# Leader-based Multi-Scale Attention Deep Architecture for Person Re-identification

Xuelin Qian†, Yanwei Fu†, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue

**Abstract**—Person re-identification (re-id) aims to match people across non-overlapping camera views in a public space. This is a challenging problem because the people captured in surveillance videos often wear similar clothing. Consequently, the differences in their appearance are typically subtle and only detectable at particular locations and scales. In this paper, we propose a deep re-id network (MuDeep) that is composed of two novel types of layers – a multi-scale deep learning layer, and a leader-based attention learning layer. Specifically, the former learns deep discriminative feature representations at different scales, while the latter utilizes the information from multiple scales to lead and determine the optimal weightings for each scale. The importance of different spatial locations for extracting discriminative features is learned explicitly via our leader-based attention learning layer. Extensive experiments are carried out to demonstrate that the proposed MuDeep outperforms the state-of-the-art on a number of benchmarks and has a better generalization ability under a domain generalization setting.

**Index Terms**—Person re-identification, multi-scale deep learning, self-attention, domain generalization

## 1 Introduction

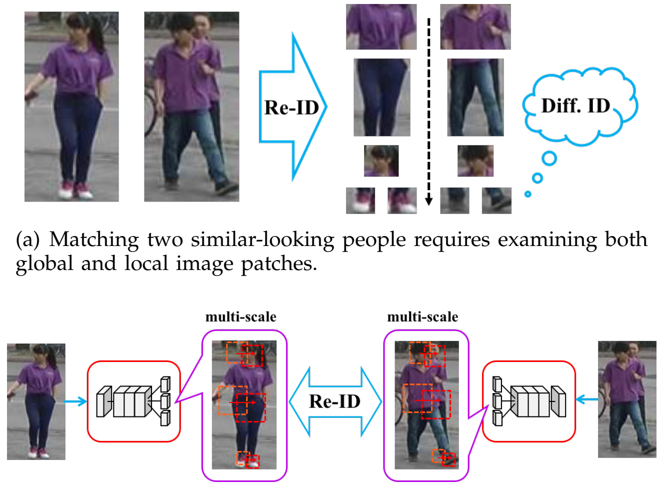
Person re-identification (re-id) is the task of matching two pedestrian images crossing non-overlapping camera views [1]. It plays an important role in a number of applications in video surveillance, including multi-camera tracking [2], [3], crowd counting [4], [5], and multi-camera activity analysis [6], [7]. 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 due to the variances in body poses, camera viewpoints, occlusion and illumination conditions. Second, in a public space, many people often wear very similar clothing (*e.g.*, dark coats in winter). Thus, the differences that can be used to distinguish between people are subtle. These subtle discrepancies exist at different locations in the image and are of different spatial scales. For instance, they could be global, *e.g.*, one person is bulkier than another, or local, *e.g.*, the two people are wearing different shoes.

Early re-id methods used hand-crafted features for person appearance representations and employed distance metric learning models as matching functions. They focused on either designing robust cross-view features [8], [9], [10], [11], [12], or learning robust distance metrics [13], [14], [15], [16], [17], [18], [19], [12], or both [20], [12], [21], [22]. Recently, inspired by the success of convolutional neural networks (CNNs) in many computer vision problems, deep CNN architectures [23], [24], [25], [26], [27], [28], [29], [30] 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.

However, most existing deep re-id models adopt a network architecture originally designed for object category recognition tasks, such as the ImageNet 1K challenge. These tasks are very different from fine-grained instance recognition tasks such as person re-id. For these tasks, it is critical to extract rich features that capture subtle instance differences at different spatial scales and locations. In particular, most existing re-id models compute features at a single scale and ignore the factor that people are often only distinguishable at particular spatial locations and scales. More specifically, these re-id models employ CNNs consisting of multiple convolutional layers for feature extraction. The final feature output of such a CNN is subject to pairwise verification or triplet ranking losses to learn a joint embedding space where the appearance of persons from different camera views is compared. With this type of architecture, features extracted at different layers become progressively more abstract. This occurs because features are extracted using filters with larger receptive fields, corresponding to larger spatial scales. Furthermore, the features computed by the final network layer typically go through a global average pooling operation. This results in the loss of spatial location information about where the features were extracted from the image.

In this work, we argue that the key to learning an effective re-id model lies in computing rich features that represent person appearance at multiple scales. This is because some people can be easily distinguished by some global features such as gender and body build,

- *Yanwei Fu is with the School of Data Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China. Email: yanweifu@fudan.edu.cn. † indicates equal contribution. Yu-Gang Jiang is the corresponding author.*
- *Xuelin Qian, Yu-Gang Jiang, Xiangyang Xue are with the School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University. Email: {xlqian15,ygj, xyxue}@fudan.edu.cn.*
- *Tao Xiang is with the School of Electric and Electronic Engineering, University of Surrey, UK. Email: t.xiang@surrey.ac.uk*



(a) Matching two similar-looking people requires examining both global and local image patches.

(b) Our MuDeep model computes features at multiple scales for re-id.

Figure 1. Computing multi-scale features is crucial for re-id and motivates our approach. In (a), to distinguish between two people wearing similar clothing, global visual cues such as body shape and clothing color are insufficient. Visual cues from local parts such as shoes and hairstyle are needed for telling them apart. Motivated by this observation, our MuDeep, as shown in (b), learns discriminative features at different spatial scales and locations (indicated by the red dashed boxes).

whilst for some others, detecting local image patches corresponding to, say a handbag of a particular color or the type of shoes, would be critical for distinguishing between two otherwise very similar-looking people. In other words, the optimal matching results are only obtainable when features at different scales are computed and combined. Such a multi-scale matching process is also likely adopted by most humans when it comes to re-id. In particular, humans typically compare two images from coarse to fine [31], [32], [33], [34], [35], [36]. More specifically, Parker *et al.* [34] presented results on a coarse-to-fine integration in complex natural scene images. These results have shown that low-to-high sequences of natural scene images are better integrated by the human perceptual system than high-to-low spatial frequency sequences. Musel *et al.* [32] believed that the human cognitive system operates in a coarse-to-fine manner in visual recognition, in terms of the various supporting results from neurophysiological, computational, and behavioral studies. Take the images of two persons in Figure 1(a) as an example. At the coarse level, the color and texture information of their clothes are very similar; humans would thus go down to finer scales to discern subtle local differences (*e.g.*, hairstyle and type of shoes) to reach a conclusion that these are two different people.

To this end, we propose a novel multi-scale deep learning model (MuDeep) for re-id. Our model aims to learn discriminative feature representations at multiple scales and automatically determine the different important spatial locations for each scale (see Figure 1(b) for a conceptual illustration and Figure 2 for a detailed architecture). The network architecture of MuDeep is designed to enable to learn features at different scales and exploit their importance for cross-camera matching. These abilities are achieved by introducing two novel types of layers: *a multi-scale stream layer* that extracts image features by analyzing the person images at multi-scale; and *a leader-based attention learning layer,* which selectively learns to refine the multi-scale data streams and generate more discriminative features. In addition to the classification loss used to globally supervise the network learning by many previous deep re-id models, we introduce a pair of classification losses to strengthen the multi-scale feature learning at two different paths, global and local. This also helps to extract features at different locations that can potentially capture subtle differences between two similar-looking people.

Apart from being more discriminative, another benefit of extracting rich scale and location sensitive features is that the features are more generalizable to unseen datasets collected from different camera networks installed in different environments. This is because by examining a richer set of features at different scales and locations, it is more likely that some of the features will be transferable to the new dataset, regardless of how different the new dataset domain is from the domain where the model is trained. In order to evaluate the generalization ability of the proposed re-id model, extensive experiments are conducted under a domain generalization (DG) setting, that is, the re-id model is only trained on a source domain and then tested on a variety of unseen domains without model updating using labeled/unlabeled data from the target domain. Under this challenging yet practical setting[1], our extensive experiments show that our model significantly outperforms state-of-the-art alternatives.

**Our contributions** are as follows. (1) We propose a novel multi-scale representation learning architecture for learning discriminative person appearance features at multiple spatial scales and locations. Critically, the multiple scales refer to different resolution levels of filters, rather than multi-scale inputs. This approach results in a lighter and more efficient model compared with previous work. (2) We propose a leader-based attention learning layer that utilizes the information computed at all scales to lead, and dynamically determines the important spatial locations in the feature extraction at each scale. This novel strategy is fundamentally different from the self-attention strategy [37], [38]. (3) Extensive experiments and performance analyses are conducted on a number of benchmark datasets to show that our model outperforms state-of-the-art deep re-id models,

[1] Collecting and annotating labeled data across camera networks is difficult even for humans, thus it is very expensive. Consequently, domain generalization is critical for the applicability of a re-id model, *i.e.*, directly applying re-id models to a new camera network without any model adaptation.

often by a significant margin. (4) We propose a more realistic domain generalization setting for person re-identification to verify the generalization ability of our model. Essentially, we argue that this is a more realistic setting in extending the study of person re-identification to real-world applications. An early and preliminary version of this work has been published in [39]. Compared with [39], some significant modifications have been made in the network architecture and evaluation experiments.

