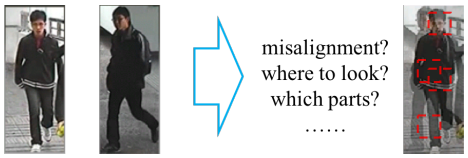
Multi-scale Deep Learning Architectures for Person Re-identification

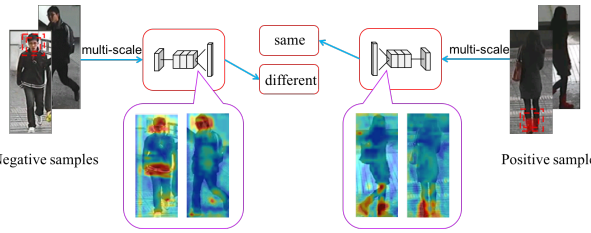
Xuelin Qian[1] Yanwei Fu[2,5,*] Yu-Gang Jiang[1,3] Tao Xiang[4] Xiangyang Xue[1,2]
[1]Shanghai Key Lab of Intelligent Info. Processing, School of Computer Science, Fudan University;
[2]School of Data Science, Fudan University; [3]Tencent AI Lab;
[4]Queen Mary University of London; [5]University of Technology Sydney;
{15110240002,yanweifu,ygj,xyxue}@fudan.edu.cn; t.xiang@qmul.ac.uk

Abstract

Person Re-identification (re-id) aims to match people across non-overlapping camera views in a public space. It is a challenging problem because many people captured in surveillance videos wear similar clothes. Consequently, the differences in their appearance are often subtle and only detectable at the right location and scales. Existing re-id models, particularly the recently proposed deep learning based ones match people at a single scale. In contrast, in this paper, a novel multi-scale deep learning model is proposed. Our model is able to learn deep discriminative feature representations at different scales and automatically determine the most suitable scales for matching. The importance of different spatial locations for extracting discriminative features is also learned explicitly. Experiments are carried out to demonstrate that the proposed model outperforms the state-of-the art on a number of benchmarks.



(a) The cognitive process a human may take to match people



(b) Our model aims to imitate the human cognitive process

Figure 1. Multi-scale learning is adopted by our MuDeep to learn discriminative features at different spatial scales and locations.

1. Introduction

Person re-identification (re-id) is defined as the task of matching two pedestrian images crossing non-overlapping camera views [11]. It plays an important role in a number of applications in video surveillance, including multi-camera tracking [2, 41], crowd counting [3, 10], and multi-camera activity analysis [54, 53]. Person re-id is extremely challenging and remains unsolved for a number of reasons. First, in different camera views, one person's appearance often changes dramatically caused by the variances in body pose, camera viewpoints, occlusion and illumination conditions. Second, in a public space, many people often wear very similar clothes (e.g., dark coats in winter). The differences that can be used to tell them apart are often subtle, which could be the global, e.g., one person is bulkier than the other, or local, e.g., the two people wear different shoes.

Early re-id methods use hand-crafted features for person appearance representation and employ distance metric learning models as matching functions. They focus on either designing cross-view robust features [7, 13, 26, 62, 32], or learning robust distance metrics [31, 33, 63, 25, 49, 60, 38, 32], or both [24, 32, 57, 59]. Recently, inspired by the success of convolutional neural networks (CNN) in many computer vision problems, deep CNN architectures [1, 48, 27, 51, 44, 50, 4] have been widely used for person re-id. Using a deep model, the tasks of feature representation learning and distance metric learning are tackled jointly in a single end-to-end model. The state-of-the-art re-id models are mostly based on deep learning; deep re-id is thus the focus of this paper.

Learning discriminative feature representation is the key objective of a deep re-id model. These features need to be computed at multiple scales. More specifically, some people can be easily distinguished by some global features such as gender and body build, whilst for some others, detecting local images patches corresponding to, say a handbag of a particular color or the type of shoes, would be critical for

*Corresponding Author

distinguishing two otherwise very similarly-looking people. The optimal matching results are thus only obtainable when features at different scales are computed and combined. Such a multi-scale matching process is likely also adopted by most humans when it comes to re-id. In particular, humans typically compare two images from coarse to fine. Taking the two images in Fig. 1(a) as an example. At the coarse level, the color and textual information of clothes are very similar; humans would thus go down to finer scales to notice the subtle local differences (*e.g*. the hairstyle, shoe, and white stripes on the jacket of the person on the left) to reach the conclusion that these are two different people.

However, most existing re-id models compute features at a single scale and ignore the factor that people are often only distinguishable at the right spatial locations and scales. Existing models typically adopt multi-branch deep convolutional neural networks (CNNs). Each domain has a corresponding branch which consists of multiple convolutional/pooling layers followed by fully connected (FC) layers. The final FC layer is used as input to pairwise verification or triplet ranking losses to learn a joint embedding space where people's appearance from different camera views can be compared. However, recent efforts [9, 39] on visualizing what each layer of a CNN actually learns reveal that higher-layers of the network capture more abstract semantic concepts at global scales with less spatial information. When it reaches the FC layers, the information at finer and local scales has been lost and cannot be recovered. This means that the existing deep re-id architectures are unsuitable for the multi-scale person matching.

