SafeNet: Scale-normalization and Anchor-based Feature Extraction Network for Person Re-identification

Kun Yuan[1,3], **Qian Zhang**[2], **Chang Huang**[2], **Shiming Xiang**[1,3], **Chunhong Pan**[1]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] Horizon Robotics
[3] University of Chinese Academy of Sciences
{kun.yuan, smxiang, chpan}@nlpr.ia.ac.cn, {qian01.zhang, chang.huang}@hobot.cc

Abstract

Person Re-identification (ReID) is a challenging retrieval task that requires matching a person's image across non-overlapping camera views. The quality of fulfilling this task is largely determined on the robustness of the features that are used to describe the person. In this paper, we show the advantage of jointly utilizing multi-scale abstract information to learn powerful features over full body and parts. A scale normalization module is proposed to balance different scales through residual-based integration. To exploit the information hidden in non-rigid body parts, we propose an anchor-based method to capture the local contents by stacking convolutions of kernels with various aspect ratios, which focus on different spatial distributions. Finally, a well-defined framework is constructed for simultaneously learning the representations of both full body and parts. Extensive experiments conducted on current challenging large-scale person ReID datasets, including Market1501, CUHK03 and DukeMTMC, demonstrate that our proposed method achieves the state-of-the-art results.

1 Introduction

Person re-identification (ReID) aims to search for the same person across different cameras with a given query image. It has attracted much attention in recent years due to its importance in many practical applications, such as security and protection monitoring in public area and content-based image retrieval [Li *et al.*, 2017a]. Despite of years of efforts, it still has many challenges, such as the large variations in person pose, illumination variance, domain gaps among different camera views, and background clutters.

The overall motivation of ReID is to obtain a location-invariant and view-free representation or learn a matching distance metric across domains of two disjoint camera views. Therefore, most existing studies typically focus on either feature representation [Kviatkovsky *et al.*, 2013; Liao *et al.*, 2015; Li *et al.*, 2017a] or matching distance metrics [Chen *et al.*, 2016; Koestinger *et al.*, 2012; Liao *et al.*, 2015]. For most ReID tasks, the acquired person images are often indistinct. Thus how to exploit the *local* detailed features is still an open problem. Most ReID methods use predefined grids to encode local structural information by different image decomposition schemes, such as horizontal stripes [Kviatkovsky *et al.*, 2013], body parts [Farenzena *et al.*, 2010], and patches [Liao *et al.*, 2015]. For example, some methods [Shi *et al.*, 2016; Geng *et al.*, 2016] split the input person image into square overlapping patches from top to bottom, and learn discriminative feature representations in local regions. However, these rigid-based methods often fail to achieve satisfactory performances from images with noisy backgrounds or body occlusion. The performances might be further degraded due to the misaligned person images which are caused by the pose variations and imperfect pedestrian detectors. In these cases, the predefined rigid grids may fail to capture correct correspondences between two pedestrian images, leading to low generalization and bad stability.

In contrast to *local* features, deep neural networks favor intrinsically in learning *global* feature representations. Recently, the methods in the family of deep global feature learning [Qian *et al.*, 2017; Li *et al.*, 2017a; Lin *et al.*, 2017b; Geng *et al.*, 2016] have shown great potential on large-scale person ReID datasets. But most popular networks [Xiao *et al.*, 2016; Zheng *et al.*, 2016; Sun *et al.*, 2017] typically stack single-scale layers to generate the representation feature. However, these methods pay less attention to some fine-grained attributes that are very useful to distinguish the pedestrian pairs with small inter-class variations. Because of the down-sample operations, small scale visual cues, such as bags, shoes and hats, might be ignored, leading to the missing of these fine-grained attributes. Thus in our opinion, these models might not be the best choice for pedestrian feature learning.

We suggest that either local or global feature learning *alone* is suboptimal. This is motivated by the human visual system that leverages both global (contextual) and local (saliency) information concurrently [Navon, 1977]. To better utilize these two types of representations, we propose a Scale-normalization and Anchor-based Feature Extraction Network (SafeNet) that learns full body and parts jointly. Given a pedestrian image, feature maps with various scales will be extracted from different convolutional layers of the backbone network. Generally, shallow layers contain much more low-level information and preserve the details of small objects. Deeper layers pay more attention to high-level information, such as gestures and clothing styles. To appropriately utilize