## 2 Related Work

### 2.1 Deep re-id models

Various deep learning architectures have been proposed to either address visual variations in pose and viewpoint [25], [40], [41], [42], or learn better relative distances of triplet training samples [43], [44], [45], [46], or learn better similarity metrics of any pairs [47], [29], [48], [23]. To cope with the data sparsity problem, Xiao *et al.* [49] proposed a single deep network built upon the inception module [50], combined multiple re-id datasets together for training, and introduced a domain guided dropout strategy to achieve domain adaptation for each individual dataset. More recently, variants of Siamese Network have been studied for person re-id [51], [52]. Pairwise and triplet comparison objectives were utilized to combine several sub-networks for person re-id in [27]. Zhong *et al.* [29] proposed a novel method to utilize hard sample mining online with triplet loss in person re-identification. Similarly, Chen *et al.* [43] improved triplet loss and proposed a deep quadruplet network. With the success of generative adversarial networks (GAN) in image generation, Wei *et al.* [53] and Zhong *et al.* [26] applied GAN in the re-id task to solve the problem of domain gap and overcame the problem of lacking labeled data in new domains. Among these existing approaches, a number are closely related which are worth mentioning and differentiating from our model.

1) Our MuDeep generalizes convolutional layers with multi-scale representation learning. In particular, we propose a multi-scale stream layer and a leader-based attention learning layer for multi-scale learning, which is clearly different from the ideas of combining multiple sub-networks [27] or channels [30] with the pairwise or triplet loss.
2) He *et al.* [54] proposed a multi-branch deep network to obtain global feature representations and local feature representations with multiple granularities. In contrast, our work applies a multi-branch architecture to not only learn the feature representations of person images at multiple scales, but also share the weights of previous layers to exploit the complementarity between multi-scale feature representations.
3) Both Shen *et al.* [55] and Guo *et al.* [51] improved the accuracy of person re-id by using a multi-level similarity, which was computed by multi-level features. Although the multi-level features are related to analyzing person images at multiple scales, our MuDeep is specially designed for multi-scale feature learning with multiple branches. Importantly, features are extracted using convolutional filters of different receptive fields at the same abstraction level, *i.e.*, in the same convolution block/layer rather than across different layers. Moreover, in order to learn scale-specific feature representations, the weights of the multiple branches are not shared between any two of them in our network.

### 2.2 Multi-scale re-id

The idea of multi-scale learning for re-id was first exploited in [56]. However, the definition of scale is different: It was defined as different levels of input resolutions rather than as in our definition, which applies multi-scale filters. Despite the similarity between terminology, very different problems are tackled in these two approaches. Compared with previous multi-scale methods in re-id, Chen *et al.* [57] adopted $m$ scale-specific networks to learn deep pyramidal features from images with different scales; however, our MuDeep has a much simpler architecture and only utilizes one network with multiple branches to extract $m$ scale-specific representations of one person image. Wang *et al.* [58] extracted multi-resolution embeddings in one network at different stages, and fused them with a simple weighted sum to solve the problem of person re-identification under resource constraints. Our MuDeep concentrates on exploiting and combining discriminative global and local information for re-id tasks and utilizes multiple branches to extract multi-scale feature representations.

Another similar multi-scale deep re-id method that we are aware of is Liu *et al.* [45]. Compared with our model, Liu *et al.* [45] proposed a quite "straightforward" multi-scale model where different down-sampled versions of the input image were fed into shallower sub-networks to extract features at different resolutions and scales. These sub-networks were combined with a deeper main network for feature fusion. With an explicit network for each scale, this network becomes computationally very expensive. In addition, scale weighting cannot be learned automatically and no spatial importance of features can be modeled as in ours.

### 2.3 Deep attention modeling

In deep representation learning, the attention mechanism [59] works in a top-down fashion and allows the salient features to dynamically come to the front as needed. This has been widely used in various computer vision tasks, including person re-identification [60], [61], [62]. Recently, a self-attention mechanism has received increasing interest as a means to dynamically focus on local salient regions for computing deep features [63], [37], [64]. In order to capture multiple attention

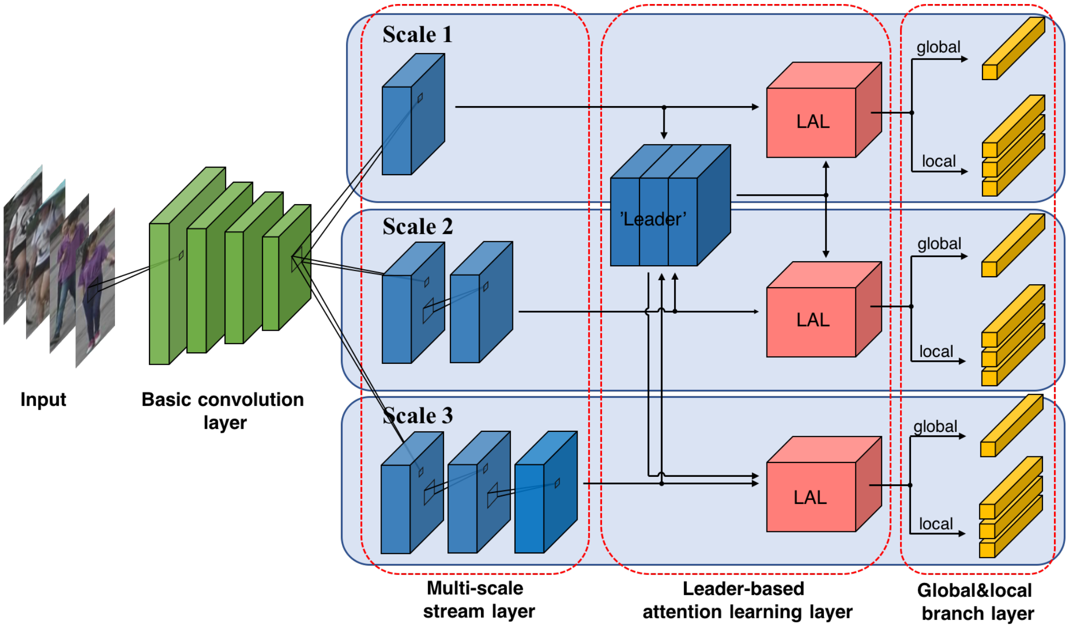


Figure 2. Overview of MuDeep architecture. The multi-scale stream layer first analyzes feature maps with multiple scales. Then the leader-based attention learning layer is followed to automatically discover and emphasize important spatial locations. Finally, the global and local branch layer is utilized to extract discriminate features from global and local parts. Note that the parameters of each scale are not shared. 'LAL' means the **L**eader-based **A**ttention **L**earning layer, with further details shown in Figure 4.

features from the low-level to the semantic-level, Liu *et al.* [24] proposed a HydraPlus-Net, which consists of three multi-directional attention modules, to extract discriminative features. To overcome the visual ambiguity, Si *et al.* [65] presented a feature sequence extraction module and a feature sequence matching module. Especially, a dual attention mechanism was used in the latter to realize intra-sequence refinement and inter-sequence alignment. Similarly, Li *et al.* [66] jointly learned the soft pixel attention and hard regional attention along with simultaneous optimization of feature representations to optimize person re-id in misaligned images.

In this work, we propose a leader-based attention learning layer to further refine the feature representations at each scale and abandon redundant and useless information. Specifically, the attention features are computed with the guidance which is summarized from the outputs of the multi-scale stram layer, to automatically discover the most discriminative feature regions among different scales. This is thus quite different from the conventional self-attention module adopted in existing deep re-id models. Importantly, our proposed layer computes attention maps not only with the input itself, but also with the features from other scale streams.

### 2.4 Domain generalization in re-id

Conventionally, the person re-id problem is formulated as a supervised learning task. Given a dataset collected from a specific camera network, existing re-id models rely on hundreds of labeled data samples per camera pair (*i.e.*, images are paired if they contain the same person) for model training. Specifically, most models are based on either supervised distance metric learning [16], [67], [68], [69], [70], or learning to rank [71], or deep learning [23]. However, the supervised setting has significantly limited the applicability of existing person re-id models in real-world scenarios. This is because with a wide deployment of CCTV surveillance cameras, it is not uncommon to have a camera network consisting of hundreds or even thousands of cameras, and thus, labeling a sufficient number of training data for each individual camera pair is not possible.

In order to overcome this problem, transfer learning can be exploited in re-id. A typical deep learning based approach would pretrain the re-id model on a source dataset with sufficient labeled training samples, followed by model fine-tuning on a target dataset with a small set of labeled data samples. To further remove the need for labeling any data from the target dataset/domain, a number of unsupervised transfer learning based person re-id approaches have been proposed. These methods assume that only unlabeled data from a target dataset/domain can be used for model adaptation. Concretely, the SVM multi-kernel learning transfer strategy [72] was adopted. Alternatively, multi-task metrics learning models [19], [73], [74] were also proposed. However, these unsupervised transfer learning based re-id models still require the collection of unlabeled examples in the target dataset to update models.

In this work, we consider the most challenging domain generalization setting for person re-id. Under this setting, no data from the target dataset is used for model updating: the model trained from labeled source datasets/domains is applied to the target domain without any modification. Thus, this is completely different from existing transfer learning based re-id settings. We show that our proposed MuDeep is naturally suitable for this setting as the features extracted at multiple scales are more likely to generalize to unseen domains.

# 3 Multi-scale Deep Architecture (MuDeep)

**Problem Definition.** Given a training dataset of $N$ persons $\mathcal{D}_{Train} = \{\mathbf{I}_k, y_k\}_{k=1}^{N}$, where $\mathbf{I}_k$ and $y_k$ are the $k$-th person's image and identity. During the training phase, we learn a function $\mathcal{F}(\cdot)$ to extract the feature of image $\mathbf{I}$, $\mathbf{f}_\mathbf{I} = \mathcal{F}(\mathbf{I})$. In the testing stage, given a pair of person images $\{\mathbf{I}_i, \mathbf{I}_j\}$ from testing dataset $\mathcal{D}_{Test}$, where the identities are not overlapping between both datasets, we need to judge whether $y_i = y_j$ or $y_i \neq y_j$.

