Multi-level and Multi-scale Horizontal Pooling Network for Person Re-identification

Yunzhou Zhang,[a,*] Shuangwei Liu,[a] Lin Qi,[a] Sonya Coleman,[b] Dermot Kerr,[b] Weidong Shi[a]

[a] College of Information Science and Engineering, Northeastern University of China, Shenyang, China, 110819
[b] Intelligent Systems Research Centre, University of Ulster, Derry, United Kingdom, BT52 1SA

Abstract. Despite recent remarkable progress, person re-identification methods either are subject to failure when key body parts are missing or are applied to a range of images of varying complexity. To mitigate these issues, we introduce a simple yet effective Multi-level and Multi-scale Horizontal Pooling Network (MMHPN) for person re-identification. The research contributions are three-fold: 1) we take a partial feature representation into account at different pooling scales and different semantic levels so that partial information is obtained to increase robustness in cases where discriminative parts are missing; 2) we propose an Adaptive Pooling Strategy (APS) as a weighted summation of global average pooling and global max pooling, which can further improve the discriminability of partial features; 3) we introduce a Part Sensitive Loss (PSL) function to reduce the effect of easily classified partition to facilitate training of the person Re-ID network. Experimental results using the Market-1501, DukeMTMC-ReID and CUHK03 datasets demonstrate that the proposed MMHPN outperforms state-of-the-art methods. Specifically, we achieve mAP scores of 83.4%, 75.1% and 65.4% on these challenging datasets.

Keywords: multi-level and multi-scale, horizontal pooling network, adaptive pooling strategy, part sensitive loss, person re-identification.

***First Author**, E-mail: zhangyunzhou@mail.neu.edu.cn

1 Introduction

Given a person-of-interest query, the goal of a person re-identification (Re-ID) algorithm is to retrieve images containing the same person in a specified pedestrian database captured across several different security cameras. Automatic person re-identification has recently attracted much attention and has become an important component in modern video surveillance systems. Despite progress in this area, person re-identification is still a challenging problem in complex situations such as occlusions, low resolution, and large variation in posture. In these situations, visual cues can be dramatically different. To address these challenges, powerful deeply-learned representations[1,2,3,4] have been widely applied and have obtained promising performances compared with hand-crafted approaches[5,6,7].

The traditional approach employed by deeply-learned representations is to extract global features from a person's body. However, the process of global feature extraction can lead to a problem that non-salient regions are easily ignored and do not contribute to improved discrimination. To address this issue, many approaches learn a discriminative partial representation and this has proven to be more effective than a global feature approach in person Re-ID accuracy. Recent state-of-the-art part-based methods for person re-identification can be categorized into three groups: 1) prior knowledge[2,8,9] such as pose estimation is utilized as structural information to locate partial regions. Nevertheless, as the off-the-shelf pose estimation model predicts unexpected body landmarks, the performance of Re-ID is inevitably influenced. 2) Attention-based methods [3,10,11] focus on enhancing features in salient parts while the selected parts lack semantic information. 3) Region based methods[4,12] produce bounding boxes for locating parts, but the proposed parts typically have fixed semantics and cannot represent all possible discriminative parts. Additionally, these methods use the output of the final convolution layer as a representation in order to distinguish the person's identity, which mainly consists of high-level semantic features and discards mid-level semantic features.

For the person re-identification task, there are two types of loss functions, the metric loss[13,14,15] and the softmax loss[1,3,4]. If an image pair belongs to the same identity or not, it is just a weak label and the metric loss may have a compromised efficiency when using a large database. In contrast, softmax loss leverages image labels to supervise the training of network parameters, which results in a superior accuracy using both the Market-1501[17] and DukeMTMC-ReID[18] datasets. However, both these loss functions give the same importance to each sample (image level) or stripe (local level), ignoring the image complexity. .

In this paper, we focus on partial discriminative features to enhance the performance of person Re-ID. Inspired by two feature learning strategies, Horizontal Pyramid Matching (HPM)[19] and Devil in the Middle (DIM)[20], we propose a simple yet effective Multi-level and Multi-scale Horizontal Pooling Network (MMHPN) to fully exploit global information in the high semantic level and partial information in the middle semantic level. Specifically, we make the following three contributions.

1) We horizontally slice the deep feature maps, produced by different convolutional layers, into various sizes of partition stripes for multi-level and multi-scale pooling as shown in Figure 1. We then learn to classify each partial stripe independently. Intuitively, integrating multi-level semantic information from different layers tends to enhance the capability of being context-invariant whilst learning multi-scale information by pyramid pooling can improve the discriminative feature of a person. We combine the strengths of the above two strategies, thus making global and local feature representation more robust and discriminative.

2) As an alternative to max/average pooling in each partition, we propose an Adaptive Pooling Strategy (APS) as a weighted summation of Global Average Pooling (GAP) and Global Max Pooling (GMP) to automatically balance the significance of each. Max pooling focusses on the salient local features but fails to exploit the available global information of the person under consideration. Average pooling represents global information yet can easily lead to over-estimated body regions. Adaptively combining them not only utilizes their complementary abilities to enhance the discrimination of features but also balances the effectiveness between global and local information.

