}_j)$ and one negative pair $(\mathcal{X}_i, \mathcal{X}_k)$, which neglects another negative pair $(\mathcal{X}_j, \mathcal{X}_k)$ in their formulation. Our triplet loss introduces the center point $\mathbf{c}_i^j$ of positive pair to help model all the pairwise relationships in each triplet unit. 2) Because of the resulting advantages in gradient back-propagation, our triplet loss can continuously minimize the intra-class distance, while the basic triplet loss is hard to achieve this goal in the training process.

Secondly, we conclude the relationships of these triplet losses as follows: 1) We can find that $\mathrm{T}_2(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \frac{1}{2}[\mathrm{T}_1(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \mathrm{T}_1(\mathbf{x}_j, \mathbf{x}_i, \mathbf{x}_k)]$, which indicates that it is important to model all the pairwise relationships in each triplet unit. 2) The symmetric loss is a generalized version of the dual triplet loss, in which it designs a novel algorithm to update u and v in the training process. 3) Our triplet loss doesn't need to use any additional algorithm to achieve a more robust performance than the symmetric triplet loss.

Now, we extend our triplet loss to the whole triplet units, which can be formulated as follows:

$$\mathcal{L}_2(\mathbf{X}; \mathbf{W}) = \frac{1}{|\mathbf{S}|} \sum\nolimits_{(\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k) \in \mathbf{S}} \mathrm{T}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k), \quad (6)$$

where $|\mathbf{S}|$ indicates the number of triplet units in $\mathbf{S}$.

[1] In order to keep our triplet loss outperforms the basic one, we need to set $\eta \in (0, 1)$, and we choose $\eta = 0.5$ in all the experiments. If $\eta = 1$, the basic triplet loss will become a special case of our method.

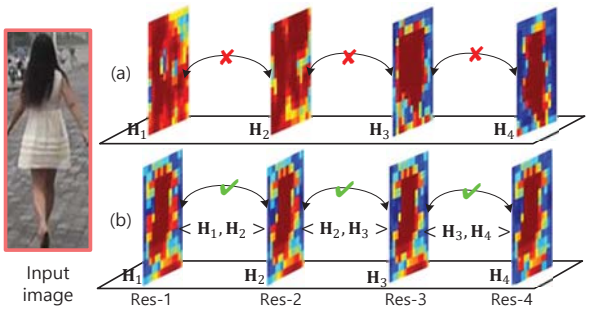

Figure 5. Illustration of the deduced heat maps from the low-level, mid-level and high-level feature maps, respectively. In particular, (a) shows the learned heat maps without applying the consistent attention regularizer, and (b) shows the learned heat maps by using our consistent attention regularizer.

Finally, we introduce the consistent attention regularizer to keep all the deduced foreground masks similar in the training process, which is defined as follows:

$$\mathcal{L}_3(\mathbf{H};\mathbf{\Theta}) = \frac{\beta}{K}\sum_{k=1}^{K}\|\mathbf{H}_{k+1} - \widehat{\mathbf{H}}_k\|_F^2 + \frac{\varpi}{K+1}\sum_{k=1}^{K+1}\|\mathbf{H}_k\|_1, \tag{7}$$

where $K+1$ denotes the number of heat maps, and β, ϖ are two constant weights. Besides, $\widehat{\mathbf{H}}_k$ is in the same size with $\mathbf{H}_{k+1}$, which is obtained by a max-pooling of $\mathbf{H}_k$ with stride 2. Because there are four residual blocks in the ResNet50, we set $K = 3$ in all the experiments. Our consistent attention regularizer is consisted of two terms, *i.e.*, the consistence term and sparsity term, in which: 1) The consistence term aims to keep these heat maps similar, which are learned from the low-level, mid-level and high-level feature maps, respectively. As a result, the high-quality foreground masks learned from the high-level feature maps can be used to help the network focus on foreground regions at the lower layers. 2) The sparsity term tends to do feature selection, which is benefit to remove some false positive responses in background. We compare two different sets of heat maps in Figure 5, from which we can see that the heat maps learned by using our consistent attention regularizer are much better than these without using this regularizer .