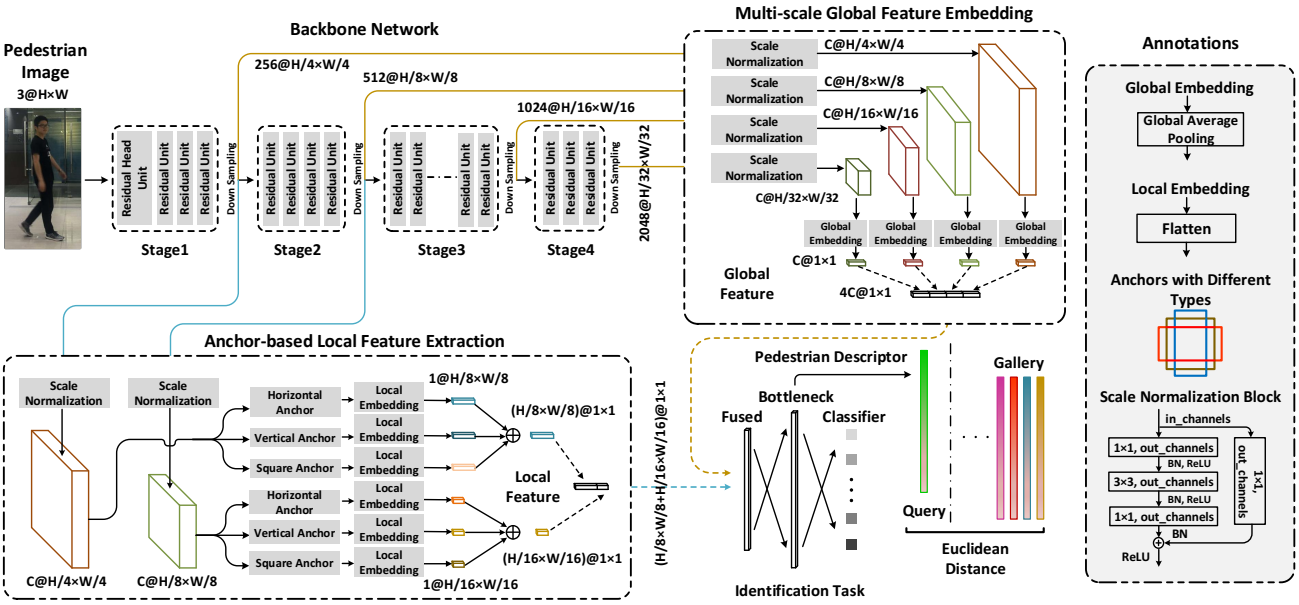

Figure 1: Illustration of the network architecture. The proposed SafeNet consists of four components: the residual-based backbone network, the global representation with the multi-scale global feature embedding part, the local representation with the anchor-based local feature extraction part and the fusion of full body and body parts for person identification tasks.

multiple scales with uneven channels, we propose a scale normalization module to balance the contributions of each scale. The global embedding is applied to the normalized features, resulting in the *global* representation. To obtain the *local* one, we propose various types of convolution kernels with different aspect ratios, called anchors, to slide across the feature maps. In this way, objects with different spatial distributions can be learned. By applying local embedding, we can obtain the local representation. Last, the global and local features are fused to form the pedestrian descriptor. During test, the Euclidean metric is adopted to compute the distance between the L2 normalized person representations for ReID.

The contributions of this paper are summarized as follows: 1)We propose a Scale-normalization and Anchor-based Feature Extraction Network (SafeNet) to effectively utilize multi-scale information and enhance the visual context for better feature representation of pedestrians; 2)Instead of using rigid grid-based methods to obtain local features, we propose to use anchors with different aspect ratios to localize area-of-interest adaptively; 3)We integrate the representation learning processes of global (full body) and local (body parts) into a unified framework. Moreover, our method can be expediently extended through replacing of the backbone network. Experimental results show that the fused pedestrian representations greatly improve the performance of person ReID.

2 Related Work

Typical person ReID methods focus on two key points: developing a powerful feature representation or learning an effective metric to make the same person be close and different ones far away. Here we mainly review the related methods.

There are many research efforts for developing better features that are partially invariant to lighting, pose, and viewpoint variations. Various features have been applied to person ReID, including color histograms [Farenzena *et al.*, 2010] and their variants local binary patterns (LBP) [Koestinger *et al.*, 2012], Gabor features [Li and Wang, 2013], and other visual appearance or contextual cues. But hand-crafted features can be easily affected by illumination variance and suffer from low generalization and bad stability.

Deep learning approaches for person ReID tend to learn person representation and similarity (distance) metric jointly. Given a pair of person images, some work learns image representations through pair-wise contrastive loss [Zheng *et al.*, 2013] or triplet ranking loss [Cheng *et al.*, 2016; Hermans *et al.*, 2017; Wang *et al.*, 2014], and use Euclidean metric for comparison. Due to current large-scale person ReID datasets, the ID-discriminative embedding feature have shown great potentials. [Xiao *et al.*, 2016] proposed the domain guided dropout to learn features over multiple datasets simultaneously with identity classification loss. [Zheng *et al.*, 2016] learned the identity discriminative embedding feature for video-based person ReID. However, most existing methods only consider layer-by-layer single-scale information and ignore abundant multi-scale information.