3) We propose a Part Sensitive Loss (PSL) to give more importance to difficult scenarios and partition stripes during training. The motivation is that easy examples or stripes should not

dominate when updating the network. Therefore, we decrease the contribution of easy examples to facilitate the training of the person Re-ID network.

We conduct performance evaluation using current benchmark person re-identification datasets and demonstrate that the proposed method can achieve state-of-the-art performance. In particular, the mAP scores on the Market-1501, DukeMTMC-ReID and CUHK03 datasets are 83.4%, 75.2%, 65.4%, respectively.

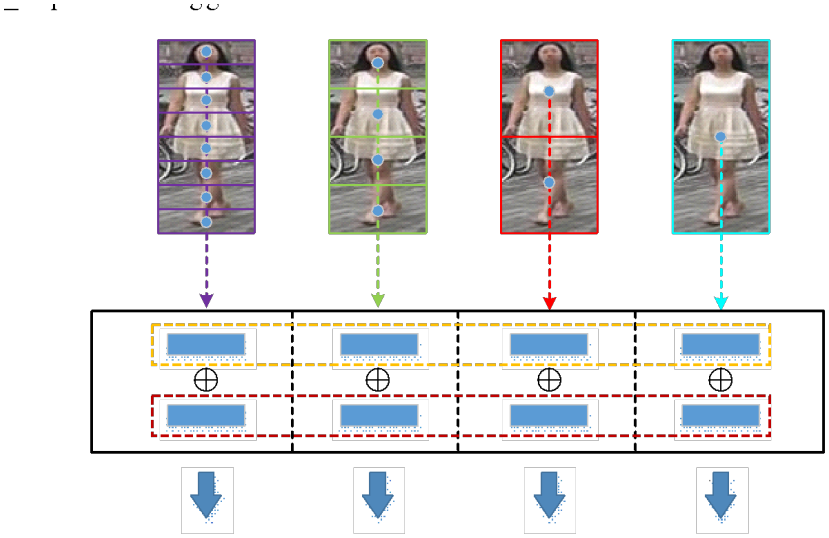

Fig. 1 Illustration of the proposed Multi-level and Multi-scale Horizontal Pooling Network. The person is split into different horizontal stripes from multiple scales and multiple semantic levels. The feature representation produced by APS of each stripe are then utilized to learn a person's identity independently. Note that $\oplus$ stands for a weighted summation between GAP and GMP.

2 Related Work

The success of deep learning methods was initially extended to the person Re-ID community in 2014, when person Re-ID works[21,22] first considered a Siamese network architecture using pairs

of images to learn the latent features of human body parts. The performance of this approach surpassed existing hand-crafted Re-ID methods. A number of methods based on deep convolutional neural networks have been proposed for simultaneously learning both feature representations and distance metrics. For example, Ahmed et al.,[13] improved the Siamese network by measuring feature similarity using a subtraction between the features of one input image and the features in a nearby location of another image. Recent works have also started to exploit the effectiveness of mid-level features. Yu et al.,[20] proposed a simple approach where the mid-layer and final-layer feature maps are fused into a single representation, followed by a softmax function to predict person identity. Compared with [20], each horizontal feature map in the proposed mid- and high-level semantic branches are supervised by an independent part sensitive loss rather than fusing all the feature maps followed by a softmax loss.

More recently, person Re-ID methods based on deep learning have demonstrated improved performance over previous approaches. Zhang et al.,[23] utilise part alignment by matching the shortest path as well as manual learning in distance metrics to facilitate global feature representation. Sun et al.,[24] cropped the feature maps into six stripes in the vertical orientation to represent local parts and concatenated them as a final feature representation. Unfortunately, it assumes that images containing people are well aligned and thus the approach is prone to errors due to outliers. To tackle this issue, Li et al.,[19] proposed a Horizontal Pyramid Matching (HPM) network to mitigate the outlier issue by incorporating a slack distance, yet it slices feature maps into stripes at the same convolutional layer. In the proposed MMHPM approach, we split feature maps into four scales at different convolutional layers to fuse mid- and high-level semantic features for effectively distinguishing person identity.

Applying this strategy, we argue for performance improvement, we can draw the conclusion that combining local representations of body parts is the most effective way to enhance the discriminative ability of the model. As discussed in Section 1, we divide deep part-based methods into three categories: The first one utilizes extra tools such as pose estimation and landmark detection[2,8] to parse pedestrians. In particular, Su et al.,[2] formulate a Pose-driven Deep Convolutional (PDC) module and feature weighted sub-network to overcome pose deformations and view variations. Secondly, several works integrate attention mechanism into person Re-Id for salient parts and report encouraging improvements[3,10,11]. Thirdly, region based methods [25,26] are exploited to locate semantic parts in several part-based methods[27,12]. Yao et al.,[12] propose Part Loss Networks which automatically generates a set of boxes as body parts in an unsupervised manner and learn each part for person classification independently. In the proposed method, we only use simple horizontal stripes in the multi semantic level and we use multi pyramid scales as part regions for local feature learning.