In this work, we propose a novel multi-scale deep learning model (MuDeep) for re-id which aims to learn discriminative feature representations at multiple scales with automatically determined scale weighting for combining them (see Fig. 1(b)). More specifically, our MuDeep network architecture is based on a Siamese network but critically has the ability to learn features at different scales and evaluating their importance for cross-camera matching. This is achieved by introducing two novel layers: *multi-scale stream layers* that extract images features by analyzing the person images in multi-scale; and *saliency-based learning fusion layer,* which selectively learns to fuse the data streams of multi-scale and generate the more discriminative features of each branch in MuDeep. The multi-scale data can implicitly serve as a way of augmenting the training data. In addition to the verification loss used by many previous deep re-id models, we introduce a pair of classification losses at the middle layers of our network, in order to strongly supervise multi-scale features learning.

2. Related Work

Deep re-id models Various deep learning architectures have been proposed to either address visual variances of pose and viewpoint [27], learn better relative distances of triplet training samples [6], or learn better similarity metrics of any pairs [1]. To have enough training samples, [48] built upon inception module a single deep network and is trained on multiple datasets; to address the specific person re-id task, the neural network will be adapted to a single dataset by a domain guided dropout algorithm. More recently, an extension of the siamese network has been studied for person re-id [50]. Pairwise and triplet comparison objectives have been utilized to combine several sub-networks to form a network for person re-id in [51]. Similarly, [4] employed triplet loss to integrate multi-channel parts-based CNN models. To resolve the problem of large variations, [44] proposed a moderate positive sample mining method to train CNN. However, none of the models developed is capable of multi-scale feature computation as our model.

More specifically, the proposed deep re-id model differs from related existing models in several aspects. (1) our MuDeep generalizes the convolutional layers with multi-scale strategy and proposed multi-scale stream layers and saliency-based learning fusion layer, which is different from the ideas of combing multiple sub-networks [51] or channels [4] with pairwise or triplet loss. (2) Comparing with [48], our MuDeep are simplified, refined and flexible enough to be trained from scratch on either large-scale dataset (*e.g*. CUHK03) or medium-sized dataset (*e.g*. CUHK01). Our experiments show that without using any extra data, the performance of our MuDeep is $12.41\%/4.27\%$ higher than that of [48] on CUHK01/CUHK03 dataset. (3) We improve the architecture of [1] by introducing two novel layers to implement multi-scale and saliency-based learning mechanisms. Our experiment results validate that the novel layers lead to much better performance than [1].

Multi-scale re-id The idea of multi-scale learning for re-id was first exploited in [29]. However, the definition of scale is different: It was defined as different levels of resolution rather than the global-to-local supporting region as in ours. Therefore, despite similarity between terminology, very different problems are tackled in these two works. The only multi-scale deep re-id work that we are aware of is [36]. Compared to our model, the model in [36] is rather primitive and naive: Different down-sampled versions of the input image are fed into shallower sub-networks to extract features at different resolution and scale. These sub-networks are combined with a deeper main network for feature fusion. With an explicit network for each scale, this network becomes computationally very expensive. In addition, no scale weighting can be learned automatically and no spatial importance of features can be modeled as in ours.

Deep saliency modelling Visual saliency has been studied extensively [19, 20]. It is typically defined in a bottom-up process. In contrast, attention mechanism [5] works

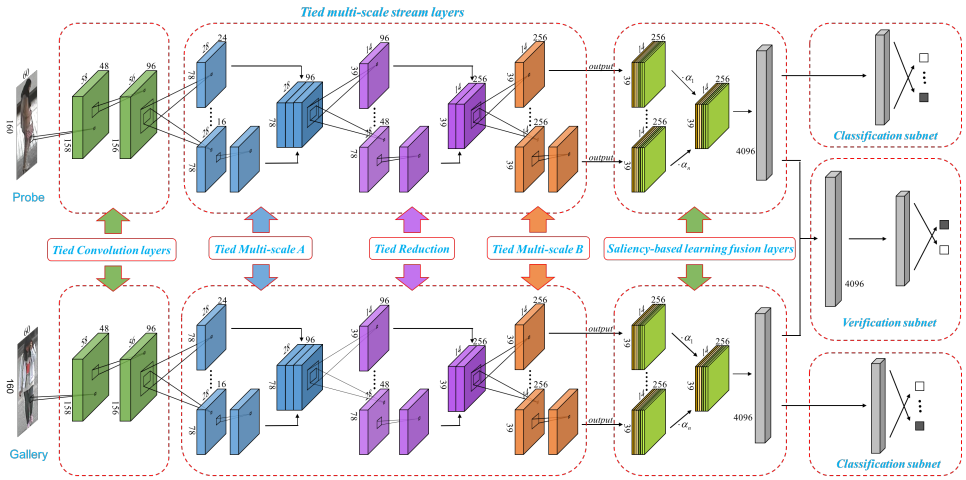


Figure 2. Overview of MuDeep architecture.