**Architecture Overview.** As shown in Figure 2, Our MuDeep consists of four components: *basic convolution layer*, *multi-scale stream layer*, *leader-based attention learning layer*, and *global and local branch layer*. The remainder of this section is structured as follows. We first give a brief introduction of the *basic convolution layer* in Sec. 3.1. Then, in Sec. 3.2, Sec. 3.3 and Sec. 3.4, we provide details on the key components of the *multi-scale stream layer*, *leader-based attention learning layer*, and *global and local branch layer*, respectively. Loss functions are formulated in Sec. 3.5. Finally, a discussion on the domain generalization setting is provided in Sec. 3.6.

## 3.1 Basic convolution layer

A basic convolution layer is applied here to extract the middle-level features of input person images, which are further processed by the multi-scale stream layer (see Sec. 3.2). Considering that the ResNet architecture [75] has achieved outstanding performance in ImageNet and has been widely adopted in re-id work [76], [77], [26], we apply ResNet-50 [75] as our basic convolution layer. More specifically, the size of the input image is resized to $384 \times 192$ using bi-linear sampling due to the requirements of the global and local branch layer (see Sec. 3.4). In order to extract middle-level features and obtain the considerable size of feature maps for further analysis, we remove the last block of ResNet-50, that is, our basic convolution layer only consists of the *con1*, *res2*, *res3* and *res4* blocks.

## 3.2 Multi-scale stream layer

As mentioned previously, humans typically compare two images from coarse to fine. In particular, if the cues at the coarse scales, such as body shape, are insufficient to verify the images of two persons, humans will further examine the fine-grained details (*e.g.*, handbags or shoes), which are more discriminative visual cues. Correspondingly, we propose a multi-scale stream layer to analyze the data stream at multi-scale. In order to capture the characteristic representations of different scales, all these multi-scale data streams do not share weights.

| Stream | Layer | Output |
|---|---|---|
| scale 1 | $3 \times 3$+$R^*$ | $2048 \times 24 \times 12$ |
| scale 2 | $3 \times 3$+$R^*$ – $3 \times 3$+$R$ | $2048 \times 24 \times 12$ |
| scale 3 | $3 \times 3$+$R^*$ – $3 \times 3$+$R$ – $3 \times 3$+$R$ | $2048 \times 24 \times 12$ |

Table 1

Details of the multi-scale stream layer. Note: (1) $3 \times 3$ indicates the convolution layer with a $3 \times 3$ kernel. (2) '+$R$' means that the type of this layer is the residual block, as shown in Figure 3. (3) '+$R^*$' indicates that there is a $1 \times 1$ convolution layer in the shortcut of the residual block to balance the number of channels.

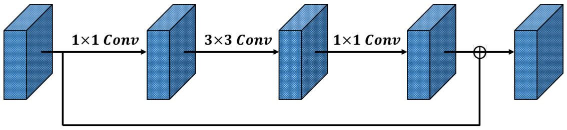


Figure 3. Illustration of our residual block.

The multi-scale stream layer analyzes the input data stream from $S = 3$ scales with receptive field sizes of $3 \times 3$, $5 \times 5$ and $7 \times 7$, respectively. To increase the depth of this layer and reduce the computation cost, we equally split the filter size of $5 \times 5$ into two $3 \times 3$ streams cascaded, and $7 \times 7$ into three $3 \times 3$ streams cascaded. Furthermore, inspired by [75], we transform the data streams with different scales into residual blocks, which can strengthen the feature representation ability at each scale. As shown in Figure 3, before analyzing the information with a $3 \times 3$ kernel, we utilize a $1 \times 1$ convolution layer to first compress and refine the essential features. Finally, another $1 \times 1$ convolution layer is applied to restore the feature maps to the original number of channels, and add it to the output of the shortcut. Details about the multi-scale stream layer are shown in Table 1.

## 3.3 Leader-based attention learning layer

With the output processed by the previous layers of each data stream, the resulting data channels at different scales may have redundant information. For example, some channels may capture relatively important information about a person, like body, while other channels may only model the background context. Therefore, a natural question to ask is, *"Where should one give more in-depth and closer attention?"*

Intuitively, this question should be answered by a "Leader", which contains more global and comprehensive information about all the feature channels and scales, and thus is best positioned to determine where to focus attention. In the light of this strategy, a leader-based attention learning mechanism is utilized here to guide the outputs of the multi-scale stream layer, and to also automatically discover and emphasize the channels with more discriminative features. Note that different from [37], [38], which utilized the self-attention mechanism, our proposed leader-based attention learning layer computes the attention maps with both the input itself, and the features from other scale streams, as illustrated in Figure 4.

Formally, we assume $\mathbf{F}_i$ represents the feature maps from the $i-$th data stream $(1 \leq i \leq 3)$; the operation of the leader-based attention learning layer in each data stream is denoted as $\overline{\mathbf{F}_i} = LAL(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$, which can be expressed as follows,

$$\mathbf{H}_g = \mathbf{W}_g{\cdot}\mathrm{Cat}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3) \tag{1}$$

$$\mathbf{H}_{f_1} = \mathbf{W}_{f_1}\mathbf{H}_g,\ \mathbf{H}_{f_2} = \mathbf{W}_{f_2}\mathbf{H}_g \tag{2}$$

$$\alpha_i = \frac{\exp\left(\mathbf{H}^i_{f_1} \otimes (\mathbf{H}_{f_2})^T\right)}{\sum_{j=1}^{C} \exp\left(\mathbf{H}^i_{f_1} \otimes \left(\mathbf{H}^j_{f_2}\right)^T\right)} \tag{3}$$

$$\overline{\mathbf{F}_i} = \alpha \otimes \mathbf{F}_i \tag{4}$$

where $\mathbf{F}_i \in \mathbb{R}^{N_b \times C \times H \times W}$; $N_b$, $C$, $H$ and $W$ mean the number of batch size, feature channels, height and width, respectively; Cat $(\cdot)$ represents the operation of concatenation; $\mathbf{W}_g \in \mathbb{R}^{C_g \times 3C \times 1 \times 1}$, $\mathbf{W}_{f_1} \in \mathbb{R}^{C \times C_g \times 1 \times 1}$ and $\mathbf{W}_{f_2} \in \mathbb{R}^{C \times C_g \times 1 \times 1}$ are the parameters of convolution layers with $1 \times 1$ kernel size. The symbol $\otimes$ in Equation 3 and Equation 4 is the batch matrix multiplication. In particular, the dimensions of $\mathbf{H}_{f_1}$ and $\mathbf{H}_{f_2}$ in Equation 2 are $N_b \times C \times H \times W$, in order to do batch matrix multiplication in Equation 3, $\mathbf{H}_{f_1}$ and $\mathbf{H}_{f_2}$ will be flattened with $N_b \times C \times HW$.

In the leader-based attention learning layer, the feature maps of $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$ are first concatenated to form the "Leader", since it needs to see images from multiple scales. Then, it goes through a convolution layer to generate guidance features $\mathbf{H}_g$ with the guidance channel $\mathbf{C}_g$, which refines useful information from the "Leader". Subsequently, we utilize the self-attention mechanism to calculate attention maps $\alpha$ from two feature spaces $\mathbf{H}_{f_1}$ and $\mathbf{H}_{f_2}$, which are computed by the convolution layers from $\mathbf{H}_g$. Especially, the $\alpha$ contains $C$ attention information to refine the input features. Finally, $\mathbf{F}_i$ is reweighted based on the attention maps $\alpha$ by a softmax function. Note that except for the $\mathbf{W}_g$, the parameters of the convolution layers are not shared within each data stream.

Because we apply random initialization for the parameters of all attention learning layers, it is possible that a bad initialization would have a detrimental effect on the previous multi-scale feature extraction, which can never be recovered. To minimize this risk, we further multiply the output of Equation 4 by a trainable parameter $\beta$ which is initialized as $0$, as follows:

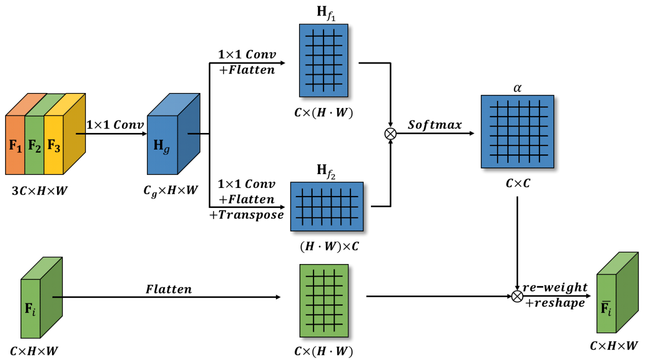


Figure 4. Structure of the leader-based attention learning layer. Note that $\otimes$ means batch matrix multiplication.

$$\overline{\mathbf{F}_i} = \beta\overline{\mathbf{F}_i} + \mathbf{F}_i \tag{5}$$

Such a way ensures that a bad initialization would have a minimal impact on model training. Importantly, the model is given a chance to find more optimal parameters for the attention model and subsequently increase the value of $\beta$.

### 3.4 Global and local branch layer

Generally, a human body can be divided into several parts according to a hierarchy, *e.g.*, two parts (upper body and lower body), three parts (head, upper body, and lower body), four parts (head, chest, abdomen, and legs) and so forth. Accordingly, some previous studies [78], [79] built an architecture with several branches to learn specific features for each part. Recently, Zhang *et al.* [77] and Wang *et al.* [54] both proposed approaches to extract global and local features for re-identification. Similar to their approaches, we design a global and local branch layer. This leads to several benefits, including better supervision of learning the discriminative features from both global and local branches, and exploitation of the relationships between global and local human body parts.