To obtain part-based representation for person ReID, Some deep learning methods [Cheng *et al.*, 2016; Shi *et al.*, 2016] used predefined rigid grids as body parts, where each part is fed into an individual branch. Different from them, our CNN model improves the classical models in two ways. First, compared with the models which only consider the single-scale feature representation, we jointly utilize multi-scale information through balance combination. Second, we propose to capture the local context knowledge by stacking convolutions of kernels with various aspect ratios. The treatment yields a mechanism that is able to explore and exploit the information hidden in the non-rigid body parts. A well-defined learning framework is constructed for simultaneously learning the rep-

resentations of the both body and those of the body parts.

3 Our Method

3.1 Problem Definition

We assume a set of n training images $\Phi = \{\boldsymbol{I}_i\}_{i=1}^{n}$ with the corresponding identity labels as $\Psi = \{y_i\}_{i=1}^{n}$. These training images capture the visual appearance of n_{id} (where $y_i \in [1, \cdots, n_{id}]$) different people under non-overlapping camera views. The focus of this approach is to learn powerful feature representations about a pedestrian in order to optimize person ReID under significant viewing condition changes across locations. The overall framework of the proposed method is shown in Figure 1. It is built upon three kinds of complementary designs detailed in the following: 1) Multi-scale global feature representations based on scale normalization; 2) Anchor-based local feature extraction; 3) Joint training of global and local representation.

3.2 Multi-scale Global Feature Embedding

For person ReID, the most important cues are visual attribute knowledge, such as colors and types of clothes. However, they have large variations in scale, shape and position, such as the hat/sunglasses at small local scale and the cloth color at larger scale. Directly using bottom-to-up single-scale convolution and pooling may not be effective to handle these complex variations. Especially, with the increasing number of layers, the small visual regions, such as hats, will be easily missed in top layers. To better learn these diverse visual cues, we propose to use multi-scale information.

As shown in Figure 1, given an input pedestrian image $\boldsymbol{I}_i$, the ResNet [He *et al.*, 2016] is adopted to generate feature maps with different scales. For each residual block which consists of stacked residual units, there are many layers producing output maps with the same size and we say these layers are in the same network *stage*. Following general operations [Qian *et al.*, 2017], only the last layer of each stage is chosen as the feature maps in particular stage. Specifically, we denote the output of these last residual units as $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_4\}$, and denote that they have the strides of $\{4, 8, 16, 32\}$ pixels with respect to the input image. These feature maps have obvious difference in the amount of channels of $\{256, 512, 1024, 2048\}$ respectively. We suggest that the straightforward concatenation of these features is inappropriate due to the uneven contributions for the person attributes. Experimental results in Section 4.3 also prove that.

To solve this problem, features maps should be normalized into the same channels with proper integration. However, single convolutional layer may not learn the transformation process adequately. Meanwhile, naive stacking layers will degrade the value of gradients in deep layers. Accordingly, we propose a scale normalization module that consists of trunk branch and shortcut. Similar to the idea in residual learning, if the trunk branch can be constructed as channel mapping, the performances should be no worse than the shortcut without it. Thus we modify the output of scale normalization module as:

$$\mathbf{g}^C = \mathcal{F}(\mathbf{g}, \{W_t\}) + W_s\mathbf{g}, \tag{1}$$

where $\mathcal{F}$ is the non-linear mapping function, W_t are the parameters of truck branch, W_s are the parameters of shortcut matrix that performs a linear projection, $\mathbf{g}^C$ is the normalized feature maps with the number of channels of C. The detailed structure of the scale normalization module is shown in the annotations of Figure 1. The truck branch tries to learn powerful representations after the number changes of channels. It consists of 3 convolutional layers with kernel sizes of $\{1\times1, 3\times3, 1\times1\}$, following by BatchNormalization and ReLU. The shortcut is a convolutional layer with 1×1 kernel. It performs the maintenance of details. And its outputs are added to the outputs of the trunk branch.

By using the scale normalization module to each scale, the normalized feature maps can be obtained and noted as $\{\mathbf{g}_1^C, \mathbf{g}_2^C, \mathbf{g}_3^C, \mathbf{g}_4^C\}$, resulting in the balance of scales. Then global average pooling is adopted to the normalized feature maps as the global embedding, resulting in feature vectors. Finally, features from different scales are concatenated to form the global representation with the length of $4C$.