Layers	Stream id	number@size	output
Multi-scale-A	1	1@3 × 3 × 96 AF –24@1 × 1 × 96 CF	78 × 28 × 96
	2	24@1 × 1 × 96 CF	
	3	16@1 × 1 × 96 CF – 24@3 × 3 × 96 CF	
	4	16@1 × 1 × 96 CF –24@3 × 3 × 96 CF– 24@3 × 3 × 24 CF	
Reduction	1	1@3 × 3 × 96 MF*	39 × 14 × 256
	2	96@3 × 3 × 96 CF*	
	3	48@1 × 1 × 96 CF – 56@3 × 3 × 48 CF– 64@3 × 3 × 56 CF*	
Multi-scale-B	1	256@1 × 1 × 256 CF	39 × 14 × 256
	2	64@1 × 1 × 256 CF –128@1 × 3 × 64 CF–256@3 × 1 × 128 CF	39 × 14 × 256
	3	64@1 × 1 × 256 CF – 64@1 × 3 × 64 CF –128@3 × 1 × 64CF– 128@1 × 3 × 128 CF– 256@3 × 1 × 128 CF	39 × 14 × 256
	4	1@3 × 3 × 256 AF* –256@1 × 1 × 256 CF	39 × 14 × 256

Table 1. The parameters of tied multi-scale stream layers of MuDeep. Note that (1) number@size indicates the number and the size of filters. (2) * means the stride of corresponding filters is 2; the stride of other filters is 1. We add 1 padding to the side of input data stream if the corresponding side of C-filters is 3. (3) CF, AF, MF indicate the C-filters, A-filters and M-filters respectively. A-filter is the average pooling filter.

in a top-down way and allows for salient features to dynamically come to the front as needed. Recently, deep soft attention modeling has received increasing interest as a means to attend to/focus on local salient regions for computing deep features [43, 42, 58, 35]. In this work, we use saliency-based learning strategy in a saliency-based learning fusion layer to exploit both visual saliency and attention mechanism. Specifically, with the multi-scale stream layers, the saliency features of multiple scales are computed in multi-channel (*e.g.* in a bottom-up way); and a per channel weighting layer is introduced to automatically discover the most discriminative feature channels with their associated scale and locations. Comparing with [35] which adopts a spatial attention model, our model is much compact and can be learned from scratch on a small re-id dataset. When comparing the two models, our model, despite being much smaller, yields overall slightly better performance: on CUHK-01 dataset ours is 8% lower than that of [35] but on the more challenging CUHK-03(detected) we got around 10% improvement over that of [35]. Such a simple saliency learning architecture is shown to be very effective in our experiments.

Our contributions are as follows: (1) A novel multi-scale representation learning architecture is introduced into the deep learning architectures for person re-id tasks. (2) We propose a saliency-based learning fusion layer which can learn to weight important scales in the data streams in a saliency-based learning strategy. We evaluate our model on a number of benchmark datasets, and the experiments show that our models can outperform state-of-the-art deep re-id models, often by a significant margin.

3. Multi-scale Deep Architecture (MuDeep)

Problem Definition. Typically, person re-id is formulated only as a verification task [24, 32, 57, 59]. In contrast, this

paper formulates person re-id into two tasks: classification [55, 56] and verification [1, 48]. Specifically, given a pair of person images, our framework will categorize them (1) either as the "same person" or "different persons" class, and (2) predict the person's identity.

Architecture Overview. As shown in Fig. 2, MuDeep has two branches to process each of image pairs. It consists of five components: *tied convolutional layers*, *multi-scale stream layers* (Sec. 3.1), *saliency-based learning fusion layer* (Sec. 3.2), *verification subnet* and *classification subnet* (Sec. 3.3). Note that after each convolutional layer or fully connected layer, batch normalization [18] is used before the ReLU activation.

Preprocessing by convolutional layers. The input pairs are firstly pre-processed by two convolutional layers with the filters (C-filters) of $48@3 \times 3 \times 3$ and $96@3 \times 3 \times 48$; furthermore, the generated feature maps are fed into a max-pooling layer with filter size (M-filter) as $1@3 \times 3 \times 96$ to reduce both length and width by half. The weights of these layers are tied across two branches, in order to enforce the filters to learn the visual patterns shared by both branches.

3.1. Multi-scale stream layers

We propose multi-scale stream layers to analyze data streams in multi-scale. The multi-scale data can implicitly serve as a way of augmenting the training data. Different from the standard Inception structure [47], all of these layers share weights between the corresponding stream of two branches; however, within each two data streams of the same branch, the parameters are not tied. The parameters of these layers are shown in Tab. 1; and please refer to Supplementary Material for the visualization of these layers.

Multi-scale-A layer analyses the data stream with the size 1×1, 3×3 and 5×5 of the receptive field. Furthermore, in order to increase both depth and width of this layer, we split the filter size of 5×5 into two 3×3 streams cascaded (*i.e.* stream-4 and stream-3 in Table 1). The weights of each stream are also tied with the corresponding stream in another branch. Such a design is in general inspired by, and yet different from Inception architectures [46, 47, 45]. The key difference lies in the factors that the weights are not tied between any two streams from the same branch, but are tied between two corresponding streams of different branches.

Reduction layer further passes the data streams in multi-scale, and halves the width and height of feature maps, which should be, in principle, reduced from 78×28 to 39×14. We thus employ Reduction layer to *gradually* decrease the size of feature representations as illustrated in Table 1, in order to avoid representation bottlenecks. Here we follow the design principle of "avoid representational bottlenecks" [47]. In contrast to directly use max-pooling layer for decreasing feature map size, our ablation study shows that the Reduction layer, if replaced by max-pooling layer, will leads to more than 10% absolute points lower than the reported results of Rank-1 accuracy on the CUHK01 dataset [28]. Again, the weights of each filter here are tied for paired streams.