Generally, two types of loss function are used as supervisory signals for person re-identification: metric and classification loss. For the first metric loss used in embedding learning, the person Re-ID is considered as a ranking problem in which a pair or triplet of images is fed into a Siamese-like network. For instance, the contrastive loss[28] is used in a verification network[13,16,17] to determine whether a pair of images is similar or not, which encourages the network to make the distance of intra-class pairs closer and push the images of inter-class further apart. This loss function is effective and is suitable for a person Re-ID task as it naturally reduces the intra-personal variations due to its retrieval nature. However, the performance of this kind of model is limited by a large pedestrian database. Whether a pair of images is similar or not, is just a weak label and it does not take full advantage of identity annotations in Re-ID. Unlike aforementioned approaches,

other works[4,21,24] treat person Re-ID not as a ranking problem, but rather as a recognition problem. Due to its strong robustness to a variety of multi-class classification tasks, softmax loss remains the overwhelming choice as a supervisory signal. Besides, integrating metric loss and classification loss may be accepted as a way to improve the overall performance of person Re-ID. In [23, 29, 30] authors adopt triplet loss and softmax loss as joint supervision to train a convolutional neural network (CNN), achieving promising performance when using the benchmark datasets. However, the aforementioned loss functions all treat both simple and complex images equally in the training phase.

3 Method

In this section, we first present an overview of the proposed Multi-level and Multi-scale Horizontal Pooling Network (MMHPN). Then we show details on the proposed Multi-level and Multi-scale Horizontal Pooling framework and Part Sensitive Loss.

3.1 Overview of network

The structure of the MMHPN is shown in Fig 2. The images containing people are input into the backbone network to extract feature maps. Then the feature maps undergo res_conv5_1, res_conv5_2 and res_conv5_3 blocks in turn (see Figure 2) and we define this straightforward stream with high-level information as a global branch. During this process, in order to make the extracted feature maps discriminative at different semantic levels, we introduce two local branches at the middle semantic level, which are immediately after the res_conv5_1 and res_conv5_2 blocks respectively. The global branch and local branches are not separated, thus they are complementary for learning feature embedding. We utilize different scales of the horizontal pooling module to capture spatial information in global and local stripes. For each horizontal stripe, we transform

the feature maps to vectors by using an APS, which automatically balances the importance between the GAP operation and GMP operation. Then the MMHPN leverages a non-share convolution layer to reduce the channel dimensions from 2048 to 256. Finally, each column feature vector is input into a classifier independently, which consists of a non-shared fully-connected (FC) layer and a softmax layer, to predict the ID of each input image. During training, the MMHPN is supervised by minimizing the summation of part sensitive losses over global and local branches of ID predictions. At the testing phase, we concatenate all feature vectors to form a 3780-dimension descriptor containing information at different semantic levels and pooling scales.

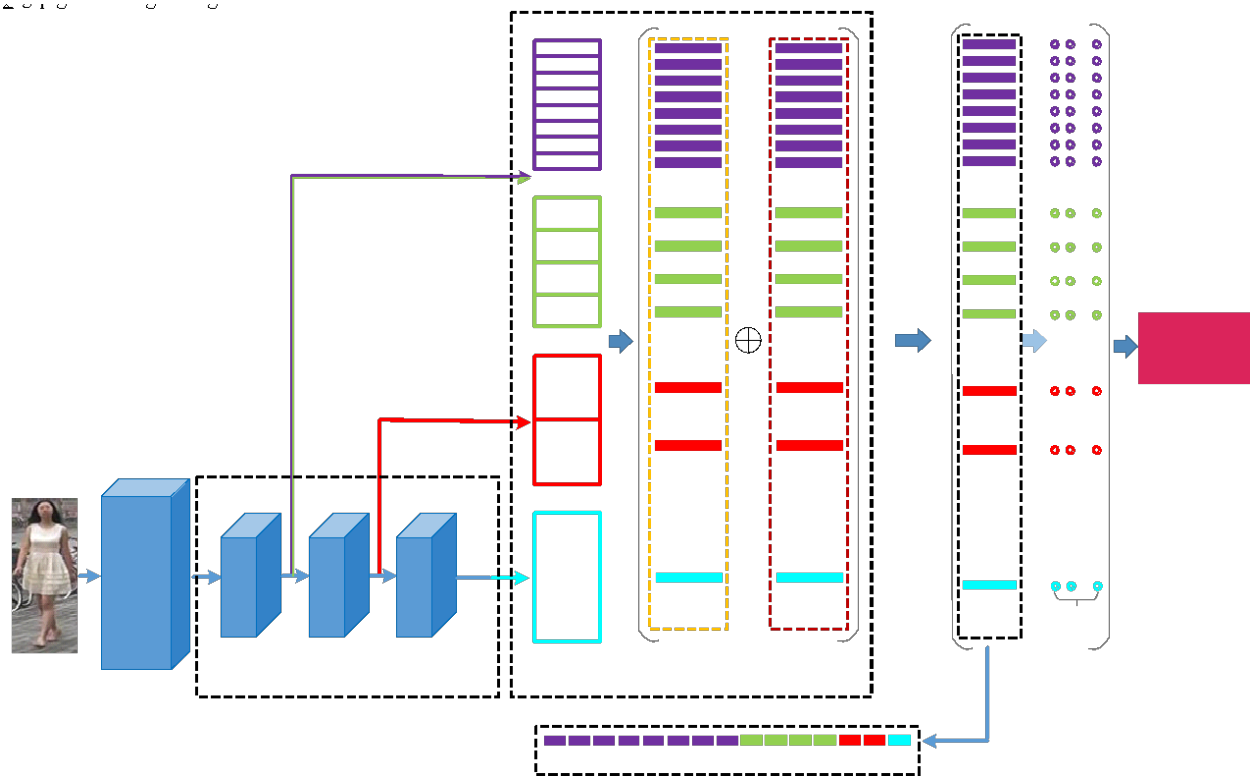

Fig. 2 Overview of Multi-level and Multi-scale Horizontal Pooling Network. The input image firstly goes forward through a ResNet-50 (before res_conv5) to extract feature maps. Then two local branches right after res_conv5_1 and res_conv5_2 in middle semantic level are introduced to boost discriminative information. Afterwards, different scales of pooling are used to produce feature representation of each stripe using adaptive pooling strategy. Finally, we assign