Figure 2 illustrates the structure of our global and local branch layer. There are two branches: a global branch and a local branch. In the global branch, global features $(C{\times}1{\times}1)$ are directly extracted by utilizing a global average pooling. The local branch uses the horizontal global average pooling, which is the operation of applying the global average pooling horizontally on feature maps[2], to extract $M$ local features $(M{\times}C{\times}1{\times}1)$. In the end, a $1{\times}1$ convolution layer is applied after all features to reduce the dimension from $C$ to $C^{'}$. Consequently, following the global and local branch layer, one global feature and $M$ local features will be obtained for each scale.

[2] In order to make sure that feature maps could be divided by different values of $M$, the size of the input image is set to $384 \times 192$.

### 3.5 Loss function

Person re-identification can be formulated as either a classification task or a verification task. For the classification task, the network usually learns to output person feature representations, that is, to classify the identity of a person. This task pays more attention to discriminative feature learning. Meanwhile, for the verification task, it aims to verify whether two persons are the same or not by decreasing the distance of positive pairs and increasing the distance of negative samples.

To combine these two complementary tasks, we utilize both classification loss and verification loss. Thus, the total loss of MuDeep can be calculated as the sum of these two kinds of losses:

$$\mathbf{L} = \lambda_1 \cdot \mathbf{L}_{cls} + \lambda_2 \cdot \mathbf{L}_{ver} \tag{6}$$

where $\lambda_1$ and $\lambda_2$ are the coefficients of each term individually.

**Classification loss.** In order to learn strong discriminative features for person appearance representations, and to supervise the learning of the leader-based attention learning layer, we first add classification loss to all global features and local features, which learn to classify different pedestrian identities with different scale features. Specifically, the softmax with $N$ output neurons are connected, where $N$ denotes the number of pedestrian identities. The total classification loss can be formulated as follows:

$$\mathbf{L}_{cls} = \sum_{i=1}^{3} \frac{1}{1+M} \left( L_g^i + \sum_{j=1}^{M} L_{l_j}^i \right) \tag{7}$$

where $M$ is the number of local features; $L_{l_j}^i$ represents the softmax loss of the $j$-th local feature in the data stream of scale $i$; and $L_g^i$ denotes the softmax loss of a global feature at scale $i$.

**Verification loss.** In order to minimize the intra-class variations and maximize the inter-class variations, the verification loss, *i.e.*, triplet loss, is employed to help optimize the network. Particularly, we adopt the HardTrip loss [44] in favor of online hard negative mining. In the training stage and given image $a$ in a batch $P \times K$ ($P$ persons with $K$ images each), we select the hardest positive and negative samples in a batch to compute triplet loss,

$$\mathbf{L}_{ver} = \sum_{i=1}^{3} \frac{1}{P \times K} \sum \left( \max d_{pos}^i - \min d_{neg}^i + m \right)_+ \tag{8}$$

where $m$ is the parameter of margin; $d^i$ denotes the Euclidean distance of global features at scale $i$ and the subscript $pos$ or $neg$ means the positive pairs or negative pairs; and the value of $(x)_+$ is not equal to zero if $x > 0$.

| Dataset | # ID | # Train | # Test | Evaluation |
|---|---|---|---|---|
| CUHK03-NP | 1467 | 767 | 700 | SQ |
| CUHK03 | 1467 | 1367 | 100 | SS |
| CUHK01 (100) | 971 | 871 | 100 | SS |
| CUHK01 (486) | 971 | 485 | 486 | SS |
| Market-1501 | 1501 | 751 | 750 | SQ |
| DukeMTMC-reID | 1402 | 702 | 702 | SQ |

Table 2
Settings of all datasets. Note: (1) SS: single-shot; (2) SQ: single-query.

### 3.6 Domain generalization

In this paper, we consider the person re-id under a domain generalization setting, where no person image has been observed so far from the target dataset. We demonstrate that a key strength of our MuDeep is that it can extract features that are generalizable to unseen domains without any model updating. This is due to two reasons. First, the rich multi-scale features cover visual cues of different scales and locations. This richness means that at least some of the extracted features would be useful in any given domain. Second, for any given instance in the training domain, our leader-based attention learning layer is able to determine the optimal ways to exploit critical features of different scales. This dynamic attention mechanism is likely to be easier to generalize than the features themselves. Therefore, an additional dimension is provided for the feature extraction process to adapt to new person images from unseen domains on the fly.

## 4 Experiments

### 4.1 Datasets and settings

**Datasets.** The proposed method is evaluated on four widely used datasets, *i.e.*, Market-1501 [92], DukeMTMC-reID [93], CUHK03 [25], and CUHK01 [94]. The settings for all the datasets are summarized in Table 2.

1) **Market-1501** is collected from six different camera views. It has 32,668 bounding boxes of 1,501 identities obtained using a Deformable Part Model (DPM) person detector. Following the standard splits in [92], we use 751 identities with 12,936 images for training and the remaining 750 identities with 19,732 images for testing.
2) **DukeMTMC-reID** is constructed from the multi-camera tracking dataset - DukeMTMC [95]. It contains 1,812 identities. Following the evaluation protocol in [93], 702 identities are used as the training set and the remaining 1,110 identities as the testing set. During testing, one query image for each identity in each camera is used for query and the remaining as the gallery set.
3) **CUHK03** 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

| Methods | DukeMTMC-reID | CUHK03(D) | CUHK03(L) | CUHK03-NP(D) | CUHK03-NP(L) | CUHK01 (100) | CUHK01 (486) |
|---|---|---|---|---|---|---|---|
| ResNet-50$^*$ | 27.87/13.94 | 16.50/– | –/– | –/– | –/– | –/– | 27.20/– |
| PN-reID [40] | 29.94/15.77 | 16.85/– | –/– | –/– | –/– | –/– | 27.58/– |
| ResNet-50 | 29.30/14.81 | 25.08/– | 26.17/– | 7.85/6.37 | 8.00/6.41 | 49.11/– | 32.14/– |
| MGN$^\dagger$ [54] | 43.27/24.99 | 28.84/– | 29.48/– | 8.50/7.37 | 9.21/8.72 | 50.88/– | 35.94/– |
| MuDeep | **47.57/27.66** | **35.16/–** | **37.90/–** | **10.29/9.10** | **11.01/9.23** | **63.13/–** | **45.38/–** |

Table 3
Rank-1/mAP accuracy of transfer learning on extensive benchmarks. Note that all models are only trained with images from Market-1501 dataset. '*' means that the results are reported from [40]. '†' indicates our implementation.

| Method | 100 test IDs | | | 486 test IDs | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-1 | Rank-5 | Rank-10 |
| KISSME [16] | 29.40 | 60.18 | 74.44 | – | – | – |
| eSDC [80] | 22.84 | 43.89 | 57.67 | 19.76 | 32.72 | 40.29 |
| mFilter [81] | – | – | – | 34.30 | 55.00 | 65.30 |
| DeepReid [25] | 27.87 | 58.20 | 73.46 | – | – | – |
| IDLA [23] | 65.00 | 88.70 | 93.12 | 47.53 | 71.50 | 80.00 |
| DeepRanking [82] | – | – | – | 50.40 | 70.00 | 84.80 |
| GOG [83] | – | – | – | 57.80 | 79.10 | 86.20 |
| EMD [28] | 69.38 | 91.03 | 96.84 | – | – | – |
| SI-CI [27] | 71.80 | 90.35 | 93.50 | – | – | – |
| MTDNet [84] | 78.50 | 96.50 | 97.50 | – | – | – |
| Spindle [85] | 79.90 | 94.40 | 97.10 | – | – | – |
| Quadruplet [43] | 81.00 | 96.50 | 98.00 | 62.55 | 83.44 | 89.71 |
| X-Corr [86] | 81.23 | 95.00 | 97.39 | 65.04 | – | 89.76 |
| NullReid [87] | – | – | – | 69.09 | 86.87 | 91.77 |
| ReID-GLILA [88] | 84.80 | 95.10 | 98.40 | – | – | – |
| CSN [51] | 88.20 | 98.20 | 99.35 | – | – | – |
| DeepAlign [89] | 88.50 | 98.40 | 99.60 | 75.00 | 93.50 | 95.70 |
| DCSL [90] | 89.60 | 97.80 | 98.90 | 76.50 | 94.20 | 97.50 |
| MC-PPMN [91] | 93.45 | 99.62 | 99.98 | 78.95 | 94.67 | 97.64 |
| MuDeep (SL) | **98.73** | **99.82** | **100** | **87.55** | **96.63** | **98.38** |

Table 4
Results on CUHK01 dataset. '-' indicates not reported.

are provided [25]: manually labeled pedestrian bounding boxes (Labeled) and bounding boxes automatically detected by the deformable-part-model detector [96] (Detected). The manually labeled images are generally of higher quality than those detected images. We use the settings of both manually *labeled* and automatically *detected* person images on the standard splits in [25]. Furthermore, we also report our results using a hard setting as proposed in [29], which is called CUHK03 new protocol (CUHK03-NP for short): 767 identities are used for training and 700 identities for testing.

4) **CUHK01** has 971 identities with 2 images per person in each camera view. As in [94], we use the images from camera A as the probe and take those from camera B as the gallery. The experiments are repeated over 10 trials. We evaluate our approach on this dataset using two settings:
   a) CUHK01 (100): we randomly select 100 identities as the test set, and the remaining identities are for training and validation;
   b) CUHK01 (486): 486 identities are randomly chosen for testing, and the remaining of 485 identities are used for training.