3.3. Optimization

We optimize the deep parameters $\mathbf{W}$, $\mathbf{\Theta}$ by using the Stochastic Gradient Descent (SGD) algorithm. For simplicity, we take $\mathbf{\Omega} = [\mathbf{W}, \mathbf{\Theta}]$ as a whole and compute the partial derivate of Eq. (3) as follows:

$$\frac{\partial\mathcal{L}(\mathbf{\Omega})}{\partial\mathbf{\Omega}} = \frac{\partial\mathcal{L}_1(\mathbf{X};\mathbf{W})}{\partial\mathbf{W}} + \alpha\frac{\partial\mathcal{L}_2(\mathbf{X};\mathbf{W})}{\partial\mathbf{W}} + \frac{\partial\mathcal{L}_3(\mathbf{H};\mathbf{\Theta})}{\partial\mathbf{\Theta}}, \tag{8}$$

where $\partial\mathcal{L}_1(\mathbf{X};\mathbf{W})/\partial\mathbf{W}$ can be easily computed by using the off-the-shelf algorithm, and $\partial\mathcal{L}_2(\mathbf{X};\mathbf{W})/\partial\mathbf{W}$ and $\partial\mathcal{L}_3(\mathbf{H};\mathbf{\Theta})/\partial\mathbf{\Theta}$ are derived in the following paragraphs.

Algorithm 1 Consistent attention gradient descent.

Input: The training data $\mathbf{X}$, learning rate τ, maximum iteration number Q, weight parameters α, β and ϖ, and margin parameter m.
Output: The network parameters $\mathbf{\Omega} = [\mathbf{W}, \mathbf{\Theta}]$.
repeat
  repeat
    1) Compute $\frac{\partial\mathcal{L}_1}{\partial\mathbf{W}}$ using the off-the-shelf algorithm;
    2) Compute $\frac{\partial\mathcal{L}_2}{\partial\mathbf{W}}$ according to Eq. (9);
    3) Compute $\frac{\partial\mathcal{L}_3}{\partial\mathbf{\Theta}}$ according to Eq. (11);
    4) Update the gradients $\frac{\partial\mathcal{L}}{\partial\mathbf{\Omega}}$ according to Eq. (8);
  until Traverse all the triplet inputs $\{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)\}$ in each min-batch;
  2. Update $\mathbf{\Omega}^{(q+1)} = \mathbf{\Omega}^{(q)} - \tau_q\frac{\partial\mathcal{L}}{\partial\mathbf{\Omega}^{(q)}}$ and $q \leftarrow q+1$.
until $q > Q$

We denote $r = m + d(\mathbf{x}_i, \mathbf{c}_i^j) + d(\mathbf{x}_j, \mathbf{c}_i^j) - d(\mathbf{x}_k, \mathbf{c}_i^j)$, then the partial derivate of our triplet loss can be formulated as follows:

$$\frac{\partial\mathcal{L}_2(\mathbf{X};\mathbf{W})}{\partial\mathbf{W}} = \begin{cases} \frac{1}{|\mathbf{S}|}\sum\limits_{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)\in\mathbf{S}}\frac{\partial\mathcal{P}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)}{\partial\mathbf{W}}, & if\ r > 0, \\ 0 & , else. \end{cases} \tag{9}$$

in which $\partial\mathcal{P}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)/\partial\mathbf{W}$ is computed as follows:

$$\begin{aligned} \frac{\partial\mathcal{P}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)}{\partial\mathbf{W}} &= 2(\mathbf{x}_i - \mathbf{c}_i^j)\cdot\frac{\partial\mathbf{x}_i - \partial\mathbf{c}_i^j}{\partial\mathbf{W}} \\ &+2(\mathbf{x}_j - \mathbf{c}_i^j)\cdot\frac{\partial\mathbf{x}_j - \partial\mathbf{c}_i^j}{\partial\mathbf{W}}. \\ &-2(\mathbf{x}_k - \mathbf{c}_i^j)\cdot\frac{\partial\mathbf{x}_k - \partial\mathbf{c}_i^j}{\partial\mathbf{W}} \end{aligned} \tag{10}$$