each stripe an independent classifier to predict a partial-level person identity. During testing, we concatenate features of stripes at different scales and semantic levels to form the final representation of each image.

3.2 Multi-level and Multi-scale Horizontal Pooling Module

3.2.1 Backbone Network

The backbone of the proposed MMHPN is ResNet-50[31], with a relatively concise architecture to obtain competitive performance compared with other person Re-ID systems[23,29] and to be consistent with previous methods[24,19] for a fair comparison. There are some slight modifications from the original ResNet-50. Firstly, the global average pooling layer and subsequent layers are removed. Additionally, the stride of the res_conv4_1 block is set to 1 which enlarges the feature maps from 1/32 to 1/16 of the original image size for more abundant spatiality and granularity of detected features. Finally, two local branches immediately after res_conv5_1 and res_conv5_2 are added to learn mid-level semantic features as illustrated in Figure 2.

3.2.2 Multi-level Semantic Module (MSM)

An effective person Re-ID model should possess the capability of extracting discriminative features at different semantic levels. However, most existing person Re-ID systems take direct advantage of deep neural networks, typically designed for object recognition, and employ a final layer output with high-level semantic features as a representation. As a result, the mid-level features are missed and cannot help to distinguish identity effectively. Therefore, it is essential for deep person Re-ID to mix mid- and high-level semantic features in a fusion module.

We adopt multi-level semantic branches in Fig. 2 where we define upper and middle branches with local information as mid-level semantic branches. We denote the lower branches that contain global information as high-level semantic branches. Moreover, the two mid-level branches are also

regarded as auxiliary classifiers connected to res_conv5_1 and res_conv5_2 blocks to increase the gradient signal, encourage mid-level discrimination, and provide additional regularization.

As shown in Fig 3, when we narrow the area of represented regions to learn local features, we can observe that the network encourages response of local attention maps starts to cluster on some salient semantic patterns which are not exploited in a global attention map. From the aforementioned observations, we can conclude that body parts at small scales tend to learn mid-level details such as head, pants and shoes, and body parts at larger scales will exploit high-level semantic information. Hence, we slice feature maps into 8, 4 and 2 stripes at local branches to enable the system to focus on mid-level information as in Fig 2. The global branch contains the complete feature map without any partition information. As a result, the global and local representation with high and mid-level semantic information respectively are combined to form the final feature descriptor for the discriminative person Re-ID model.

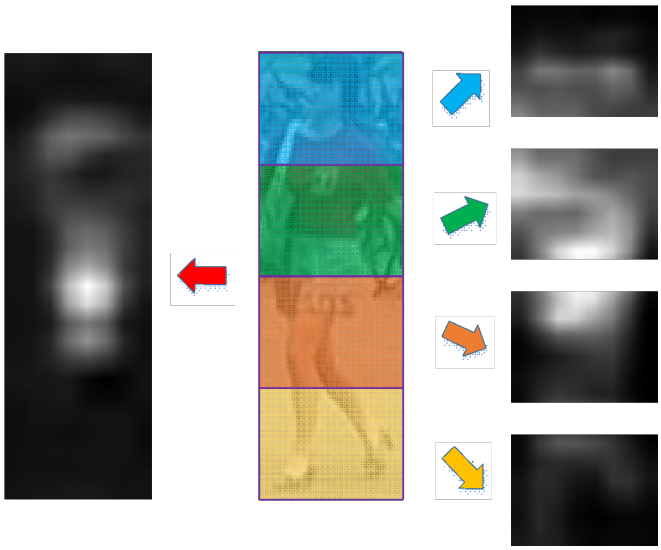

Fig. 3 Attention maps in different scales produced by the last convolutional output of different models. Second Column: a person image. First Column: global attention map by IDE model. Third Column: four local attention maps corresponding to four spilt stripes of the person image, produced from part-based model.

3.2.3 Multi-scale Horizontal Pooling Module (MHPM)

Due to the requirement of a fully connected layer needing a fixed-length vector, the convolutional neural network is obliged to accept a fixed-size input image, which is often contradictory to pedestrian images as well as images of the body parts which typically are of varying sizes or ratios. As a result, the recognition accuracy is compromised due to the limited scale of the input image.

To eliminate the requirement for arbitrarily sized input images, we adopt the concept of HPM[19] which splits and pools the feature maps at four scales. As shown in Fig 3, the proposed multi-scale pooling module has four scales and the output feature maps of each scale are split into 8, 4, 2, and 1 stripe(s) respectively and the adaptive pooling strategy is applied to each stripe. By applying multi-scale horizontal pooling, we can obtain both a fixed-dimensional vector for body parts with varying sizes, and also capture a discriminative representation of partition from global to local, from high-level to mid-level and from coarse to fine. Moreover, global and local cues are stacked in four scales together, which contributes to making the final predictions more reliable, especially when the key parts are missing.