**Implementation details.** We implement our model based on the PyTorch framework. We use the weights of ResNet-50 pretrained on ImageNet [102] for finetuning. We set the channels $C_g$ of the guidance features as $512$, and set $C^{'}$ as 512. During training, the images are resized to $384 \times 192$, as the same aspect ratio for images in Market-1501. We also apply random horizontal flipping and random erasing [103] for data augmentation. Each batch is sampled with size $P \times K = 12 \times 4$. For triplet loss, we set the margin $m$ to 1, and set $\lambda_1 : \lambda_2 = 2 : 1$. Additionally, SGD is utilized as the optimizer to train networks with momentum $0.9$, and the weight decay factor for L2 regularization is set to $0.0005$. We train our network for $100$ epochs with the initial learning rate set to $0.01$, and reduce the learning rate by a factor of 0.1 every 40 epochs. The proposed MuDeep model gets converged in two hours with Market-1501 training images on two NVIDIA TITAN Xp GPUs. During testing, all features from the global and local branch layer are concatenated as the final re-id features.

**Evaluation metrics.** In terms of standard evaluation metrics, we report the Rank-1, Rank-5 and Rank-10 accuracies with the single-shot setting, and report Rank-1 and mAP accuracies with the single-query setting.

**Evaluation settings.** The evaluation experiments on all datasets are conducted under two settings.

1) Supervised Learning (SL) setting: the models are

| Method | Detected | | | Labeled | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-1 | Rank-5 | Rank-10 |
| SDALF [8] | 4.87 | 21.17 | 35.06 | 5.60 | 23.45 | 36.09 |
| eSDC [80] | 7.68 | 21.86 | 34.96 | 8.76 | 24.07 | 38.28 |
| LMNN [68] | 6.25 | 18.68 | 29.07 | 7.29 | 21.00 | 38.28 |
| XQDA [12] | 46.25 | 78.90 | 88.55 | 52.20 | 82.23 | 92.14 |
| LDM [97] | 10.92 | 32.25 | 48.78 | 13.51 | 40.73 | 52.13 |
| DeepReid [25] | 19.89 | 50.00 | 64.00 | 20.65 | 51.50 | 66.50 |
| IDLA [23] | 44.96 | 76.01 | 83.47 | 54.74 | 86.50 | 93.88 |
| SI-CI [27] | 52.17 | 84.30 | 92.30 | – | – | – |
| EMD [28] | 52.09 | 82.87 | 91.78 | 61.32 | 88.90 | 96.44 |
| G-Dropout [49] | – | – | – | 72.58 | 91.59 | 95.21 |
| Gated_Sia [52] | 68.10 | 88.10 | 94.60 | – | – | – |
| X-Corr [86] | 72.04 | 92.10 | 96.00 | 72.43 | 92.50 | 95.51 |
| MTDNet [84] | 74.68 | 95.99 | 97.47 | – | – | – |
| Quadruplet [43] | 75.53 | 95.15 | **99.16** | – | – | – |
| JLML [98] | 80.60 | 96.90 | 98.70 | 83.20 | 98.00 | 99.40 |
| DeepAlign [89] | 81.60 | 97.30 | 98.40 | 85.40 | 97.60 | 99.40 |
| DPFL [57] | 82.00 | – | – | 86.70 | – | – |
| Verif-Identif.+LSRO [93] | 84.60 | 97.60 | 98.90 | – | – | – |
| PDC [41] | 78.29 | 94.83 | 97.15 | 88.70 | 98.61 | 99.24 |
| CSN [51] | 86.45 | 97.50 | 99.10 | 87.50 | 97.85 | 99.45 |
| HP-net [24] | – | – | – | 91.80 | 98.40 | 99.10 |
| MC-PPMN [91] | 81.88 | 96.56 | 98.58 | 86.36 | 98.54 | 99.66 |
| HAP2S_P [99] | 88.90 | 98.40 | 99.10 | 90.40 | **99.50** | **99.90** |
| Mancs [100] | 92.40 | **98.80** | **99.40** | 93.80 | 99.30 | 99.80 |
| SGGNN [101] | – | – | – | 95.30 | 99.10 | 99.60 |
| ReID-GLILA [88] | 90.90 | 98.20 | – | 92.50 | 98.80 | – |
| Aligned-ReID [77] | – | – | – | 91.90 | 98.70 | 99.40 |
| MuDeep (SL) | **93.70** | 98.53 | 98.95 | **95.84** | 99.42 | 99.61 |

Table 5
Results on CUHK03 dataset. '-' indicates not reported.

| Method | Detected | | Labeled | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| BOW+XQDA [92] | 6.40 | 6.40 | 7.90 | 7.31 |
| LOMO+XQDA [12] | 12.80 | 11.50 | 14.80 | 13.60 |
| IDE [104] | 21.30 | 19.70 | 22.20 | 21.00 |
| IDE+XQDA [104] | 31.10 | 28.20 | 32.00 | 29.60 |
| PAN [105] | 36.30 | 34.00 | 36.90 | 35.00 |
| SVDNet [106] | 41.50 | 37.30 | – | – |
| HA-CNN [66] | 41.70 | 38.60 | 44.40 | 41.00 |
| MLFN [107] | 52.80 | 47.80 | 54.70 | 49.20 |
| PCB [108] | 61.30 | 54.20 | – | – |
| PCB+RPP [108] | 63.70 | 57.50 | – | – |
| DaRe(De)+RE [58] | 63.30 | 59.00 | 66.10 | 61.60 |
| MGN [54] | 66.80 | 66.00 | 68.00 | 67.40 |
| Mancs [100] | 65.50 | 60.50 | 69.00 | 63.90 |
| MuDeep (SL) | **71.93** | **67.21** | **75.64** | **70.54** |

Table 6
Results on CUHK03-NP dataset. '-' indicates not reported.

trained on the training set of one dataset, and evaluated on the corresponding testing set.

2) Domain Generalization (DG) setting: the models are trained on Market-1501 dataset, and directly evaluated on the testing sets of the other re-id datasets, *i.e.*, CUHK03, CUHK01, and DukeMTMC-reID. The DG setting is especially useful in real-world scenarios, where a pre-trained model needs to be deployed to a new camera network without any model fine-tuning. This setting thus tests the generalizability of a re-id model.

## 4.2 Domain generalization results

We first report our results obtained under the DG settings on three datasets, *i.e.*, DukeMTMC-reID, CUHK03, and CUHK01. All models are trained on Market-1501 dataset, and directly used for evaluation on other testing sets. For the evaluation of generalization ability, we compare our results against those of ResNet-50 baselines as well as two very recent and state-of-the-art re-id models PN-reID [40] and MGN [54] in Table 3.

First, except for CUHK03-NP, we can observe that our model gets improvements over those of ResNet-50 baselines and PN-reID/MGN by at least $10\%$. These results thus show that our model has the potential to be truly generalizable to new re-id data from new camera networks when operating in a "plug-and-play" mode.

Second, it can be seen from Table 3 that CUHK03-NP is a truly challenging setting, however, our model still improves the performance by $3\%$, which demonstrates that our proposed approach facilitates the person re-id in the domain generalization setting.

Third, we have to say that the results under the domain generalization setting are much lower than those in the supervised setting. On the one hand, this is because of the intrinsic difficulty of the domain generalization setting. On the other hand, it shows that domain generalization in person re-id is still an important and not fully-resolved problem.

| Methods | Rank-1 | mAP |
|---|---|---|
| TMA [109] | 47.90 | 22.3 |
| SCSP [110] | 51.90 | 26.40 |
| DNS [87] | 61.02 | 35.68 |
| Gated_Sia [52] | 65.88 | 39.55 |
| HP-net [24] | 76.90 | – |
| Spindle [85] | 76.90 | – |
| PIE [42] | 79.33 | 55.95 |
| Verif.-Identif. [111] | 79.51 | 59.87 |
| DLPAR [89] | 81.00 | 63.40 |
| DeepTransfer [112] | 83.70 | 65.50 |
| Verif-Identif.+LSRO [93] | 83.97 | 66.07 |
| PDC [41] | 84.14 | 63.41 |
| DML [113] | 87.70 | 68.80 |
| SSM [114] | 82.20 | 68.80 |
| JLML [98] | 85.10 | 65.50 |
| PN-reID [40] | 89.43 | 72.58 |
| CSA [26] | 89.49 | 71.55 |
| MLFN [107] | 90.00 | 74.30 |
| HA-CNN [66] | 91.20 | 75.70 |
| DuATM [65] | 91.42 | 76.62 |
| Deep-Person [115] | 92.31 | 79.62 |
| Aligned-ReID [77] | 92.62 | 82.31 |
| SGGNN [101] | 92.30 | 82.80 |
| HAP2S_P [99] | 84.59 | 69.43 |
| Mancs [100] | 93.10 | 82.30 |
| ReID-GLILA [88] | 93.30 | 81.80 |
| PCB+RPP [108] | 93.81 | 81.62 |
| MGN [54] | **95.70** | **86.90** |
| MuDeep (SL) | 95.34 | 84.66 |

Table 7
Results on Market-1501 dataset. '-' indicates not reported. Note that all results are reported without re-ranking [29] for a fair comparison.

| Methods | Rank-1 | mAP |
|---|---|---|
| LOMO+XQDA [12] | 30.80 | 17.00 |
| ResNet50 [75] | 65.20 | 45.00 |
| Basel. +LSRO [93] | 67.70 | 47.10 |
| AttIDNet [116] | 70.69 | 51.88 |
| PN-reID [40] | 73.58 | 53.20 |
| SVDNet [106] | 76.70 | 56.80 |
| CSA [26] | 78.32 | 57.61 |
| HA-CNN [66] | 80.50 | 63.80 |
| MLFN [107] | 81.00 | 62.80 |
| DuATM [65] | 81.82 | 64.58 |
| Deep-Person [115] | 80.91 | 64.83 |
| HAP2S_P [99] | 75.94 | 60.64 |
| SGGNN [101] | 81.10 | 68.20 |
| PCB+RPP [108] | 83.31 | 69.20 |
| Mancs [100] | 84.90 | 71.80 |
| MGN [54] | **88.70** | **78.40** |
| MuDeep (SL) | 88.19 | 75.63 |

Table 8
Results on DukeMTMC-reID dataset. All results are reported without re-ranking [29].