The partial derivate of our consistent attention regularizer is computed as follows:

$$\frac{\partial\mathcal{L}_3(\mathbf{H};\mathbf{\Theta})}{\partial\mathbf{\Theta}} = \frac{\beta}{K}\sum_{k=1}^{K}\ell_c(\mathbf{H}_{k+1},\widehat{\mathbf{H}}_k) + \frac{\varpi}{K+1}\sum_{k=1}^{K+1}\ell_s(\mathbf{H}_k), \tag{11}$$

where $\ell_c(\mathbf{H}_{k+1},\widehat{\mathbf{H}}_k)$ and $\ell_s(\mathbf{H}_k)$ are computed as follows:

$$\ell_c(\mathbf{H}_{k+1},\widehat{\mathbf{H}}_k) = 2(\mathbf{H}_{k+1} - \widehat{\mathbf{H}}_k)\cdot\frac{\partial\mathbf{H}_{k+1} - \partial\widehat{\mathbf{H}}_k}{\partial\mathbf{\Theta}}, \tag{12}$$

$$\ell_s(\mathbf{H}_k) = \mathrm{sign}(\mathbf{H}_k)\cdot\frac{\partial\mathbf{H}_k}{\partial\mathbf{\Theta}}, \tag{13}$$

where $\mathrm{sign}(\cdot)$ denotes the sign function, in which $\mathrm{sign}(z) = 1$ if $z > 0$, and otherwise $\mathrm{sign}(z) = -1$.

Because our method needs to back-propagate gradients to learn a discriminative feature representation by using our consistent attention regularizer, we name it as the consistent attention gradient descent algorithm. Algorithm 1 shows the overall implementation of our training process.

Index	Network	Losses	Market1501				DukeMTMC-reID		CUHK03			
			Single-Query		Multi-Query		Single-Query		Labeled		Detected	
			Top 1	mAP	Top 1	mAP	Top 1	mAP	Top 1	Top 5	Top 1	Top 5
1	ResNet.	S	87.5	72.8	91.2	79.4	78.3	62.1	72.1	91.2	66.5	88.4
2	ResNet.	BT	87.0	72.4	91.3	79.5	77.6	61.8	73.2	92.2	68.1	89.6
3	ResNet.	S+BT	89.1	75.0	92.4	81.0	79.7	64.9	76.8	93.8	74.8	93.0
4	ResNet.	IT	89.7	75.8	92.9	81.4	79.2	64.5	77.1	94.2	74.1	92.9
5	ResNet.	S+IT	93.4	79.2	94.2	82.5	82.1	68.4	82.4	96.6	78.4	94.5
6	ResNet.(AM)	S	87.8	73.0	91.6	79.8	78.9	63.6	74.1	92.8	70.9	90.9
7	ResNet.(AM)	BT	87.1	72.5	91.2	79.5	78.1	62.0	76.5	93.6	72.9	91.8
8	ResNet.(AM)	S+BT	89.4	75.4	92.5	81.1	81.2	68.1	81.1	95.8	77.8	94.3
9	ResNet.(AM)	IT	90.2	76.6	93.3	82.0	79.8	65.2	81.3	96.1	78.1	94.4
10	ResNet.(AM)	S+IT	93.9	79.5	94.6	82.9	82.6	69.1	88.4	97.8	85.5	96.6
11	ResNet.(AM)	S+CA	89.3	75.4	92.7	81.2	81.6	68.4	78.5	94.6	75.1	93.2
12	ResNet.(AM)	BT+CA	88.9	74.9	92.7	80.9	80.9	67.9	80.1	95.4	76.9	93.8
13	ResNet.(AM)	S+BT+CA	92.1	78.6	93.8	82.5	83.5	70.4	86.6	97.2	82.4	96.0
14	ResNet.(AM)	IT+CA	93.3	79.2	95.2	83.7	83.1	70.2	89.1	98.1	87.1	97.3
15	ResNet.(AM)	S+IT+CA	96.1	84.7	98.2	87.3	86.3	73.1	96.9	99.6	93.2	99.2

Table 1. Matching rates(%) of different variants of our method on the three benchmark datasets, in which 1) AM: Attention Module; 2) S: Softmax Loss; 3) BT: Basic Triplet Loss; 4) IT: Improved Triplet Loss; 5) CA: Consistent Attention Regularizer.