3.2.4 Adaptive Pooling Strategy

Average pooling usually exploits the global extent of the body, yet it is prone to over-estimate the unrelated background, while max pooling focusses on the most salient local region but lacks discriminative information. Intuitively, adaptively integrating GAP and GMP is appropriate to automatically balance the effectiveness between global and local information in each stripe. To this end, we propose an Adaptive Pooling Strategy (APS) using the weighted summation of the GAP and GMP operations. Specifically, the proposed MMHPN has four scales and in each scale the feature maps inherited from backbone network *B* are divided into a specific number of stripes

horizontally and equally. We denote the *j*-th stripe in the *i*-th scale as $B_{i,j}$. Note that the indexes, *i* and *j*, are incremented from top to bottom. Then, each stripe is pooled by a weighted sum of the GAP and GMP operations to obtain the feature vector $V_{i,j}$.

$$V_{i,j} = avgpool(\mathrm{B}_{i,j}) + \omega_i \,\mathrm{maxpool}(\mathrm{B}_{i,j}), \tag{1}$$

where ω_i denotes learnable parameters and is shared in the same scale. We also try to set each stripe in the same branch to share the parameter ω, but the Re-Id model fails to converge due to the obvious difference of stripe scale in the upper branch. With this design, each partition at different scales could voluntarily choose to focus more on either global information or local information by adaptively adjusting the learning parameter ω. Then a convolutional layer is employed to reduce the channel dimensions from 2048 to 256 in each scale. Finally, these reduced dimension vectors $R_{i,j}$, with the same index *i,* are concatenated to obtain the final feature descriptor for the pedestrian images.

3.3 Part Sensitive Loss

With the consideration of strong robustness to a variety of multi-class classification tasks, we employ softmax loss in the proposed MMHPN to unleash the discriminative capability of the deep representation. As discussed in Section 3.1, three branches at different semantic levels have complementary strengths to learn discriminative descriptors for pedestrian images. In order to maximize these complementary effects, the three branches of the network are trained jointly to distinguish person identity in a global and local feature learning manner. We utilize a non-shared

fully connected layer as classifier for each stripe. Specifically, we input each feature column vector $R_{i,j}$ into a corresponding classifier $FC_{i,j}$ and employ a softmax layer to predict person identity. In the training phase, we consider each person as a class, and the MMHPN as a function that maps a given image to a set of predictions $\hat{z_{i,j}}$. Each $\hat{z_{i,j}}$ can be formulated as:

$$\hat{z_{i,j}} = \frac{\exp((W_{i,j}^{y})^{T} R_{i,j}(I))}{\sum_{n=1}^{N} \exp((W_{i,j}^{n})^{T} R_{i,j}(I))}, \tag{2}$$

where y is the ground truth for person identity of input image I, N is denoted as the total number of person identities in the training dataset and $W_{i,j}$ is defined as the learned weight in $FC_{i,j}$. The loss function on this sample is computed by the sum of softmax loss of the predicted probability $\hat{z_{i,j}}$.

$$L_{soft\max} = -\sum_{m=1}^{M} \sum_{i,j} \log(\hat{z}_{i,j}^{m}), \tag{3}$$

where M is the size of the mini-batch in the training phase.

However, one notable problem of using the softmax loss is that each example or stripe is given the same importance during the training process. This results in the softmax loss ignoring how complex the images examples or stripes are. The easy examples $\hat{z_{i,j}} \geq 0.5$ incur the loss with non-trivial magnitude. When summed over a large number of simple examples, these small loss values overwhelm the valuable, whilst, rare and complex examples with sophisticated illumination, deformation and scale variation that can be learned to improve robustness and further enhance the generalization ability. As a consequence, easily classified examples or stripes comprise the

majority of the softmax loss and dominate the gradient. Hence, we propose a Part Sensitive Loss (PSL) to encourage the network to focus more on the complex examples or stripes and decrease the contribution of simples ones when training the classifier. The predicted probability $\hat{z}_{i,j}$ of each stripe with the same index *i* and differing index *j* is employed with a softmax function to obtain the easy and hard degree $d_{i,j}$ across the whole body. Thus $d_{i,j}$ can be defined as:

$$d_{i,j} = \frac{\exp \hat{z}_{\mathrm{i,j}}}{\sum_{p=1}^{P} \exp \hat{z}_{i,\mathrm{p}}}, \tag{4}$$

where *P* is the corresponding number of stripes at each scale. From Eq. (4), we can observe that as the value of predicted probability $\hat{z}_{i,j}$ increases, the hard and easy degree $d_{i,j}$ increases similarly, which means the corresponding stripe in the whole body can be classified more easily. When $d_{i,j}$ for a specific stripe is significantly larger than other stripes, we should reduce the weight of it to decrease the effectiveness of this easy stripe and focus on the hard stripes. Hence, we align the proposed approach to be consistent with focal loss[33] and attach a modulating factor in terms of the hard and easy degree $d_{i,j}$ to the softmax loss. To be specific, we build the PSL upon the softmax loss:

$$L_{part\ sensitive} = \sum_{m=1}^{M} \sum_{i,j} \log(\hat{\mathrm{z}}_{i,j}^{\mathrm{m}})(1-\mathrm{d}_{i,j})^{\alpha_i}, \tag{5}$$

where α_i is a tunable focussing parameter with a range value of [0, 2]. The stripes with the same index *i* and differing index *j* employ the same tunable parameter α_i. We note two properties of the