### 4.3 Comparisons against the state-of-the-art

**Evaluations on CUHK01.** The results of CUHK01 dataset are provided in Table 4. It is a small re-id dataset requiring strong learning and discriminative feature extraction ability of re-id models. Moreover, the setting of 486 test IDs is more challenging because this provides fewer training samples. Our MuDeep still obtains the best performance with 98.73% and 87.55% on Rank-1 accuracy under two different settings, which beats all state-of-the-art methods; and is about 10% higher than the second best method [90]. This further shows the advantages of our framework.

**Evaluations on CUHK03.** On the CUHK03 dataset, our results are compared with the state-of-the-art methods under both manually labeled and automatically detected settings in Table 5. The setting of automatically detected is more realistic, but more difficult than manually labeled. Firstly, our MuDeep outperforms the methods of using hand-crafted features and recent deep learning models, including the method of [54], which shows the advantages of our proposed multi-scale stream layer and leader-based attention learning layer. Secondly, compared with CSN [51], a method of computing the visual similarities at different levels of the whole network and leveraging Spatial Transformer Networks to extract meaningful parts from the feature maps, our results are 7.25% and 8.34% higher at Rank-1 accuracy in the *Detected* setting and *Labeled* setting, respectively. More importantly, we also evaluate the effectiveness and scalability of our model using a new setting as proposed in [29]: 767 identities for training and 700 identities for testing, which is a more challenging re-id task. As shown in Table 6, our results provide the best performance, and are about 5% higher than the second best method [54], which utilized a multi-branch architecture to extract global and local features.

**Evaluations on Market-1501.** We also evaluate our approach on Market-1501, which is one of the largest re-id datasets. As shown in Table 7, our MuDeep can achieve competitive performance compared with the method of MGN. Except for MGN, our model outperforms all the other methods. Particularly, we note that our results are significantly better than the third best model by a margin of 1.5% and 2.7% in Rank-1 and mAP, respectively. This validates the efficacy of our architecture and suggests that the proposed multi-scale stream layer and leader-based attention learning layer can help extract discriminative features for person re-id. Moreover, compared with the HA-CNN method, which designed complex attention modules and utilized multi-branch architectures to extract discriminative features, our approach improves the performance from 91.20% to 95.39% in Rank-1, and from 75.70 to 84.37 in mAP. This suggests that our framework can better analyze the multi-scale patterns from data, exploit the complementarity between different scales, and highlight the important regions with attention, due to the novel multi-scale stream layer and the leader-based attention learning layer.

**Evaluations on DukeMTMC-reID.** We further evaluate our approach on DukeMTMC-reID. Similar to Market-1501, the person images in this dataset are captured on campus, but have more occlusion and complex background, which means that this benchmark is more challenging. Table 8 shows that our MuDeep results

| Datasets | Market-1501 | | | | CUHK01(100 test IDs) | | | CUHK03-detected | | | CUHK03-labeled | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | mAP | R1 | R5 | R10 | R1 | R5 | R10 | R1 | R5 | R10 | R1 | R5 | R10 |
| MDLA [39] | - | - | - | - | 79.01 | 97.00 | 98.96 | 75.64 | 94.36 | 97.46 | 76.87 | 96.12 | 98.41 |
| ResNet+MDLA [39] | 79.62 | 91.78 | 96.94 | 98.04 | 96.02 | 99.45 | **100** | 89.25 | 97.57 | **99.03** | 92.36 | 99.05 | 99.52 |
| MuDeep | **84.66** | **95.34** | **98.16** | **98.73** | **98.73** | **99.82** | **100** | **93.70** | **98.53** | 98.95 | **95.84** | **99.42** | **99.61** |

Table 9
Analysis of improvements. “ResNet+MDLA [39]” means that we replace the backbone with ResNet50 and incorporate the residual blocks into the multi-scale stream layer of [39].

| Methods | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| MuDeep w.o. “Leader” | 84.45 | 95.17 | **98.18** | 98.72 |
| MuDeep | **84.66** | **95.34** | 98.16 | **98.73** |

Table 10
Results of training models with/without the “Leader” on Market-1501 dataset.

| Method | Rank-1 | Rank-5 | mAP |
|---|---|---|---|
| -scale -atten -local | 87.02 | 94.62 | 73.30 |
| -atten -local | 89.55 | 95.52 | 76.54 |
| -scale | 94.59 | 98.07 | 82.98 |
| -local | 93.64 | 97.68 | 82.54 |
| -atten | 95.07 | 98.15 | 84.42 |
| MuDeep | **95.34** | **98.16** | **84.66** |

Table 11
Results of comparing variants of MuDeep on Market-1501 dataset. Note that “-scale” denotes that we only use one scale-specific stream layer instead of the multi-scale; “-atten” represents that our MuDeep is trained without the leader-based attention learning layer; and “-local” indicates that MuDeep doesn’t have local branches in the global and local branch layer.

in a lower accuracy than [54] by a small margin, but still outperforms other competitors by clear margins, with 88.19% and 75.63% accuracy in Rank-1 and mAP, respectively. This verifies the importance of multi-scale information analysis and attention modeling in person re-id tasks.

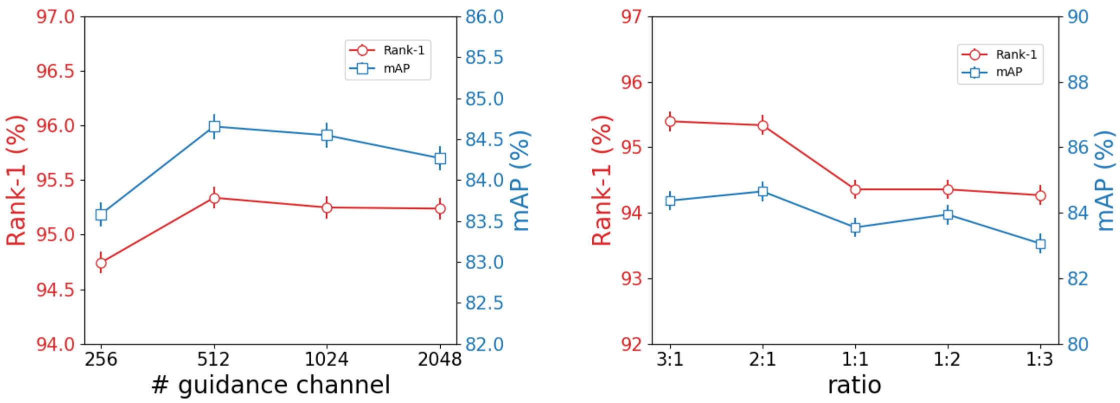
### 4.4 Ablation study

**Analysis of improvements compared with [39].** This ablation study would clearly show how much the proposed new multi-scale learning architecture and leader-based attention learning layer contribute to the final performance. We conduct experiments by replacing the backbone in [39] with ResNet-50 and incorporating residual blocks into the multi-scale stream layer of [39]. For a fair comparison, we also use triplet loss [44] as in MuDeep, instead of the original contrastive loss used in [39]. The results are shown in Table 9, where “MDLA” and “MuDeep” refer to the method in [39] and the model proposed in this paper, respectively; and “ResNet+MDLA” is the contrast experiment described above (*i.e.*, with the same ResNet50 backbone as MuDeep). The results suggest that using the ResNet50 backbone indeed helps, but the gap between our MuDeep and “ResNet+MDLA” is still significant, indicating that the proposed improvements in this paper provide clear contributions to the final model performance.

**Analysis of different components.** To further investigate the contributions of three key components: *multi-scale stream layer*, *leader-based attention learning layer* and *global and local branch layer*, we compare variants of our model by removing different components and tested on Market-1501 dataset. The results in Table 11 show that our full model has the best performance over all variants. Thus, we conclude that all three components are helpful and the combination of the three can further boost the performance.

**Analysis of contribution of the “Leader”.** In order to ensure that the attention maps produced from each scale will not be confused by local information, we propose a leader-based attention learning layer that computes the attention maps with both the input itself and the features from other scale streams in Sec. 3.3. To verify whether the “Leader” works as expected, we conduct experiments by training models with or without Equation 1. The results shown in Table 10 indicate that our proposed leader-based attention learning layer indeed provides clear contributions to the final result.

**Analysis of hyper-parameters.** In Figure 5, we analyze the performance when applying different values of the hyper-parameters (*e.g.*, $C_g$, $\lambda_1$ and $\lambda_2$). First, the guidance channel $C_g$ is a very important hyper-parameter, which controls the information capacity of $\mathbf{H}_g$. As we mentioned in Sec. 3.3, $\mathbf{H}_g$ is computed as the guidance feature to guide the network about where to look closer at each scale. Therefore, a bottleneck layer is adopted here to refine the information of the “Leader”, which thus outputs $\mathbf{H}_g$ with the guidance channel $C_g$. We evaluate our model with different values of $C_g$ on Market-1501 dataset, as illustrated in Figure 5(a). With an increase of $C_g$, the performance first shows an upward trend and then decreases. It indicates that if the value of $C_g$ is too small, the refined features will not be representative enough as a good “Leader”. On the contrary, if $C_g$ is too large, a number of redundant features or trivial background information will still be retained. The best performance is achieved when $C_g$ is set at an appropriate value, *i.e.*, 512. Second, we evaluate various combinations of $\lambda_1$ and $\lambda_2$. These two hyper-parameters balance



(a)

(b)

Figure 5. Analysis of hyper-parameters. (a) Results of choosing different values of the guidance channel $C_g$; (b) Results of using different ratios of $\lambda_1$ and $\lambda_2$.