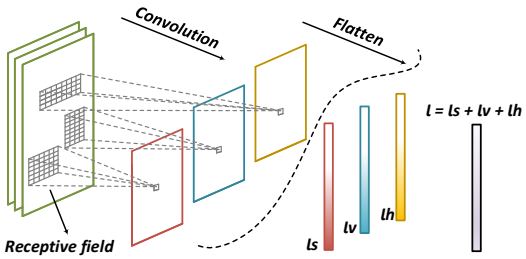

Figure 2: Details of the local feature extraction. Anchors with various aspect ratios are adopted, corresponding to different receptive fields.

3.3 Anchor-based Local Feature Extraction

It has been shown in many methods [Cheng *et al.*, 2016; Shi *et al.*, 2016; Liao *et al.*, 2015] that part-based representation is useful for person ReID. But most part-based methods roughly decompose the extracted pedestrian into predefined rigid body parts which approximately correspond to head, shoulder, upper-body, upper-leg and lower-leg, respectively. However, due to the unsatisfying pedestrian detection algorithms and large pose variations, the methods of using rigid body parts for local feature extraction is not the optimal solution. This motivates us to learn the local representation according to the spatial distributions of objects.

We assume that objects with particular spatial distributions will be activated in varying degrees under different receptive fields [Theunissen *et al.*, 2001]. Based on this assumption, we design a anchor-based local feature extraction structure. As shown in Figure 2, for a given feature map, convolutional features are computed with different kernels, corresponding to different receptive fields. We define various types of convolutional kernels, called anchors, with aspect ratios of {1:2, 2:1, 1:1}. These aspect ratios represents objects with spatial distributions of horizontal, vertical and square, respectively. Besides, it is not necessary to have multi-scale anchors because anchors will slide over more than one scale.

The unified structure is shown in Figure 1. First, the scale normalization modules are adopted to obtain local information. Distinctively, the normalized feature maps are denoted by $\mathbf{l}^C$. Second, the anchors are applied to each scale, resulting

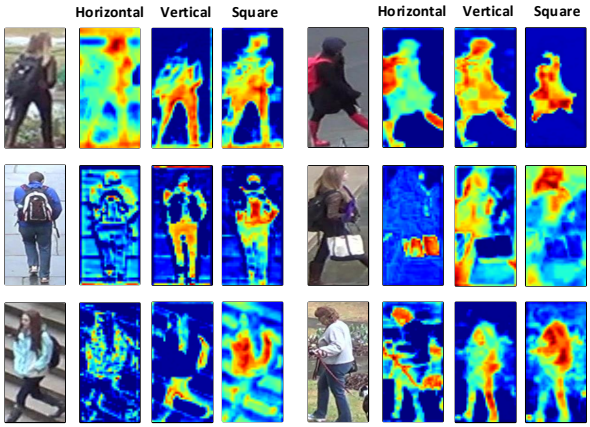

Figure 3: Heat maps of the learned local features. Red regions are highly activated. Blue regions attract less attentions. Diverse anchors have different area-of-interests. Generally, backgrounds attract less attentions than pedestrians.

in different local activations. It can be denoted by:

$$\mathbf{l}_k = \sum_{n=1}^{C} \mathbf{l}^n \otimes \text{anchor}_k^n, \quad k \in \{\text{horizontal}, \text{vertical}, \text{square}\}, \tag{2}$$

where n is the number of channel and the corresponding anchor, k stands for the type of anchors and $\otimes$ denotes the convolution operation. Local information can be obtained by the summation of the learned feature maps $\mathbf{l} = \mathbf{l}_h + \mathbf{l}_v + \mathbf{l}_s$. Finally, $\mathbf{l}$ is flattened to a vector as the local feature for each scale. For multiple scales, features of different scales are concatenated to form the local representation. Since shallow layers have small strides with respect to the input image, we consider that local details will be kept most in these layers.

We demonstrate some visualization results of $\{\mathbf{l}_h, \mathbf{l}_v, \mathbf{l}_s\}$ in Figure 3. These prove our assumption and show the activated regions under different receptive fields. As can be seen, the horizontal anchors focus on horizontally distributed objects, such as walking legs, bags and skirts. Regions with legs, arms and hairs will be activated with the vertical anchors. As for the square ones, body trucks, knapsacks and heads will be noticed. Notably, this is not unalterable since there exist various types of poses and clothes. What's more, most backgrounds can be eliminated, resulting in more robust local feature learning. These anchors have different concerns and come together to form the local representations.