PSL: 1) When a stripe is considered to be a hard stripe and $d_{i,j}$ is small, the modulating factor ($1-d_{i,j}$) is close to 1 and the loss is not affected. As $d_{i,j}$ tends to 1, the factor ($1-d_{i,j}$) approaches 0, and the loss of the easily classified stripe in the whole body is given a reduced weight. 2) The focussing parameter α_i can reflect the extent to which easy stripes are down-weighted. When the focussing parameter α_i is 0, the Part Sensitive Cross loss is equivalent to the softmax loss. As the focussing parameter α_i is increased, the effect of the modulating factor ($1-d_{i,j}$) is enhanced similarly. Intuitively, the modulating factor ($1-d_{i,j}$) decreases the loss contribution from the easily classified stripe and enforces the requirement where each stripe could receive a lower loss. For instance, when the focussing factor is 2, the part sensitive loss of an easy stripe with $d_{i,j}$ =0.1 is 1% of the softmax loss. This in turn casts a high importance to a hard misclassified stripe whose loss is scaled down by at most 4 times when $d_{i,j}$ <=0.5 and $\alpha_i = 2$.

4 Experiment

We evaluate the proposed method using three benchmark datasets Market-1501[17], DukeMTMC-reID[18,34] and CUHK03[21].

4.1 Dataset and Evaluation Protocol

4.1.1 Market 1501 Dataset

The dataset contains 32,368 pedestrian images with 1,501 different identities captured by six manually installed cameras. The dataset is divided into a training set and a test set, the training set

contains 12,936 images of 751 identities, the testing set contains 3,368 query images and 19,732 images of 750 identities in the gallery. On average, each person has 3.6 corresponding images taken from different angles. These images can be divided into two categories, namely, clipping images and DPM[35] automatically detecting pedestrian image.

4.1.2 DukeMTMC-ReID Dataset

The DukeMTMC-ReID dataset is composed of 36,411 images of 1,812 identities from 8 high-resolution cameras where 1,404 identities appear in more than two cameras and the remaining 408 identities are used as distraction images. Among the 1,404 identities, the dataset randomly selected 16,522 images of 702 identities as a training set, and the remaining 702 are categorized into a test set, including 2,228 query images and 17,661 gallery images. DukeMTMC-ReID is considered to be one of the most challenging Re-ID datasets so far, because it has many common situations with high similarity and also contains huge differences in person with the same identity.

4.1.3 CUHK03 Dataset

This CUHK03 dataset includes 1,467 labeled persons from the CUHK campus, with a total of 14,097 images. Each identity is captured by two disjoint cameras, and each identity has approximately 4.8 corresponding images per view. The annotation of this dataset includes manually tagged pedestrian bounding boxes and automatic detections by DPM. We conduct the performance evaluation based on the latter.

4.1.4 Evaluation Protocol

We perform a standard evaluation protocol on each dataset. In order to evaluate the performance of the proposed person Re-ID method, we report the Cumulative Matching Characteristics (CMC) in terms of Rank-1 accuracy and mean Average Precision (mAP) for all candidate datasets. CMC

represents the accuracy of the pedestrian search and is accurate when there is only one ground truth in each query. However, when multiple ground truths exist in the gallery CMC may not have sufficient discrimination and often uses mAP to reflect recalls. For the DukeMTMC-ReID and CUHK03 datasets, the evaluation is performed in a single query mode. As for Market-1501, the experiments are conducted both using single query and multiple-query settings. Meanwhile, to simplify the evaluation procedure using the CUHK03 dataset, we adopt the new protocol used in [36]. Note that all the results are reported without using the re-ranking proposed in [36].

4.2 Implementation Details

The proposed MMHPN model is trained and fine-tuned using the Pytorch framework. For the backbone network, we adopt the ResNet-50 model with weights from a pre-trained ImageNet. During training, we only employ horizontal flipping to train pedestrian images for data augmentation. in order to obtain an appropriately sized feature map for multi-scale and multi-level pyramid pooling, all the training images are resized to 384x128 pixels. We set the mini-batch size to 64 for all experiments and trained the model for 60 epochs in total. With respect to the learning rate strategy, we set the learning rate to 0.1, and decay it to 0.01 after 30 epochs. As for the optimizer, stochastic gradient descent (SGD) with momentum 0.9 and weight decay factor 0.0005 was selected to update the parameters in each mini-batch. The focussing factor α_i in PSL is set to 0.3, 0.5, 1.5, 2.0 respectively.

4.3 Comparison with State-of-the-Art Methods

We compare the proposed method, MMHPN, with current state-of-the-art approaches using three benchmark datasets to demonstrate the improved performance achieved by the proposed method. The experimental results are detailed in the following sub-sections.