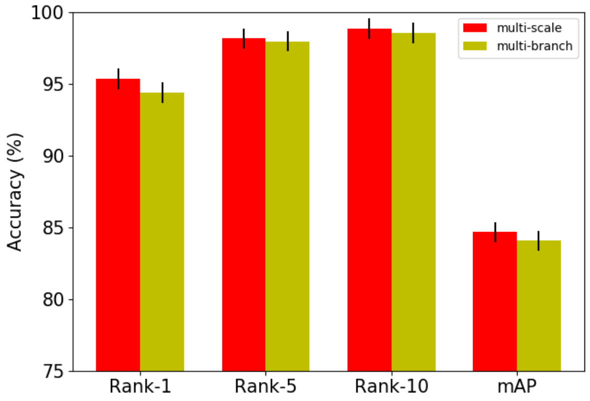


Figure 6. Results of comparing "multi-scale" with "multi-branch' on Market-1501 dataset.

the importance between classification loss and triplet loss. If $\lambda_1 : \lambda_2 > 1$, the features tend to have more distinct representations of a person. If $\lambda_1 : \lambda_2 < 1$, the features have a potential for concentrating on maximizing inter-class and minimizing intra-class distances. Figure 5(b) shows that it is beneficial for re-id feature learning when both classification loss and triplet loss are used, and for assigning the former a higher weight.

**Analysis of the number of scales ($S$) and parts ($M$).** The multi-scale stream layer is proposed here to identify two person images from coarse to fine with multi-scale features, and the global and local branch layer is adopted to exploit the complementarity between global and local (part) features. Intuitively, if the number of scales ($S$) and parts ($M$) is small, for example being one, then some fine information may not be considered for identification, however, too many scales and parts require huge computation resources. Therefore, what is the optimum number of scales and parts? To answer this question, we conduct extensive experiments to gradually increase the number from 1 to 4, as shown in Table 12. Note that we do not take more than 5 scales or parts into consideration because of the large memory cost, and especially, $M = 1$ is equivalent to using only the global feature. A number of observations can be made from Table 12. (1) From left to right, the improvements in accuracy become smaller and smaller, or even degrade. From top to bottom, the performance gradually improves, as expected. (2) Comparing the results of $S = 1, M = 1$ with $S = 1, M = 2$, we can find that there is a large jump in Rank-1 accuracy when both global and local features are used, which indicates the importance of our proposed global and local branch layer. (3) Furthermore, comparing the second observation with the results of $M = 1$ and $M = 2$, the gain gradually becomes tiny with an increase of $S$, which intuitively shows that our multi-scale stream layer helps to analyze the data from coarse to fine, and from global to local with multiple scales; but too many scales would not be helpful. (4) We notice that the result of $S = 4, M = 1$ is a special case in the above observations. We explain that having a value of $S$ that is too large and missing the complementarity between global features and local features may make the results inconsistent. Lastly, (5) to balance the efficiency and effectiveness, the best optimal combination we apply in this work is $S = 3$ and $M = 3$. For better computational efficiency, we can choose $S = 2$ and $M = 3$ as an alternative.

**Multi-scale vs. multi-branch.** Our multi-scale stream layer is designed based on a multi-branch structure to exploit features from multiple scales. Consequently, there may be a doubt about whether the contribution of our proposed layer comes from the multi-branch instead of the multi-scale. To this end, we formulate two contrastive models: one is our MuDeep; another is similar to MuDeep, except that all branches in the multi-scale stream layer have the same structure of scale 3 in Table 1. The results in Figure 6 show that our MuDeep is approximately 1% higher than multi-branch in both

| Rank-1/mAP | | $M$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| $S$ | 1 | 89.10/74.98 | 94.51/83.11 | 94.59/82.98 | 94.38/81.78 |
| | 2 | 90.14/76.61 | 94.53/83.70 | **95.33/84.54** | 94.53/83.18 |
| | 3 | 93.64/82.54 | 94.65/84.20 | ***95.34/84.66*** | 94.68/83.65 |
| | 4 | 82.15/66.06 | 94.92/84.53 | 94.92/84.03 | 94.74/83.21 |

Table 12
Results of comparing a various number of scales and parts on Market-1501 dataset.

| Methods | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| Saliency+Fusion | 82.56 | 93.88 | 97.35 | 98.27 |
| Guidance+Attention+Fusion | 83.01 | 94.77 | 98.09 | 98.72 |
| MuDeep | **84.66** | **95.34** | **98.16** | **98.73** |

Table 13
Results of comparing the leader-based attention learning layer with the saliency-based learning fusion layer on Market-1501 dataset.

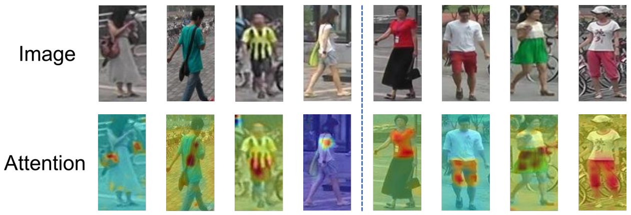


Figure 7. Some examples of small scale and large scale features learned by our model. A warmer color indicates a higher activation value.

Rank-1 and mAP accuracy. This indicates that our proposed multi-scale stream layer efficiently explores data from coarse to fine to help solve the re-id problem.

**Leader-based attention learning layer vs. saliency-based learning fusion layer.** One of our contributions is to improve the saliency-based learning fusion layer in [39] and propose a leader-based attention learning layer. In order to verify gains in performance, we first do some summaries about these two attention layers. Thus, the saliency-based learning fusion layer consists of two parts: "Saliency" and "Fusion". The former aims to choose essential features and remove redundant information, and the latter merges filtered features for final re-identification. The motivation of this module is to select discriminative features from different scales. However, once the model is trained, the weight for each scale is fixed, which is clearly suboptimal as the weights should be adaptive according to the model inputs. To overcome this shortcoming, we propose a leader-based attention learning layer. This is also composed of two parts: "Guidance" and "Attention". The former is applied to fuse comprehensive information to guide each scale branch so that the attention is not be confused by partial details, and the latter is a self-attention mechanism which refines features at each scale. Then, we conduct experiments to analyze the contribution of these two attention layers. More specifically, we first replace the leader-based attention learning layer with the previously proposed saliency-based learning fusion layer. It (referred to as "Saliency+Fusion") achieves 93.88% and 82.56% accuracies on Rank-1 and mAP, respectively, which is inferior to the results reported in this paper by a margin of 1.46% and 2.1%, respectively, as shown in Table 13. This validates the efficacy of our proposed leader-based attention learning layer. Furthermore, we also add the part of "Fusion" after the leader-based attention learning layer, and it (referred to as "Guidance+Attention+Fusion") obtains 0.89% and 0.54% higher accuracies at Rank-1 and mAP than the "Saliency+Fusion". This result further suggests that our proposed leader-based attention learning layer brings clear benefits.

**Visualization of attention mechanism.** In order to illustrate the intuition for the leader-based attention mechanism, eight query examples are shown in Figure 7. We have small scale feature activations in the first four examples, and large scale feature activations in the last four examples. It is shown that the large scale features are most concentrated in the regions containing homogeneous colors and textures, *e.g.*, the torso of the person. In contrast, the small scale features are focused on the regions containing small objects, such as accessories that make the person stand out, *e.g.*, the area of a short sleeve, a handbag strap, and unique patterns on clothing.

## 5 Conclusion

In this paper, inspired by the human cognitive process, we have proposed a novel deep architecture – MuDeep. Different from previous re-id approaches, MuDeep exploits person features that are extracted from multiple scales. Further, a novel leader-based attention learning layer is proposed to utilize all information as the leader to dynamically guide analyses of important regions for each specific-scale data stream. Extensive experiments on several benchmarks show that our model achieves state-of-the-art performance. Finally, a more realistic domain generalization setting is considered in our work. We show that under this setting, the advantages of our model over state-of-the-art alternatives are even more pronounced. This suggests that the multi-scale features extracted using MuDeep are more generalizable to novel domains.

## Acknowledgments

This work was supported in part by NSFC Projects (61702108, 61622204, 61572138), STCSM Project (16JC1420400), Shanghai Municipal Science and Technology Major Projects (2017SHZDZX01, 2018SHZDZX01) and ZJLab.