Multi-scale-B layer serves as the last stage of high-level features extraction for the multiple scales of 1×1, 3×3 and 5×5. Besides splitting the 5×5 stream into two 3×3 streams cascaded (*i.e.* stream-4 and stream-3 in Table 1). We can further decompose the 3×3 C-filters into one 1×3 C-filter followed by 3×1 C-filter [45]. This leads to several benefits, including reducing the computation cost on 3×3 C-filters, further increasing the depth of this component, and being capable of extracting asymmetric features from the receptive field. We still tie the weights of each filter.

3.2. Saliency-based learning fusion layer

This layer is proposed to fuse the outputs of multi-scale stream layers. Intuitively, with the output processed by previous layers, the resulting data channels have redundant information: Some channels may capture relative important information of persons, whilst others may only model the background context. The saliency-based learning strategy is thus utilized here to automatically discover and emphasize the channels that had extracted highly discriminative patterns, such as the information of head, body, arms, clothing, bags and so on, as illustrated in Fig. 3. Thus, we assume $\mathbf{F}_{i\star}$ represents the input feature maps of $i-$th stream $(1 \leq i \leq 4)$ in each branch and $\mathbf{F}_{ij}$ represents the $j-$th channel of $\mathbf{F}_{i\star}$, *i.e.* $(1 \leq j \leq 256)$ and $\mathbf{F}_{ij} \in \mathbb{R}^{39\times14}$. The output feature maps denoted as $\mathbf{G}$ will fuse the four streams; $\mathbf{G}_j$ represents the $j-$th channel map of $\mathbf{G}$, which is computed by:

$$\mathbf{G}_j = \sum_{i=1}^{4} \mathbf{F}_{ij} \cdot \alpha_{ij} \quad (1 \leq j \leq 256) \tag{1}$$

where α_{ij} is the scalar for $j-$th channel of $\mathbf{F}_{i\star}$; and the saliency-weighted vector $\alpha_{i\star}$ is learned to account for the importance of each channel of stream $\mathbf{F}_{i\star}$; $\alpha_{i\star}$ is also tied.

A fully connected layer is appended after saliency-based learning fusion layer, which extracts features of 4096-dimensions of each image. The idea of this design is 1) to concentrate the saliency-based learned features and reduce dimensions, and 2) to increase the efficiency of testing.

3.3. Subnets for person Re-id

Verification subnet accepts feature pairs extracted by previous layers as input, and calculate distance with *feature difference layer*, which followed by a fully connected layer of 512 neurons with 2 softmax outputs. The output indicates the probability of "same person" or "different persons". Feature difference layer is employed here to fuse the features of two branches and compute the distance between two images. We denote the output features

of two branches as $\mathbf{G}^1$ and $\mathbf{G}^2$ respectively. The feature difference layer computes the difference $\mathbf{D}$ as $\mathbf{D} = \left[\mathbf{G}^1 - \mathbf{G}^2\right] . * \left[\mathbf{G}^1 - \mathbf{G}^2\right]$. Note that (1) '$.*$' indicates the element-wise multiplication; the idea behind using element-wise subtraction is that if an input image pair is labelled "same person", the features generated by multi-scale stream layers and saliency-based learning fusion layers should be similar; in other words, the output values of feature difference layer should be close to zero; otherwise, the values have different responses. (2) We empirically compare the performance of two difference layer operations including $\left[\mathbf{G}^1 - \mathbf{G}^2\right] . * \left[\mathbf{G}^1 - \mathbf{G}^2\right]$ and $\left[\mathbf{G}^1 - \mathbf{G}^2\right]$. Our experiment shows that the former achieves 2.2% higher performance than the latter on Rank-1 accuracy on CUHK01.

Classification subnet In order to learn strong discriminative features for appearance representation, we add classification subnet following saliency-based learning fusion layers of each branch. The classification subnet is learned to classify images with different pedestrian identities. After extracting 4096-D features in saliency-based learning fusion layers, a softmax with *N* output neurons are connected, where *N* denotes the number of pedestrian identities.

4. Experiments

4.1. Datasets and settings

Datasets. The proposed method is evaluated on three widely used datasets, *i.e*. CUHK03 [27], CUHK01 [28] and VIPeR [12]. The CUHK03 dataset includes $14,096$ images of $1,467$ pedestrians, captured by six camera views. Each person has 4.8 images on average. Two types of person images are provided [27]: manually labelled pedestrian bounding boxes (labelled) and bounding boxes automatically detected by the deformable-part-model detector [8] (detected). The manually labelled images generally are of higher quality than those detected images. We use the settings of both manually *labelled* and automatically *detected* person images on the standard splits in [27] and report the results in Sec. 4.2 and Sec. 4.3 respectively. CUHK01 dataset has 971 identities with 2 images per person of each camera view. As in [28], we use as probe the images from camera A and take those from camera B as gallery. Out of all data, we select randomly 100 identities as the test set. The remaining identities for training and validation. The experiments are repeated over 10 trials. For all the experiments, we train our models from the scratch. VIPeR has 632 pedestrian pairs in two views with only one image per person of each view. We split the dataset and half of pedestrian pairs for training and the left for testing as in [1] over 10 trials. In addition, we also evaluate proposed method on two video-based re-id datasets, *i.e.*, iLIDS-VID dataset [52] and PRID-2011 dataset [15]. The iLIDS-VID dataset contains 300 persons, which are captured by two non-overlapping cameras. The sequences range in length from 23 to 192 frames, with an average number of 73. The PRID-2011 dataset contains 385 persons for camera view A; 749 persons for camera view B, with sequences lengths of 5 to 675 frames. These two camera views have no nonoverlapping. Since the primary focus of this paper is on image-based person re-id, we employ the simplest feature fusion scheme for video re-id: Given a video sequence, we compute features of each frame which are aggregated by max-pooling to form video level representation. In contrast, most of the state-of-the-art video-based re-id methods [40, 37, 52, 23, 30, 22] utilized the RNN models such as LSTM to perform temporal/sequence video feature fusion from each frame.