4.3.1 Evaluation using Market1501

The compelling results using the Market-1501 dataset are shown in Table 1. the proposed MMHPN achieves a mAP of 83.4% and Rank1 accuracy of 94.6%, which improves the former method by 0.4% on the Rank1 accuracy and 0.7% on the mAP in a single query mode. The metric is only a little higher than HPM[19] and the mAP and Rank1 accuracy of the HPM we implemented are just 81.6% and 93.5% respectively. In addition, it should be noted that we do not adopt any post-processing operation such as a re-ranking approach[36], which will further enhance performance for person Re-ID especially with respect to mAP. With the multiple query setting on this dataset, we also obtain similar performance improvements, achieving 1.4% and 2.8% improvements for Rank1 and mAP respectively. The HPM has the closest performance to the proposed MMHPN, which also utilizes part-based feature learning for person Re-ID. Nevertheless, there are two main disadvantages of HPM: 1) it only makes use of various scale pyramid pooling on the last feature map, which merely leverages high-level semantic features to form the final descriptor; 2) it gives the same importance to each partition in the whole human body, which ignores the effectiveness of complex images when updating the gradient. In contrast, the proposed MMHPN approach conducts horizontal pyramid pooling on different feature maps to fuse mid- and high-level features for further enhancement of discriminative information. In addition, we focus more on difficult partitions and reduce the contribution of simple ones by using part sensitive loss.

Table 1 Comparison of state-of-the-art results on Market1501 with Single Query setting and Multiple Query setting.

Methods	Single Query		Multiple Query	
	Rank1	mAP	Rank1	mAP
Spindle[9]	76.9	-	-	-
MSCAN[4]	80.3	57.5	86.8	66.7
DLPA[27]	81.0	63.4	-	-
SVDNet[25]	82.3	62.1	-	-
TripletLoss[16]	84.9	69.1	90.5	76.4
Part loss[12]	88.2	69.3	-	-
Multi-Scale[2]	88.9	73.1	92.3	80.7

DIM[20]	89.9	75.6	93.3	82.4
HA-CNN[3]	91.2	75.7	93.8	82.8
GP-reid[32]	92.2	81.2	94.7	87.3
PCB[24]	92.3	77.4	-	-
Deep-Person[29]	92.3	79.6	94.5	85.1
Aligned-ReID[23]	92.6	82.3	-	-
PCB+RPP[24]	93.8	81.6	-	-
HPM[19]	94.2	82.7	-	-
HPM (Ours)	93.5	81.6	95.8	87.1
MMHPN	**94.6**	**83.4**	**96.1**	**88.2**

4.3.2 Evaluation using DukeMTMC-ReID

Comparison between MMHPN and state-of-the-art approaches using the DukeMTMC-ReID dataset is given in Table 2. This is a challenging dataset due to the pedestrian images being captured from eight different cameras and the bounding box size varies dramatically across different camera views. Nevertheless, the proposed MMHPN achieves 87.8% for Rank-1 accuracy and 75.1% for mAP, outperforming all the state-of-the-art methods and achieving an even better improvement of 1.5% Rank1 accuracy as well as 0.9% mAP compared with other approaches. Beyond HPM, the best model PCB[19] conducts a powerful post-processing method named Refined Part Pooling (RPP) to re-assign outliers from the human body partition. Combined with any post-processing operation, we believe that the performance is expected to be further enhanced.

Table 2 Comparison of state-of-the-art results on DukeMTMC-reID

Methods	Rank1	mAP
SVDNet[25]	76.7	56.8
Multi-Scale[2]	79.2	60.6
DIM[20]	80.4	63.9
HA-CNN[3]	80.5	63.8
Deep-person[29]	80.9	64.8
PCB[24]	81.8	66.1
PCB+RPP[24]	83.3	69.2
GP-reid[32]	85.2	72.8
HPM[19]	86.6	74.3
HPM (Ours)	85.6	71.7
MMHPN(Ours)	**87.8**	**75.1**

4.3.3 Evaluation using CUHK03

Table 3 shows evaluation results using the CUHK03 dataset where pedestrian bounding boxes automatically detected by DPM are used both in the training and testing phase. Under this setting, the proposed MMHPN achieves the state-of-the art result of 68.2% for Rank-1 accuracy and 65.4% for mAP, which outperforms all other published methods by a large margin. We attribute this

significant result to the effectiveness of part sensitive loss, which encourages the network to mine hard examples thereby reducing the risk that the deep network is prone to overfit with simple partition examples, especially on this small scale dataset.

Table 3 Comparison of state-of-the-art results on CUHK03 detected set

Methods	Rank1	mAP
LOMO+XQDA[7]	12.8	11.5
SVDNet[25]	41.5	37.3
DIM[20]	47.1	43.5
HA-CNN[3]	41.7	38.6
PCB[24]	61.3	54.2
PCB+RPP[24]	63.7	57.5
HPM[19]	63.1	57.5
HPM (Ours)	64.8	61.6
MMHPN	**68.2**	**65.4**

4.3.4 Qualitative Results

Figure 4 presents some queries and the corresponding heatmap as well as the top-10 ranking results. From the first two results, we can observe that MMHPN can represent robust and discriminative information of pedestrian identities regardless of the varying pose, gait and illumination. Note that the third pedestrian image is captured under low-resolution such that a certain amount of important information is lost. However, by some detailed cues such as an orange handbag and cyan T-shirt, the majority of person Re-ID ranking results are accurate and high quality. In the last query the person is pushing a bicycle and their body is partly occluded by the bicycle However, we can obtain all the corresponding captured images except for the one in which the man rides the bicycle. It can be seen that the proposed MMHPN model has an incredible capability in guaranteeing accurate person Re-ID results.