3.4 Joint Training of Local and Global Features

The global and local branches have complementary strengths for learning discriminative pedestrian descriptors. To leverage these complementaries, we jointly train the whole network to predict person identity for both part-based and global feature learning. The joint training is built upon a multi-class person identification task. Structurally, a two-layer fully connected block is used, where features are named as *fused*, *bottleneck* and *classifier*. The *bottleneck* is used to ease the influence of straightforward concatenation as well as enhance the ability of representation. The Softmax loss is used:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(W_{y_i}^T f_i + b_{y_i})}{\sum_{j=1}^{N_c} \exp(W_j^T f_i + b_j)} \tag{3}$$

Anchors	kernel	stride	padding
horizontal	(4, 8)	(2, 2)	(1, 3)
vertical	(8, 4)	(2, 2)	(3, 1)
square	(6, 6)	(2, 2)	(2, 2)

Table 1: Details of Three Types of Anchors

where f_i is the *classifier* feature of the i-th sample, y_i is the identity, N is the number of samples, W_j and b_j is the weight and bias of the classifier for the j-th identity, respectively .

4 Experiments

4.1 Implementation Details

The proposed model is built on PyTorch framework. The backbone network is the ResNet-50/101 model pre-tained on ImageNet. We follow common dataset augmentation strategies with different scales and aspect ratios to train our model. We randomly crop each resized image with scale in the interval [0.64, 1.0] and aspect ratio in [2,3]. Then the cropped patch is resized to 256×128. To prevent overfitting, randomly horizontal flip with a probability of 0.5 is also applied. During testing phase, images are simply resized to 256×128. Before feeding the input image to the network, we subtract the mean value and divide the standard deviation. The SGD optimizer is used to minimize the loss function $\mathcal{L}$ given in Eq 3. The initial learning rate for the backbone network is 0.01 and the rest parts are 0.001. The total number of training epochs for all conducted experiments are set to 80. The mini-batch size is set to 128 for all experiments.

As for the local anchors, we give the details in Table 1. Experiments in Section 4.3 prove the effectiveness of the designed sizes. Visualization results also indicate these anchors can capture most person parts adaptively.

4.2 Datasets and Evaluation Results

Datasets and Evaluation Protocol

We use three widely used person ReID benchmark datasets, CUHK03 [Li *et al.*, 2014], Market-1501 [Zheng *et al.*, 2015] and DukeMTMC-ReID [Ristani *et al.*, 2016], for performance evaluations. All the datasets contain a set of persons, each of whom has several images captured by different cameras. The following is brief descriptions of these datasets:

CUHK03: It contains $1,360$ identities and $12,164$ person images which are captured by six surveillance cameras in campus. Each identity is captured by two disjoint cameras. It offers a 20-split dividing, resulting in a training set with 1260 ids and a testing set with 100 ids. The average of 20-split is adopted as final results.

Market-1501: It contains $1,501$ identities which are captured by six manually set cameras. There are $32,368$ pedestrian images in total. As official setting, 751 ids are used for training and the rest 750 ids are used for testing. The query contains 3368 images.

DukeMTMC-ReID: Constructed from the multi-camera tracking dataset DukeMTMC, it contains $1,812$ identities. 702 identities are used as the training set and the remaining $1,110$ identities as the testing set. It contains 36411 images in total. 2228 images are used as queries.

Methods	Labeled			Detected		
	R1	R5	R10	R1	R5	R10
XQDA [Liao *et al.*, 2015]	52.2	82.2	92.1	46.3	78.9	88.6
MLAPG [Liao and Li, 2015]	57.9	87.1	94.7	51.2	83.6	92.1
DNS [Zhang *et al.*, 2016a]	62.5	90.1	94.8	54.7	84.7	94.8
SS-SVM [Zhang *et al.*, 2016b]	57.0	85.7	94.3	51.2	80.8	89.6
EDM [Shi *et al.*, 2016]	61.3	88.9	96.4	52.1	82.9	91.8
OL-MANS [Zhou *et al.*, 2017a]	61.7	88.4	95.2	62.7	87.6	93.8
MSCAN [Li *et al.*, 2017a]	74.2	94.3	97.5	68.0	91.0	95.4
MuDeep [Qian *et al.*, 2017]	76.9	96.1	98.4	75.6	94.4	97.5
JLML [Li *et al.*, 2017b]	83.2	**98.0**	**99.4**	80.6	**96.9**	**98.7**
Single scale	76.0	94.3	97.8	71.2	91.3	95.3
Multiple scales (Global only)	84.3	93.6	96.1	82.3	92.0	95.8
SafeNet (ResNet-50)	**86.3**	97.8	99.1	**84.0**	96.3	98.2
SafeNet (ResNet-101)	**87.2**	**98.1**	**99.3**	**84.1**	**97.2**	**98.4**

Table 2: Evaluation on CUHK03 using labeled pedestrian bounding boxes and automatic detections by DPM.