Experimental settings. On the CUHK03 dataset, in term of training set used, we introduce two specific experimental settings; and we report the results for both settings: (a) **Jointly**: as in [48], under this setting the model is firstly trained with the image set of both labelled and detected CUHK03 images, and for each task, the corresponding image set is used to fine-tune the pre-trained networks. (b) **Exclusively**: for each of the "labelled" and "detected" tasks, we only use the training data from each task without using the training data of the other task.

Implementation details. We implement our model based on the Caffe framework [21] and we make our own implementation for the proposed layers. We follow the training strategy used in [1] to first train the network without classification subnets; secondly, we add the classification subnets and freeze other weights to learn better initialization of the identity classifier; finally we train classification loss and verification loss simultaneously, with a higher loss weight of the former. The training data include positive and negative pedestrian pairs. We augment the data to increase the training set size by 5 times with the method of random 2D translation as in [27]. The negative pairs are randomly sampled as twice the number of positive pairs. We use the stochastic gradient descent algorithm with the mini-batch size of 32. The learning rate is set as 0.001, and gradually decreased by $1/10$ every 50000 iterations. The size of input image pairs is[1] $60 \times 160 \times 3$. Unless specified otherwise, the dropout ratio is set as 0.3. The proposed MuDeep get converged in $9 \sim 12$ hours on re-id dataset on a NVIDIA TITANX GPU. Our MuDeep needs around 7GB GPU memory. Code and models will be made available on the first author's webpage.

Competitors. We compare with the deep learning based methods including DeepReID [27], Imp-Deep [1], En-Deep [55], and G-Dropout [48], Gated_Sia [50], EMD [44], SI-CI [51], and MSTC[2] [36], as well as other non-deep competitors, such as Mid-Filter [62], and XQDA [32], LADF [31],

[1] To make a fair comparision with [1], the input images are resized to $60 \times 160 \times 3$.

[2] We re-implement [36] for evaluation purpose.

eSDC [61], LMNN [16], and LDM [14].

Evaluation metrics. In term of standard evaluation metrics, we report the Rank-1, Rank-5 and Rank-10 accuracy with single-shot setting in our paper. For more detailed results using *Cumulative Matching Characteristics* (CMC) curves, please refer to the Supplementary Material.

4.2. Results on CUHK03-Detected

On the CUHK03-Detected dataset, our results are compared with the state-of-the-art methods in Table 2.

Firstly and most importantly, our best results – MuDeep (jointly) outperforms all the other baselines at all ranks. Particularly, we notice that our results are significantly better than both the methods of using hand-crafted features and the recent deep learned models. This validates the efficacy of our architectures and suggests that the proposed multi-scale and saliency-based learning fusion layer can help extract discriminative features for person re-id.

Secondly, comparing with the Gated_Sia [50] which is an extension of Siamese network with the gating function to selectively emphasize fine common local patterns from data, our result is 7.54% higher at Rank-1 accuracy. This suggests that our framework can better analyze the multi-scale patterns from data than Gated_Sia [50], again thanks to the novel multi-scale stream layers and saliency-based learning fusion layers.

Finally, we further compare our results on both "Jointly" and "Exclusively" settings. The key difference is that in the "jointly" settings, the models are also trained with the data of CUHK03-Labelled, *i.e.* the images with manually labelled pedestrian bounding boxes. As explained in [27], the quality of labelled images is generally better than those of detected images. Thus with more data of higher quality used for training, our model under the "Jointly" setting can indeed beat our model under the "Exclusively" setting. However, the improved margins between these two settings at all Ranks are very small if compared with the margin between our results and the results of the other methods. This suggests that our multi-scale stream layers have efficiently explored and augmented the training data; and with such multi-scale information explored, we can better train our model. Thanks to our multi-scale stream layers, our MuDeep can still achieve good results with less training data, *i.e.* under the Exclusively setting.

4.3. Results on CUHK03-Labelled

The results of CUHK03-Labelled dataset are shown in Table 3 and we can make the following observations.

Firstly, in this setting, our MuDeep still outperforms the other competitors by clear margins on all the rank accuracies. Our result is 4.27% higher than the second best one – G-Dropout[48]. Note that the G-Dropout adopted the domain guided dropout strategies and it utilized much more training data in this task. This further validates that our multi-scale stream layers can augment the training data to exploit more information from medium-scale dataset rather than scaling up the size of training data; and saliency-based learning fusion layers can better fuse the output of multi-scale information which can be used for person re-id.