4. Experiments

4.1. Settings

Datasets. We conduct experiments on three large-scale datasets, *i.e.*, the Market1501 [56], DukeMTMC-reID [27] and CUHK03 [18]. The Market1501 dataset contains 32,668 images, including 12,936 training samples from 751 identities, and 19,732 testing samples from 750 identities, respectively. The DukeMTMC-reID dataset is consisted of 1,812 identities captured from 8 different cameras, in which 16,522 images from 702 identities are used as training samples, 2,228 images of another 702 identities are used as queries, and the remaining 17,661 noisy images are also used for the gallery set. The CUHK03 dataset contains 13,164 images of 1,467 identities, in which samples of 1,367 identities are randomly chosen for training, and the samples of remaining identities are used for testing.

Implementation. In our implementation, we first resize the input images into 256×128, then followed by a random cropping and flipping for data augmentation. The batch size is 32, the learning rate is $\tau = 0.01$ and decayed by 0.1 at every 10 epochs. The weight parameters are set as $\alpha = \beta = 0.1$ and $\varpi = 0.01$, and the margin parameter is chosen as $m = 1.0$. Once the the network is trained, we simply use it to extract features from the testing images and formulate the person Re-ID as a nearest neighbor search problem.

4.2. Ablation Study

Variants. To evaluate how much our method improves the final results, we design 15 experiments on each dataset, as shown in Table 1, which can well support the following conclusions: 1) The multi-task learning framework is more effective than the single-task learning framework in learning discriminative features; 2) The improved triplet loss is superior than the basic triplet loss in supervising the feature learning; 3) The attention subnetwork can slightly improve the network's representation capability; 4) The consistent attention regularizer can guide the attention subnetwork to better explore the foreground regions of input images. As a result, we incorporate our three contributions in a multi-task learning framework to learn a discriminative feature representation for person Re-ID. In the next paragraph, we will explain the above conclusions in detail.

For clarity, we try to check the above conclusions based on the performances on the Market1501 dataset using the single-query evaluation. To evaluate how much the multi-task learning framework outperforms the single-task learning framework, we can compare the experimental results as listed in indexes 1, 2 and 3; indexes 1, 4 and 5; indexes 6, 7 and 8; indexes 6, 9 and 10; indexes 11, 12 and 13; and indexes 11, 14 and 15, from which we can find that the multi-task learning framework can significantly improve the person Re-ID result in all the six situations. Take the experimental results in indexes 1, 2 and 3 for an example, the

Methods	Labeled		Detected	
	Top 1	Top 5	Top 1	Top 5
LDNS [51] (CVPR2016)	62.6	90.5	54.7	84.8
PDC [36] (ICCV2017)	88.7	98.6	78.3	94.8
DLPA [54] (ICCV2017)	85.1	97.6	–	–
SVDNet [37] (ICCV2017)	–	–	81.8	95.2
DCAF [17] (CVPR2017)	74.2	94.3	68.0	91.0
SSM [1] (CVPR2017)	76.6	94.6	72.7	92.4
DPFL [6] (CVPR2017)	86.7	82.8	82.0	78.1
JLML [19](IJCAI2017)	83.2	98.0	80.6	96.9
PRGP [39] (CVPR2018)	91.7	98.2	–	–
DGRW [30] (CVPR2018)	94.9	98.7	–	–
BraidNet [44] (CVPR2018)	88.2	98.7	85.9	98.5
AACN [48] (CVPR2018)	91.4	98.9	89.5	97.7
GCSL [4] (CVPR2018)	90.2	98.5	88.8	97.2
SGGNN [31] (ECCV2018)	95.3	99.1	–	–
PN-GAN [26] (ECCV2018)	79.8	96.2	–	–
Our Method	96.9	99.6	93.2	99.2

Table 2. The matching rates(%) comparison with the state-of-the-art methods on the CUHK03 dataset, in which '–' means they do not report the corresponding results.