Methods	Single Query		Multi Query	
	R1	mAP	R1	mAP
LDNS [Zhang *et al.*, 2016a]	55.4	29.8	71.6	46.0
Gated S-CNN [Varior *et al.*, 2016]	65.9	39.6	76.0	48.5
P2S [Zhou *et al.*, 2017b]	70.7	44.3	85.8	55.7
CRAFT [Chen *et al.*, 2018]	71.8	45.5	79.7	54.3
CADL [Lin *et al.*, 2017a]	73.8	47.1	80.8	55.6
MSCAN [Li *et al.*, 2017a]	80.3	57.5	86.8	66.7
LSRO [Zheng *et al.*, 2017b]	83.9	66.1	88.4	76.1
DeepTransfer[Geng *et al.*, 2016]	83.7	65.5	89.6	73.8
TriNet [Hermans *et al.*, 2017]	84.9	69.1	90.5	76.4
JLML [Li *et al.*, 2017b]	85.1	65.5	89.7	74.5
Single scale	81.7	59.8	87.6	67.4
Multiple scales (Global only)	86.4	68.3	91.5	78.0
SafeNet (ResNet-50)	**90.2**	**72.7**	**93.1**	**81.6**
SafeNet (ResNet-101)	**91.5**	**75.4**	**94.7**	**84.2**

Table 3: Comparison with state-of-the-art results on Market-1501. Both single and multi query results are reported.

We follow the standard evaluation protocol. The cumulative matching characteristics (CMC) at rank-1 and mean average precision (mAP) are adopted for performance evaluation on Market-1501 and DukeMTMC-ReID. The precisions of rank-1, rank-5, and rank-10 are reported for CUHK03. Following most related works, the evaluation on CHUK03 and DukeMTMC-ReID is performed under single query. Both single and multiple query settings are used for Market-1501. Re-ranking is *not* adopted in our method.

Evaluation on CUHK03

Table 2 shows the comparisons against 9 existing methods on CUHK03. It is evident that our method outperforms existing methods in all categories on both labelled and detected bounding boxes, surpassing the 2nd best performers MuDeep and JLML on corresponding labelled images in R1 by 9.4% and 3.1%. Compared with MSCAN that also utilize multi-scale, the performance of our is better, improving R1 by 12.1% and 13.7% for labeled and detected. Both labeled and detected ones indicate the robustness and competitiveness of our approach in mining local and global discriminative features. Higher results can be obtained through replacing the backbone network with deeper one. We demonstrate the query matching results in Figure 4. Remarkably, our methods can handle the situations that suffer from occlusion or the badly-lighted environment.

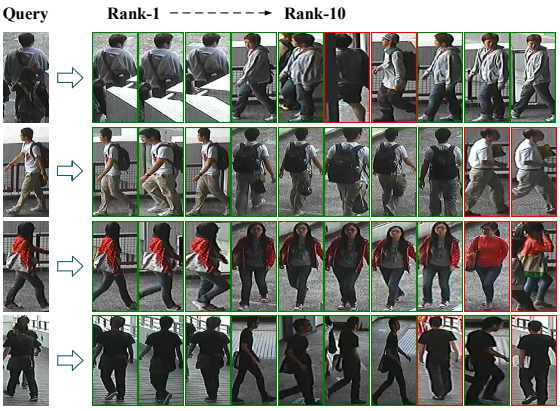

Figure 4: Pedestrian matching results on the CUHK03 test dataset. Green box denotes the correct matching result.

Methods	R1	mAP
LOMO+XQDA [Liao *et al.*, 2015]	30.8	17.1
LSRO [Zheng *et al.*, 2017b]	67.7	47.1
AttIDNet [Lin *et al.*, 2017b]	70.7	51.2
PAN [Zheng *et al.*, 2017a]	71.6	51.5
ACRN [Schumann and Stiefelhagen, 2017]	72.6	52.0
SVDNet [Sun *et al.*, 2017]	76.7	56.8
DPFL [Chen *et al.*, 2018]	79.2	**60.6**
Single scale	73.9	51.9
Multiple scales (Global only)	80.8	52.5
SafeNet (ResNet-50)	**82.7**	57.0
SafeNet (ResNet-101)	**83.6**	**58.4**

Table 4: Quantitative comparison with state-of-the-art methods on DukeMTMC-ReID.

Evaluation on Market-1501

As depicted in Table 3, the proposed method achieves the SOTA results on Market-1501. Compared with existing multi-scale methods, such as MSCAN and TriNet, the performance of our model is substantially better, e.g. improving R1 by 9.9% and 5.3% for single query. Compared with full body-based network Gated S-CNN, the proposed network structure can better capture pedestrian features. Concretely, our method improves R1 by 5.1% to JLML in single query and 2.4% in multiple query. By using deeper backbone network ResNet101, we can achieve much higher R1 by 91.5%, and mAP by 75.4% in single query and R1 by 94.7% in multiple query. These results show consistent superiority and robustness of the proposed model over the existing methods.