Secondly, we can draw a similar conclusion as from the CUHK03-Detected results: Our MuDeep with "Jointly" setting is only marginally better than that with "Exclusively" setting. This means that more available related training data can always help improve the performance of deep learning model; and this also validates that our multi-scale stream layers and saliency-based fusion layers can help extract and fuse the multi-scale information and thus cope well with less training data.

Dataset	"Detected"		
Rank	1	5	10
SDALF[7]	4.87	21.17	35.06
eSDC [61]	7.68	21.86	34.96
LMNN [16]	6.25	18.68	29.07
XQDA [32]	46.25	78.90	88.55
LDM [14]	10.92	32.25	48.78
DeepReid [27]	19.89	50.00	64.00
MSTC [36]	55.01	–	–
Imp-Deep [1]	44.96	76.01	83.47
SI-CI [51]	52.17	84.30	92.30
Gated_Sia [50]	68.10	88.10	94.60
EMD [44]	52.09	82.87	91.78
MuDeep (Jointly)	**75.64**	**94.36**	**97.46**
MuDeep (Exclusively)	75.34	94.31	97.40

Table 2. Results of CUHK03-Detected dataset.

Dataset	"Labelled"		
Rank	1	5	10
SDALF[7]	5.60	23.45	36.09
eSDC [61]	8.76	24.07	38.28
LMNN [16]	7.29	21.00	32.06
XQDA [32]	52.20	82.23	92.14
LDM [14]	13.51	40.73	52.13
DeepReid [27]	20.65	51.50	66.50
Imp-Deep [1]	54.74	86.50	93.88
G-Dropout[48]	72.60	92.30*	94.30*
EMD [44]	61.32	88.90	96.44
MuDeep (Jointly)	**76.87**	**96.12**	**98.41**
MuDeep (Exclusively)	76.34	95.96	98.40

Table 3. Results of CUHK03-Labelled dataset. Note that: * represents the results reproduced from [48] with the model trained only by CUHK03 dataset.

4.4. Results on CUHK01 and VIPeR

CUHK01 dataset. Our MuDeep is trained only on CUHK01 dataset without using extra dataset. As listed in

Table 4, our approach obtains 79.01% on Rank-1 accuracy, which can beat all the state-of-the-art; and is 7.21% higher than the second best method [51]. This further shows the advantages of our framework.

VIPeR dataset. This dataset is extremely challenging due to small data size and low resolution. In particular, this dataset has relatively small number of distinct identities and thus the positive pairs for each identity are much less if compared with the other two datasets. Thus for this dataset, our network is initialized by the model pre-trained on CUHK03-Labelled dataset with the "Jointly" setting. The training set of VIPeR dataset is used to fine-tune the pre-trained network. We still use the same network structure except changing the number of neurons on the last layer in the classification subnet. The results on VIPeR dataset are listed in Table 5. The results of our MuDeep remains competitive and outperforms all compared methods.

Qualitative visualization. We give some qualitative results of visualizing the saliency-based learning maps on CUHK01. The visualization of our saliency-based learning maps is computed from saliency-based learning fusion layer in Fig. 3. Given one input pair of images (left of Fig. 3), for each branch, the saliency-based learning fusion layer combines four data streams into a single data stream with selectively learning learned from the saliency of each data stream. The heatmaps of three channels are visualized for each stream and each branch is shown on the right side of Fig. 3. Each row corresponds to one data stream; and each column is for one channel of heatmap of features. The weight α (in Eq. (1)) for each feature channel is learned and updated accordingly with the iteration of the whole network. The three channels illustrated in Fig. 3 have a high reaction (i.e. high values) in the heatmaps on discriminative local patterns of the input image pair. For example, the first and third columns highlight the difference of clothes and body due to the existence of different visual patterns learned by our multi-scale stream layers. Thus these two channels have relative higher α weights, whilst the second column models the background patterns which are less discriminative for the task of person re-id and results in lower α value. Our saliency-based learning fusion layer can automatically learn the optimal α weights from training data.

4.5. Ablation study

Multi-scale stream layers. We compare our multi-scale stream layers with three variants of Inception-v4 [45] on CUHK01 dataset. Specifically, Inception-v4 has Inception A and Inception B modules, both of which are compared against here. Furthermore, we also compare Inception A+B structure which is constructed by connecting Inception A, Reduction, and Inception B modules. The Inception A+B structure is the most similar one to our multi-scale stream layers except that (1) we modify some parameters; (2) the

Rank	1	5	10
KISSME [34, 25]	29.40	60.18	74.44
SDALF[7]	9.90	41.21	56.00
eSDC [61]	22.84	43.89	57.67
LMNN [16]	21.17	49.67	62.47
LDM [14]	26.45	57.68	72.04
DeepReid [27]	27.87	58.20	73.46
G-Dropout [48]	66.60	–	–
MSTC [36]	64.12	–	–
Imp-Deep [1]	65.00	88.70	93.12
SI-CI [51]	71.80	91.35*	95.23*
EMD [44]	69.38	91.03	96.84
MuDeep	**79.01**	**97.00**	**98.96**

Table 4. Results of CUHK01 dataset. *: reported from CMC curves in [51].