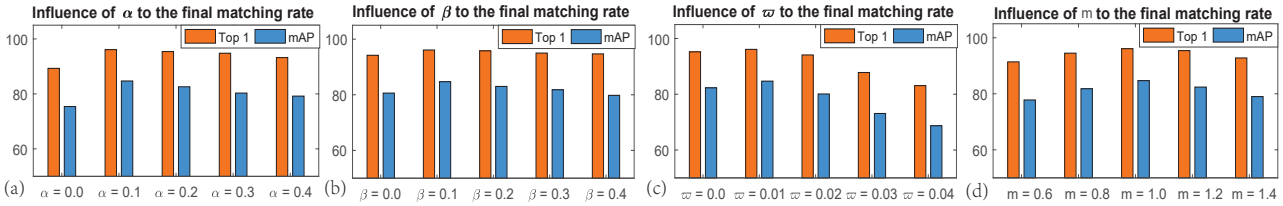

Figure 6. Influences of different parameter settings to the final matching rates. Specifically, we compare the Top 1 and mAP performances of our method on the Market1501 dataset using the single-query evaluation, in which the detailed influences of α, β, ϖ and m are illustrated in (a) to (d), respectively.

Methods	Single-query		Multi-query	
	Top 1	mAP	Top 1	mAP
LDNS [51] (CVPR2016)	61.0	35.6	71.6	46.0
PDC [36] (ICCV2017)	84.1	63.4	–	–
SVDNet [37] (ICCV2017)	82.3	62.1	–	–
DLPA [54] (ICCV2017)	81.0	63.4	–	–
DPFL [6] (CVPR2017)	88.6	72.6	92.3	80.7
PRGP [39] (CVPR2018)	81.2	–	–	–
MLFN [2] (CVPR2018)	90.0	74.3	92.3	82.4
HA-CAN [20] (CVPR2018)	91.2	75.7	93.8	82.8
DGRW [30] (CVPR2018)	92.7	82.5	–	–
DuATM [32] (CVPR2018)	91.4	76.6	–	–
MGCAN [34] (CVPR2018)	83.8	74.3	–	–
BraidNet [44] (CVPR2018)	83.7	69.5	–	–
AACN [48] (CVPR2018)	85.9	66.9	76.8	59.3
GCSL [4] (CVPR2018)	93.5	81.6	–	–
PCB [38] (ECCV2018)	93.8	81.6	–	–
SGGNN [31] (ECCV2018)	92.3	82.8	–	–
PN-GAN [26] (ECCV2018)	89.4	72.6	92.9	80.2
MGN [42] (ACM MM2018)	95.7	86.9	96.9	90.7
Our Method	96.1	84.7	98.2	87.3

Table 3. The matching rates(%) comparison with the state-of-the-art methods on the Market1501 dataset, in which '–' means they do not report the corresponding results.

Methods	Top 1	Top 5	Top10	mAP
SVDNet [37] (ICCV2017)	75.9	86.4	89.5	56.3
DLPA [54] (ICCV2017)	81.0	63.4	–	–
GAN [58] (ICCV2017)	67.7	–	–	47.1
DPFL [6] (CVPR2017)	79.2	–	–	60.6
MLFN [2] (CVPR2018)	81.0	–	–	62.8
HA-CAN [20] (CVPR2018)	80.5	–	–	60.8
DGRW [30] (CVPR2018)	80.7	88.5	90.8	66.4
DuATM [32] (CVPR2018)	81.8	90.2	–	64.6
BraidNet [44] (CVPR2018)	76.4	–	–	59.5
AACN [48] (CVPR2018)	76.8	–	–	59.3
GCSL [4] (CVPR2018)	84.9	–	–	69.5
PCB [38] (ECCV2018)	83.3	90.5	92.5	69.2
SGGNN [31] (ECCV2018)	81.1	88.4	91.2	68.2
PN-GAN [26] (ECCV2018)	73.6	–	88.8	53.2
MGN [42] (ACM MM2018)	88.7	-	-	78.4
Our Method	86.3	92.3	95.2	73.1

Table 4. The matching rates(%) comparison with the state-of-the-art methods on the DukeMTMC-reID dataset, in which '–' means they do not report the corresponding results.