Evaluation on DukeMTMC-ReID

The comparison with state-of-the-art methods is depicted in Table 4, our method also outperforms all approaches. Compared with the AttIDNet which utilizes additional attributes information, our method outperforms in Rank-1 by 12.0%. Compared to DPFL which learns multi-scale feature using image pyramid inputs, our method achieve an improvement of 3.5% by Rank-1 under much less computations. By using ResNet-101, the higher results can be obtained 83.6% by Rank-1. This also demonstrates the robustness and expandability of our method when replacing the backbone network.

scales	R1	R5	R10	mAP
$\mathbf{g}_4$	81.7	92.3	95.0	59.8
$\mathbf{g}_4, \mathbf{g}_3$	84.0	94.1	96.3	63.8
$\mathbf{g}_4, \mathbf{g}_3, \mathbf{g}_2$	86.4	94.6	96.6	68.0
$\mathbf{g}_4, \mathbf{g}_3, \mathbf{g}_2, \mathbf{g}_1$	86.4	94.4	96.5	68.3
$\mathbf{g}_4^C, \mathbf{g}_3^C, \mathbf{g}_2^C, \mathbf{g}_1^C$ (C=128)	87.9	95.2	97.0	70.0
$\mathbf{g}_4^C, \mathbf{g}_3^C, \mathbf{g}_2^C, \mathbf{g}_1^C$ (C=256)	88.1	95.2	96.8	71.1
$\mathbf{g}_4^C, \mathbf{g}_3^C, \mathbf{g}_2^C, \mathbf{g}_1^C$ (C=512)	**88.7**	**95.5**	**97.1**	**71.6**
multi-global (C=256) + $\mathbf{l}_1$	89.8	96.1	**97.5**	**72.9**
multi-global (C=256) + $\mathbf{l}_1, \mathbf{l}_2$	**90.2**	**96.2**	**97.5**	72.7
multi-global (C=256) + $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$	88.8	95.3	97.3	70.8
multi-global (C=256) + $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3, \mathbf{l}_4$	88.1	95.4	97.1	67.9

Table 5: Ablation Analysis of Different Components.

4.3 Ablation Analysis and Discussion

We further evaluate several variants of our network to verify the effectiveness of each individual component. Without loss of generality, the ablation study is performed on Market-1501 under single query, using the same settings in Sec. 4.1.

Effectiveness of Multi-scale Information

The proposed method relies on multi-scale contents. To evaluate the contributions of different components, we test different combinations and parameters C. As shown in Table 5, compared with single scale, the four-scale representation can improve the result by 4.7% on R1 and 7.5% on mAP. But $\mathbf{g}_4$ will domain the pedestrian feature by straightforward fusion because of the 2048-dimension. By adopting scale normalization, different scales are balanced during fusion, resulting in boosting of R1 to 88.7%. Larger C leads to better performance but consumes higher computation complexity. Experiments show the advantages of multi-scale features over single scale counterparts in ReID. Moreover, direct multi-scale concatenation may result in suboptimal optimization. Scale normalization is an effective approach to handle this problem.

Effectiveness of Local Part-based Context

To evaluate the local parts, we test under different scales. In Table 5 (below), better results can be obtained in comparison with global branch by 2.1% on R1 and 1.6% on mAP. Results in Table 2 and Table 4 also suggest our improvements. The main reason is that the pedestrians detected by DPM consist much more background and the part-based representation can better reduce the influences of background clutters. The best results are gotten when only the first two scales are used. It can be explained that shallow layers have small strides with respect to the input image and keep more details about local information. Besides, anchors with different sizes are tested, including large ones {(6,12), (12,6), (9,9)}, middle ones {(4,8), (8,4), (6,6)} and small ones {(2,4), (4,2), (3,3)}. Results in Figure 5(a) show the middle ones are more effective. We also make additional experiments to verify the importance of each anchor. The ablation of horizontal one leads to R1 by 89.5% and mAP by 72.0%. The absence of vertical one achieves R1 by 88.9% and mAP by 71.5%. The lack of square one leads to R1 by 88.2% and mAP by 71.0%. These prove each anchor is indispensable and complementary.