Rank	1	5	10
kCCA[34]	30.16	62.69	76.04
Mid-Filter [62]	29.11	52.34	65.95
RPLM [17]	27.00	55.30	69.00
MtMCML [38]	28.83	59.34	75.82
LADF [31]	30.22	64.70	78.92
XQDA [32]	40.00	68.13	80.51
Imp-Deep [1]	34.81	63.61	75.63
G-Dropout [48]	37.70	–	–
MSTC [36]	31.24	–	–
SI-CI [51]	35.76	67.40	83.50
Gated_Sia [50]	37.90	66.90	76.30
EMD [44]	40.91	67.41	79.11
MuDeep	**43.03**	**74.36**	**85.76**

Table 5. Results on the VIPeR dataset

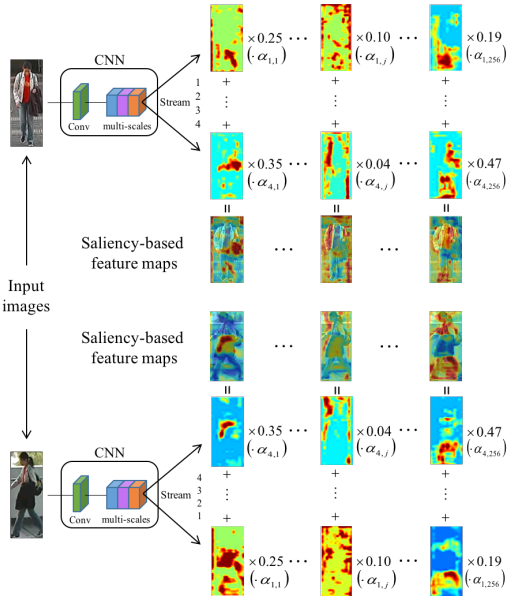


Figure 3. Saliency Map of $\mathbf{G}$ in Eq (1).

weights of our layers are tied between the corresponding streams of two branches. Such weight tieing strategy enforces each paired stream of our two branches to extract the common patterns. With the input of each image pair, we use Inception A, Inception B, and Inception A+B as the based network structure to predict whether this pair is the "same person" or "different persons". The results are compared in Table 6. We can see that our MuDeep architecture has the best performance over the other baselines. This shows that our MuDeep is the most powerful at learning discriminative patterns than the Inceptions variants, since the multi-scale stream layers can more effectively extract multi-scale information and the saliency-based learning fusion layer facilitates the automatic selection of the important feature channels.

Saliency-based learning fusion layer and classification subset. To further investigate the contributions of the fusion layer and the classification subset, we compare three variants of our model with one of or both of the two components removed on CUHK01 dataset. In Table 7, the "– Fusion" denotes our MuDeep without using the fusion layer; and "–ClassNet" indicates our MuDeep without the classification subnet; and "– Fusion – ClasNet" means that MuDeep has neither fusion layer nor classification subnet. The results in Table 7 show that our full model has the best performance over the three variants. We thus conclude that both components help and the combination of the two can further boost the performance.

Rank	1	5	10
Inception A	60.11	85.30	92.44
Inception B	67.31	92.71	97.43
Inception A+B	72.11	91.90	96.45
MuDeep	**79.01**	**97.00**	**98.96**

Table 6. Results of comparing with different inception models on the CUHK01 dataset.

Rank	1	5	10
– Fusion	77.88	96.81	98.21
– ClassNet	76.21	94.47	98.41
– Fusion – ClasNet	74.21	92.10	97.63
MuDeep	**79.01**	**97.00**	**98.96**

Table 7. Results of comparing with the variants of MuDeep on the CUHK01 dataset. Note that "–Fusion" means that saliency-based learning fusion layer is not used; "–ClassNet" indicates that the classification subnet is not used in the corresponding structure.

Dataset	PRID-2011			iLIDS-VID		
Rank	1	5	10	1	5	10
RCNvrid[40]	**70**	**90**	**95**	**58**	**84**	**91**
STA [37]	64	87	90	44	72	84
VR [52]	42	65	78	35	57	68
SRID [23]	35	59	70	25	45	56
AFDA [30]	43	73	85	38	63	73
DTDL [22]	41	70	78	26	48	57
DDC [15]	28	48	55	–	–	–
MuDeep	65	87	93	41	70	83

Table 8. Results on the PRID-2011 and iLIDS-VID datasets

4.6. Further evaluations

Multi-scale vs. Multi-resolution. Due to the often different camera-to-object distances and the resultant object image resolutions, multi-resolution re-id is also interesting on its own and could potentially complement multi-scale re-id. Here a simplest multi-resolution multi-scale re-id model is formulated, by training the proposed multi-scale models at different resolutions followed by fusing the different model outputs. We consider the original resolution (60×160) and a lower one (45×145). The feature fusion is done by concatenation. We found that the model trained at the lower resolution achieves lower results and when the two models are fused, the final performance is around $1-2\%$ lower than the model learned at the original resolution alone. Possible reasons include (1) All three datasets have images of similar resolutions; and (2) more sophisticated multi-resolution learning model is required which can automatically determine the optimal resolution for measuring similarity given a pair of probe/gallery images.

Results on video-based re-id. Our method can be evaluated on iLIDS-VID and PRID-2011 datasets. Particularly, two datasets are randomly split into 50% of persons for training and 50% of persons for testing. We follow the evaluation protocol in [15] on PRID-2011 dataset and only consider the first 200 persons who appear in both cameras. We compared our model with the results reported in [15, 52, 40]. The results are listed in Table 8. These results are higher than those in [15, 52], but lower than those in [40] (only slightly on PRID-2011)[3]. These results are quite encouraging and we expect that if the model is extended to a CNN-RNN model, better performance can be achieved.