S+T outperforms S and T by 1.6% and 2.1% in Top 1, and 2.2% and 2.6% in mAP, respectively. For the improvements by our triplet loss, we compare the results between indexes 2 and 4; between indexes 3 and 5; between indexes 7 and 9; between indexes 8 and 10; between 11 and 14; and between 13 and 15, respectively. The results explain that the improved triplet loss is superior than the basic triplet loss in learning discriminative features. For instance, the results obtained by our triplet loss outperform that achieved by the basic triplet loss by 3.1% in Top 1 and 4.1% in mAP, when we compare the performances between indexes 7 and 9. From the results listed in Block 1 (as shown in indexes 1 to 5) and Block 2 (as shown in indexes 5 to 10), we can see that the improvements by the attention subnetwork is insignificant, because it is hard to directly deduce attention from the low-level feature maps. Specifically, the improvements are only 0.3%, 0.1%, 0.3%, 0.5% and 0.5% in Top 1, and 0.2%, 0.1%, 0.4%, 0.8% and 0.3% in mAP, when we compare the corresponding results between Block 1 and Block 2, respectively. When the consistent attention regularizer is used to help deduce attention, the results can be significantly improved. Specifically, the improvements are 1.5%, 1.8%, 2.7%, 3.1% and 2.2% in Top 1, and 2.4%, 2.4%, 3.2%, 2.6% and 5.2% in mAP, when we compare the corresponding results between Block 2 and Block 3 (as shown in indexes 11 to 15), respectively.

Parameters. As in most of the deep learning methods, the performance of our method is also highly dependent on the weight parameters α, β and ϖ, and the margin parameter m. In order to clarify this influence, we design four sets of experiments to evaluate how the parameter setting effects the final person Re-ID performance. Specifically, we only change one parameter and keep the others fixed in each set of experiments, so as to evaluate how the varying parameter effects the final performance. For simplicity, we conduct the experiments on the Market1501 datasets and evaluate the results using the single-query evaluation. The results are shown in Figure 6, from which we find that: 1) The experimental results are robust to α, β and m, in which a large variation range is allowed to maintain the final person Re-ID performance in a relatively high level. 2) The experimental results are slight sensitive to ϖ, because the sparsity is hard to control in the training process. If ϖ is large, some of the useful information may be filtered out, therefore the person Re-ID performance will be seriously affected. If ϖ is small, the ability of feature selection will be weaken, which is also not benefit to further improve the final performance. Taking the two situations into account, we prefer to choose a small ϖ in our experiments.