Choice of Final Pedestrian Descriptor

In our method, there exist four candidates to be the pedestrian descriptor: 1) the global feature, 2) the fused feature, 3) the bottleneck and 4) the classifier. As shown in Figure 5(b), the fused feature outperforms the global one because of additional local information. The bottleneck can ease the influence caused by fusion, resulting in more discriminative representations and an improvement by 1.9% on R1. However, when the classifier is used, it suffers a 1.6% mAP drop. One possible reason is that the identification loss makes the features near the classification layer focus more on the difference of training identities. Such feature might be discriminative for identities in training, but is not useful for the unseen identities during test. These comparison motivate us to use the bottleneck feature as the final pedestrian descriptor.

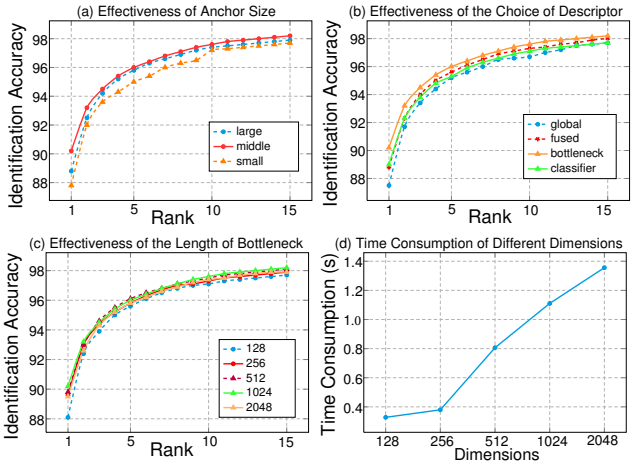

Figure 5: CMC and time consumption curves.

Effectiveness of Bottleneck

As shown in Figure 5(c), the dimension of the bottleneck affects the performance slightly. Relatively, the dimensions of $\{512, 1024\}$ achieve better performance. Time consumption of computing the pairwise distance matrix between query and gallery are also reported in Figure 5(d). It indicates that fewer dimensions lead to insufficient expression and larger ones suffer from huge time-consumption during the pedestrian retrieval. On balance, the dimension of 1024 is selected.

5 Conclusion

In this paper, we have proposed the SafeNet for person ReID which demonstrates the advantage of jointly utilizing multi-scale abstract information to learn powerful features over full body and parts. The proposed scale normalization module is an effective way to balance different scales with residual-based integration. By designing anchors with different aspect ratios, the local context knowledge hidden in non-rigid body parts can be obtained as the complement to the global feature. The well-defined framework can simultaneously learn the representations of the both body and those of the body parts. Extensive comparative evaluations on current challenging large-scale person ReID datasets demonstrate that the proposed method achieves the state-of-the-art results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 91646207 and 61573352.

References

[Chen *et al.*, 2016] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016.

[Chen *et al.*, 2018] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 40(2):392–408, 2018.

[Cheng *et al.*, 2016] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[Geng *et al.*, 2016] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint:1611.05244*, 2016.

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

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

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[Kviatkovsky *et al.*, 2013] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *TPAMI*, 35(7):1622–1634, 2013.

[Li and Wang, 2013] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[Li *et al.*, 2014] 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.

[Li *et al.*, 2017a] 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*, 2017.

[Li *et al.*, 2017b] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, pages 770–778, 2017.

[Liao and Li, 2015] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[Lin *et al.*, 2017a] Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, volume 6, 2017.

[Lin *et al.*, 2017b] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint:1703.07220*, 2017.

[Navon, 1977] David Navon. Forest before trees: The precedence of global features in visual perception. *CP*, 9(3):353–383, 1977.

[Qian *et al.*, 2017] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. *ICCV*, 2017.

[Ristani *et al.*, 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016.

[Schumann and Stiefelhagen, 2017] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPR Workshop*, 2017.

[Shi *et al.*, 2016] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, pages 732–748, 2016.

[Sun *et al.*, 2017] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, pages 3820–3828, 2017.

[Theunissen *et al.*, 2001] F. E. Theunissen, S. V. David, N. C. Singh, A Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *NCNS*, 12(3):289–316, 2001.

[Varior *et al.*, 2016] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016.

[Wang *et al.*, 2014] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *ICCV*, pages 1386–1393, 2014.

[Xiao *et al.*, 2016] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *ICCV*, pages 1249–1258, 2016.

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

[Zhang *et al.*, 2016b] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, pages 1278–1287, 2016.

[Zheng *et al.*, 2013] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 35(3):653–668, 2013.

[Zheng *et al.*, 2015] 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.

[Zheng *et al.*, 2016] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[Zheng *et al.*, 2017a] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint:1707.00408*, 2017.

[Zheng *et al.*, 2017b] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zhou *et al.*, 2017a] Jiahuan Zhou, Pei Yu, Wei Tang, and Ying Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *CVPR*, 2017.

[Zhou *et al.*, 2017b] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, volume 6, 2017.