5. Conclusion

We have identified the limitations of existing deep re-id models in the lack of multi-scale discriminative feature learning. To overcome the limitation, we have presented a novel deep architecture – MuDeep to exploit the multi-scale and saliency-based learning strategies for re-id. Our model has achieved state-of-the-art performance on several benchmark datasets.

Acknowledgments. This work was supported in part by two NSFC projects ($\#U1611461$ and $\#U1509206$) and one project from STCSM ($\#16JC1420401$).

[3]We also note that our results are better than those of most video-based re-id specialist models listed in Table 1 of [40].

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2, 3, 4.1, 1, 4.2, 4.3, 4.4, 4.4

[2] J. Berclaz, F. Fleuret, and P. Fua. Multi-camera tracking and atypical motion detection with behavioral maps. In *ECCV*, 2008. 1

[3] A. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, pages 545 –551, 2009. 1

[4] D. Cheng, Y. Gong, S. Zhou, JinjunWang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 1, 2

[5] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. In *Nature reviews neuroscience,*, 2002. 2

[6] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. In *Pattern Recognition*, 2015. 2

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 4.2, 4.3, 4.4

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32:1627–1645, 2010. 4.1

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. *CoRR*, 2015. 1

[10] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *CVPR*, 2009. 1

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014. 1

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE PETS Workshop*, 2007. 4.1

[13] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1

[14] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 4.1, 4.2, 4.3, 4.4

[15] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 4.1, 4.6, 4.6

[16] M. Hirzer, P. M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *IEE AVSS*, 2012. 4.1, 4.2, 4.3, 4.4

[17] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 4.4

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[19] L. Itti and C. Koch. Computational modelling of visual attention. *Nat Rev Neurosci*, 2(3):194–203, Mar 2001. 2

[20] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, Jan 2001. 2

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014. 4.1

[22] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4516–4524, 2015. 4.1, 4.6

[23] S. Karanam, Y. Li, and R. J. Radke. Sparse re-id: Block sparsity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2015. 4.1, 4.6

[24] S. Khamis, C. Kuo, V. Singh, V. Shet, and L. Davis. Joint learning for attribute-consistent person re-identification. In *ECCV workshop*, 2014. 1, 3

[25] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 4.4

[26] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE TPAMI*, 2013. 1

[27] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 4.1, 4.2, 4.3, 4.4

[28] W. Li, R. Zhao, and X.Wang. Human re-identification with transferred metric learning. In *ACCV*, 2012. 3.1, 4.1

[29] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, December 2015. 2

[30] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, volume 1, page 2, 2015. 4.1, 4.6

[31] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *ECCV*, 2014. 1, 4.1, 4.4

[32] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 3, 4.1, 4.2, 4.3, 4.4

[33] G. Lisanti, I. Masi, A. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 2014. 1

[34] G. Lisanti, I. Masi, and A. D. Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *ICDSC*, 2014. 4.4, 4.4

[35] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. In *IEEE TIP*, 2016. 2

[36] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *ACM Multimedia*, 2016. 2, 4.1, 2, 4.2, 4.4, 4.4

[37] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3810–3818, 2015. 4.1, 4.6

[38] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. In *IEEE TIP*, 2014. 1, 4.4

[39] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, 2015. 1

[40] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016. 4.1, 4.6, 4.6, 3

[41] T. Mensink, W. Zajdel, and B. Krose. Distributed em learning for appearance based multi-camera tracking. In *ICDSC*, 2007. 1

[42] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 2

[43] P. Sermanet, A. Frome, and E. Real. Attention for fine-grained categorization. *arXiv*, 2014. 2

[44] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei1, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 1, 2, 4.1, 4.2, 4.3, 4.4, 4.4

[45] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In *arxiv*, 2016. 3.1, 4.5

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 3.1

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Zbigniew-Wojna. Rethinking the inception architecture for computer vision. In *arxiv*, 2015. 3.1

[48] X. T, W.Ouyang, H. Li, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1, 2, 3, 4.1, 4.3, 3, 4.4, 4.4

[49] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person reidentification by regularized smoothing kiss metric learning. *IEEE TCSVT*, 2013. 1

[50] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 1, 2, 4.1, 4.2, 4.4

[51] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 1, 2, 4.1, 4.2, 4.4, 4, 4.4

[52] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014. 4.1, 4.6, 4.6

[53] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson. Trajectory analysis and semantic region modeling using a non-parametric bayesian model. In *CVPR*, 2008. 1

[54] X. Wang, K. Tieu, and W. Grimson. Correspondence-free multi-camera activity analysis and scene modeling. In *CVPR*, 2008. 1

[55] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016. 3, 4.1

[56] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 3

[57] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 1, 3

[58] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2

[59] Z. Zhang, Y. Chen, and V. Saligrama. A novel visual word co-occurrence model for person re-identification. In *ECCV workshop*, 2014. 1, 3

[60] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 1

[61] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 4.1, 4.2, 4.3, 4.4

[62] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 1, 4.1, 4.4

[63] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE TPAMI*, 2013. 1