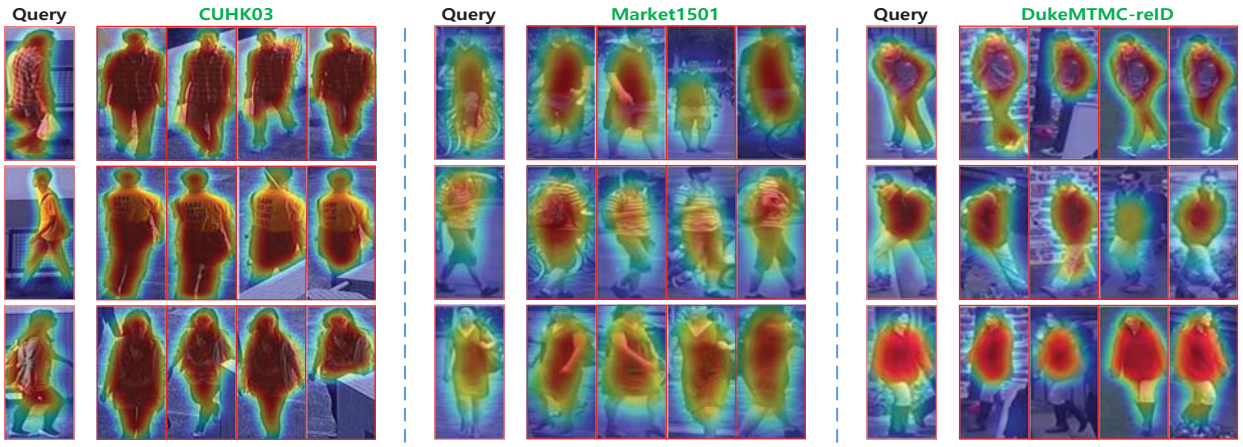

Figure 7. Visualization of the averaged heat maps on the CUHK03, Market1501 and DukeMTMC-reID datasets. From the results we can see that the network can focus on foreground regions at the lower layers by using the consistent attention regularizer.

Losses	CUHK03		Market1501		DukeMTMC	
	Top 1	Top 5	Top 1	mAP	Top 1	mAP
BT	88.6	97.2	92.1	78.6	83.5	70.4
DT	90.3	98.2	93.5	79.8	84.2	70.9
ST	92.8	98.6	94.2	80.3	85.0	71.5
Our Triplet	96.9	99.6	96.1	84.7	86.3	73.1

Table 5. Results of four different triplet losses on three benchmark datasets, in which 'BT' denotes the basic triplet loss, 'DT' means the dual triplet loss and 'ST' indicates the symmetric triplet loss.

Visualization. Our consistent attention regularizer can effectively keep these foreground masks similar, which are deduced from the low-level, mid-level and high-level feature maps, respectively. As a result, our network will focus its attention on foreground regions at the lower layers. We visualize the averaged heat maps on the three datasets, as shown in Figure 7, from which we can find that most of the network's attention has been focused on the foreground regions across the lower to higher layers. Therefore, the resulting features will be very robust to target misalignment and background clutter.

4.3. Comparison Results

Firstly, we compare our method with many state-of-the-art competitors on the CUHK03, Market1501 and DukeMTMC-reID datasets, as shown in Table 2 to Table 4. From the result we can see that: 1) Our method has achieved the best result on the CUHK03 dataset, in which it outperforms the previous best performed SGGNN [31] by 1.6% in Top 1; 2) Our method performs closely to MGN [42] on the Market1501 and DukeMTMC-reID datasets, in which our method is better in Top 1 and the MGN is better in mAP. The reason comes from two aspects: 1) Our network is much lighter, while the MGN needs to take three part branch networks to extract features; 2) Our triplet loss doesn't use any hard mimining strategy, while the MGN further applies the batchhard triplet loss [11] improve the final results. From this point of view, our method can achieve a competitive result in a very simple yet effective way.

Secondly, we compare the performances of four different triplet losses, as shown in Table 5, on the three datasets. From the results we can conclude that: 1) The dual triplet loss outperforms the basic triplet loss, and the symmetric triplet loss outperforms the dual triplet loss on all the three datasets, which indicate that it is an effective way to revise the gradient back-propagation in minimizing the intra-class distances. 2) Our triplet loss outperforms the symmetric triplet loss on all the three datasets, because it doesn't need to introduce any additional algorithm to help update weights in the training process.

5. Conclusion

In this paper, we propose a simple yet effective feed-forward attention network to learn discriminative features from the foreground regions for person Re-ID. Specifically, a novel consistent attention regularizer is designed to drive the foreground masks similar, which are deduced from the low-level, mid-level and high-level feature maps, respectively. As a result, the network will focus on the foreground regions at the lower layers, and the network can effectively deal with the target misalignment and background clutter at the higher layers. Besides, a novel triplet loss is introduced to enhance the feature learning capability, which can jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Extensive experimental results on the Market1501, DukeMTMC-reID and CUHK03 datasets have shown that our method outperforms most of the state-of-the-art approaches.