## References

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

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

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

[4] A. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *ICCV*, pp. 545 –551, 2009. 1

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

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

[7] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric bayesian model," in *CVPR*, 2008. 1

[8] 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, 4.1

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

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

[11] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014. 1

[12] 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, 4.1, 4.1, 4.3

[13] 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

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

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

[16] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012. 1, 2.4, 4.1

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

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

[19] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," in *IEEE TIP*, 2014. 1, 2.4

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

[21] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person reidentification using kernel-based metric learning methods," in *ECCV*, 2014. 1

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

[23] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015. 1, 2.1, 2.4, 4.1, 4.1

[24] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *ICCV*, 2017. 1, 2.3, 4.1, 4.3

[25] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014. 1, 2.1, 4.1, 4.1, 3, 4.1

[26] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, 2018. 1, 2.1, 3.1, 4.3, 4.3

[27] 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, 2.1, 1, 4.1, 4.1

[28] 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, 4.1, 4.1

[29] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 3652–3661, IEEE, 2017. 1, 2.1, 3, 7, 8, 4.3

[30] 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, 1

[31] M. Mermillod, N. Guyader, and A. Chauvin, "The coarse-to-fine hypothesis revisited: evidence from neuro-computational modeling," *Brain and Cognition*, vol. 57, no. 2, pp. 151–157, 2005. 1

[32] B. Musel, A. Chauvin, N. Guyader, S. Chokron, and C. Peyrin, "Is coarse-to-fine strategy sensitive to normal aging?," *PloS one*, vol. 7, no. 6, p. e38493, 2012. 1

[33] P. Vuilleumier, J. L. Armony, J. Driver, and R. J. Dolan, "Distinct spatial frequency sensitivities for processing faces and emotional expressions," *Nature neuroscience*, vol. 6, no. 6, p. 624, 2003. 1

[34] D. M. Parker, J. R. Lishman, and J. Hughes, "Role of coarse and fine spatial information in face and object processing.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 6, p. 1448, 1996. 1

[35] D. M. Parker, J. R. Lishman, and J. Hughes, "Temporal integration of spatially filtered visual images," *Perception*, vol. 21, no. 2, pp. 147–160, 1992. 1

[36] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," *Psychological science*, vol. 5, no. 4, pp. 195–200, 1994. 1

[37] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017. 1, 2.3, 3.3

[38] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018. 1, 3.3

[39] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *ICCV*, 2017. 1, 4.3, 9, 4.4, 4.4

[40] X. Qian, Y. Fu, W. Wang, T. Xiang, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," *ECCV*, 2018. 2.1, 3.6, 3, 4.2, 4.3, 4.3

[41] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3980–3989, IEEE, 2017. 2.1, 4.1, 4.3

[42] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *arXiv preprint arXiv:1701.07732*, 2017. 2.1, 4.3

[43] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. CVPR*, vol. 2, 2017. 2.1, 4.1, 4.1

[44] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. 2.1, 3.5, 4.4

[45] 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. 2.1, 2.2

[46] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," in *Pattern Recognition*, 2015. 2.1

[47] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2.1

[48] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *IEEE International Conference on Computer Vision*, 2017. 2.1

[49] X. T, W.Ouyang, H. Li, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016. 2.1, 4.1

[50] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *arxiv*, 2016. 2.1

[51] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," *CVPR*, 2018. 2.1, 3, 4.1, 4.1, 4.3

[52] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016. 2.1, 4.1, 4.3

[53] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *CVPR*, 2018. 2.1

[54] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," *ArXiv e-prints*, Apr. 2018. 2, 3.4, 3.6, 4.1, 4.2, 4.3, 4.3, 4.3, 4.3

[55] C. Shen, Z. Jin, Y. Zhao, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua, "Deep siamese network with multi-level similarity perception for person re-identification," in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1942–1950, ACM, 2017. 3

[56] 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. 2.2

[57] Y. Chen, X. Zhu, S. Gong, *et al.*, "Person re-identification by deep learning multi-scale representations," *ICCVW*, 2017. 2.2, 4.1

[58] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8042–8051, 2018. 2.2, 4.1

[59] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," in *Nature reviews neuroscience,*, 2002. 2.3

[60] 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. 2.3

[61] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 369–378, 2018. 2.3

[62] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," *arXiv preprint arXiv:1708.02286*, 2017. 2.3

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017. 2.3

[64] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," *arXiv preprint arXiv:1804.09337*, 2018. 2.3

[65] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," *CVPR*, 2018. 2.3, 4.3, 4.3

[66] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, vol. 1, p. 2, 2018. 2.3, 4.1, 4.3, 4.3

[67] G. Lisanti, I. Masi, and A. D. Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *ICDSC*, 2014. 2.4

[68] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *IEE AVSS*, 2012. 2.4, 4.1

[69] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012. 2.4

[70] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Asian conference on Computer vision*, pp. 501–512, Springer, 2010. 2.4

[71] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1846–1855, 2015. 2.4

[72] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," in *Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, pp. 25–32, ACM, 2013. 2.4

[73] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 8, pp. 1447–1460, 2016. 2.4

[74] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016. 2.4

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2015. 3.1, 3.2, 4.3

[76] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2017. 3.1

[77] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017. 3.1, 3.4, 4.1, 4.3

[78] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li, "Constrained deep metric learning for person re-identification," *arXiv preprint arXiv:1511.07545*, 2015. 3.4

[79] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," *arXiv preprint arXiv:1407.4979*, 2014. 3.4

[80] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013. 4.1, 4.1

[81] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 144–151, 2014. 4.1

[82] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *TIP*, 2016. 4.1

[83] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1363–1372, 2016. 4.1

[84] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification.," in *AAAI*, vol. 1, p. 3, 2017. 4.1, 4.1

[85] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1085, 2017. 4.1, 4.3

[86] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Advances in Neural Information Processing Systems*, pp. 2667–2675, 2016. 4.1, 4.1

[87] L. Zhang and T. X. S. Gong, "Learning a discriminative null space for person re-identificatio," in *CVPR*, 2016. 4.1, 4.3

[88] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 54–70, 2018. 4.1, 4.1, 4.3

[89] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017. 4.1, 4.1, 4.3

[90] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification.," in *IJCAI*, pp. 3545–3551, 2016. 4.1, 4.3

[91] C. Mao, Y. Li, Y. Zhang, Z. Zhang, and X. Li, "Multi-channel pyramid person matching network for person re-identification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 4.1, 4.1

[92] L. Zheng, L. Shen, L. Tian, S.Wang, J.Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015. 4.1, 1, 4.1

[93] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017. 4.1, 2, 4.1, 4.3, 4.3

[94] W. Li, R. Zhao, and X.Wang, "Human re-identification with transferred metric learning," in *ACCV*, 2012. 4.1, 4

[95] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera

tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 2

[96] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, pp. 1627–1645, 2010. 3

[97] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009. 4.1

[98] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *IJCAI*, 2017. 4.1, 4.3

[99] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 188–204, 2018. 4.1, 4.3, 4.3

[100] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 365–381, 2018. 4.1, 4.1, 4.3, 4.3

[101] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 486–504, 2018. 4.1, 4.3, 4.3

[102] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009. 4.1

[103] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017. 4.1

[104] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016. 4.1

[105] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *arXiv preprint arXiv:1707.00408*, 2017. 4.1

[106] Y. Sun, L. Zheng, D. Weijian, and W. Shengjin, "Svdnet for pedestrian retrieval," in *ICCV*, 2017. 4.1, 4.3

[107] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *CVPR*, vol. 1, p. 2, 2018. 4.1, 4.3, 4.3

[108] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," *arXiv preprint arXiv:1711.09349*, 2017. 4.1, 4.3, 4.3

[109] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person reidentification," in *ECCV*, 2016. 4.3

[110] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016. 4.3

[111] Z. Zheng, L. Zheng, , and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," in *arXiv:1611.05666*, 2016. 4.3

[112] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," in *arXiv:1611.0524*, 2016. 4.3

[113] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," *CVPR*, 2018. 4.3

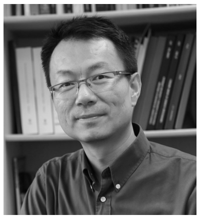
[114] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, vol. 6, p. 7, 2017. 4.3

[115] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *arXiv preprint arXiv:1711.10658*, 2017. 4.3, 4.3

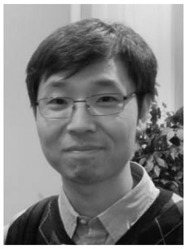
[116] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017. 4.3

**Xuelin Qian** received the BSc degree in Mathematics and Applied Mathematics from Xidian University in 2015. He is studying the fourth year of PhD in Fudan University and working in the area of machine learning and computer vision. His research interest is face recognition, object detection, and person re-identification.

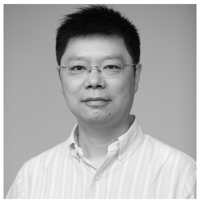
**Yanwei Fu** received the Ph.D. degree from Queen Mary University of London in 2014, and the M.Eng. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2011. He held a post-doctoral position at Disney Research, Pittsburgh, PA, USA, from 2015 to 2016. He is currently a tenure-track Professor with Fudan University. His research interests are image and video understanding, and life-long learning.



**Tao Xiang** is a Professor of Computer Vision and Machine Learning and Distinguished Chair at the University of Surrey. He is also a Principal Researcher at Samsung AI Centre, Cambridge where he leads the Body Behaviour Group. Xiang's research in computer vision has focused on video surveillance, daily activity analysis, and sketch analysis. He also has interests in large-scale machine learning problems including zero/few-shot learning and domain generalization.



**Yu-Gang Jiang** is a Professor of Computer Science at Fudan University and Director of Fudan-Jilian Joint Research Center on Intelligent Video Technology, Shanghai, China. He is interested in all aspects of extracting high-level information from big video data, such as video event recognition, object/scene recognition, and large-scale visual search. His work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015 ACM SIGMM Rising Star Award, and the research award for outstanding young researchers from NSF China. He is currently an associate editor of ACM TOMM, Machine Vision and Applications (MVA) and Neurocomputing. He holds a PhD in Computer Science from City University of Hong Kong and spent three years working at Columbia University before joining Fudan in 2011.



**Xiangyang Xue** received the BS, MS, and PhD degrees in communication engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively. He is currently a professor of computer science with Fudan University, Shanghai, China. His research interests include computer vision, multimedia information processing, and machine learning.