Acknowledgement

This work is jointly supported by the National Key Research and Development Program of China under Grant No. 2017YFA0700800, and the National Natural Science Foundation of China Grant No. 61629301.

References

[1] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, July 2017. 6

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, volume 1, page 2, 2018. 7

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, June 2018. 2

[4] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, pages 8649–8658, 2018. 6, 7

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *AAAI*, pages 3988–3994, 2017. 1, 2

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *CVPR*, pages 2590–2600, 2017. 6, 7

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, June 2018. 2

[8] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 48(10):2993–3003, 2015. 1, 2, 4

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 3

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 8

[12] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, pages 1173–1182, 2016. 2

[13] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 2

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3

[15] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gkmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, June 2018. 2

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[17] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017. 2, 3, 6

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 6

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *AAAI*, pages 2194–2200, 2017. 6

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, June 2018. 1, 2, 7

[21] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3

[22] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, June 2018. 1

[23] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. *arXiv preprint arXiv:1807.11164*, 2018. 2

[24] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 2

[25] Wenjie Pei, Tadas Baltrušaitis, David MJ Tax, and Louis-Philippe Morency. Temporal attention-gated model for robust sequence classification. In *CVPR*, pages 820–829. IEEE, 2017. 2

[26] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, September 2018. 6, 7

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 6

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv preprint arXiv:1801.04381*, 2018. 2

[29] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, June 2018. 1

[30] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, pages 2265–2274, 2018. 6, 7

[31] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, September 2018. 6, 7, 8

[32] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, June 2018. 7

[33] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, page 1470. IEEE, 2003. 1

[34] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, June 2018. 1, 2, 7

[35] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, pages 5552–5561, 2017. 2

[36] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3980–3989. IEEE, 2017. 6, 7

[37] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, Oct 2017. 6, 7

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, September 2018. 1, 2, 7

[39] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, pages 5794–5803, 2018. 1, 2, 6, 7

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[41] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 2

[42] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282. ACM, 2018. 7, 8

[43] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367. Citeseer, 2010. 1

[44] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, pages 1470–1478, 2018. 6, 7

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, June 2018. 2

[46] Xing Wei, Yue Zhang, Yihong Gong, and Nanning Zheng. Kernelized subspace pooling for deep local descriptors. In *CVPR*, June 2018. 1

[47] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016. 2

[48] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018. 6, 7

[49] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, Oct 2017. 2

[50] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *ICCV*, Oct 2017. 1

[51] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016. 6, 7

[52] Shun Zhang, Yihong Gong, Jia-Bin Huang, Jongwoo Lim, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. Tracking persons-of-interest via adaptive discriminative features. In *ECCV*, pages 415–433. Springer, 2016. 4

[53] Shizhou Zhang, Qi Zhang, Xing Wei, Yanning Zhang, and Yong Xia. Person re-identification with triplet focal loss. *IEEE Access*, 6:78092–78099, 2018. 1

[54] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, Oct 2017. 1, 2, 6, 7

[55] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, volume 6, 2017. 2

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2, 6

[57] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015. 2

[58] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *TOMM*, 14(1):13, 2017. 1, 2, 7

[59] Sanping Zhou, Jinjun Wang, Deyu Meng, Yudong Liang, Yihong Gong, and Nanning Zheng. Discriminative feature learning with foreground attention for person re-identification. *TIP*, 2019. 1

[60] Sanping Zhou, Jinjun Wang, Deyu Meng, Xiaomeng Xin, Yubing Li, Yihong Gong, and Nanning Zheng. Deep self-paced learning for person re-identification. *PR*, 76:739–751, 2018. 2

[61] Sanping Zhou, Jinjun Wang, Rui Shi, Qiqi Hou, Yihong Gong, and Nanning Zheng. Large margin learning in set-to-set similarity comparison for person reidentification. *TMM*, 20(3):593–604, 2017. 1

[62] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person reidentification. In *CVPR*, volume 6, 2017. 1, 4

[63] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785. IEEE, 2017. 2