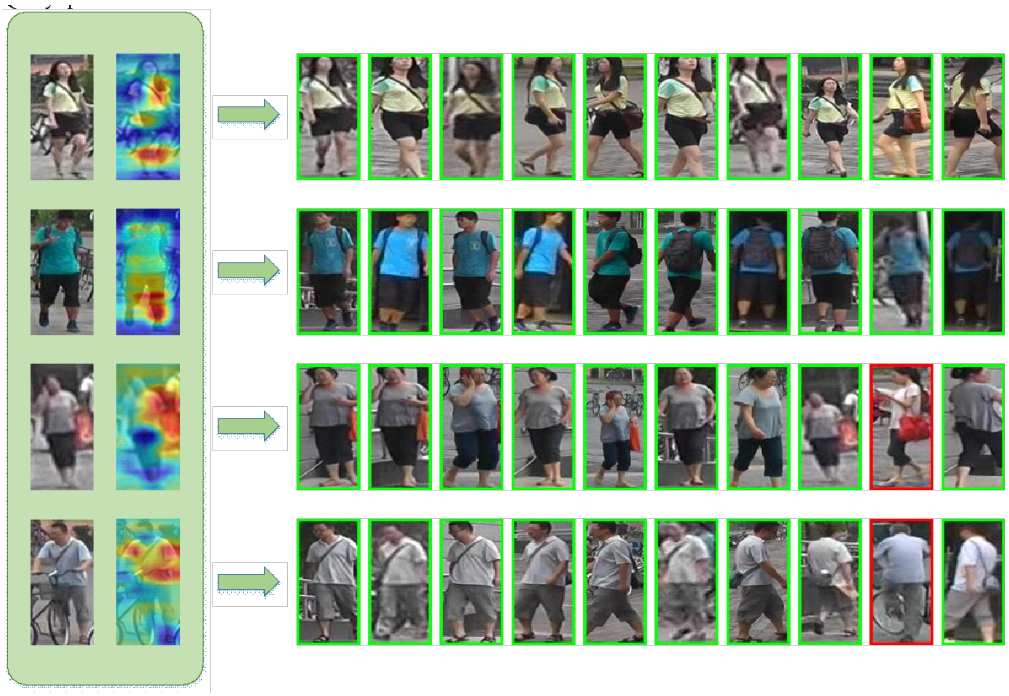

Fig. 4 Top-10 ranking list and discriminative heatmap for some given query images on Market1501 dataset by MMHPN. The images within green rectangles have the same identity as the query image, and thise with red rectangles do not.

4.4 Ablation Study

To verify the validity of each individual component in the proposed MMHPN, we conduct several ablation studies using the Market-1501, DukeMTMC-reID and CUHK03 datasets.

4.4.1 Effectiveness of Multi-level and Multi-scale Horizontal Pooling Network

In order to evaluate the effectiveness of the Multi-level Semantic Module (MSM) and the Multi-scale Horizontal Pooling Module (MHPM) within MMHPN, we remove two mid-level branches and just preserve the high-level branch with single scale as our baseline. Here, the MSM separated from MMHPN represents the three branches immediately after res_conv5_1, res_conv5_2 and res_conv5_3 and only contains one complete feature map. Additionally, MHPM separated from MMHPN means that multi-scale pooling is operated at the same convolutional layer. As depicted

in Table 4, we can observe that the baseline model with added MSM results in an obvious performance improvement over the three datasets. The Rank1 accuracy and mAP using Market-1501 are significantly improved from 85.5% and 68.2% to 90.1% and 77.3% respectively, which indicates that combining mid-level and high-level features can further enhance discriminative information to distinguish a person's identity. In addition, when MHPM is added to the baseline model, we obtain an even better performance improvement. Finally, we integrate MSM and MHPM to form MMHPN as introduced in Sec 3.2, which also achieves promising Rank1 accuracy and mAP.

Table 4 Effectiveness of Multi-level and Multi-scale Horizontal Pooling Network on three datasets.

Model	Market1501		DukeMTMC-reID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
Baseline	86.1	69.9	78.1	60.3	42.7	41.6
Baseline+MSM	90.1	77.3	82.5	65.8	55.3	53.8
Baseline+MHPM	93.5	81.6	85.6	71.7	66.8	63.6
Baseline+MMHPN	**93.6**	**82.3**	**85.7**	**72.5**	**67.1**	**64.1**

4.4.2 Effectiveness of the number of body scales

Figure 5 shows the performance of the Multi-scale Horizontal Pooling Module (MHPM) with different body scales, e.g. 1, 2, 4, 8, 12. As body scale increases, the mAP and Rank1 accuracy generally improve. Compared with one complete feature map, dividing the feature map into 8 parts achieves an improvement of 9.6% and 6.9% on mAP and Rank1 accuracy using the Market1501 dataset. When the feature map is divided into 12 parts, it does not bring obvious improvement, but has additional costs. Hence, we adopt four body scales in the MHPM. In addition, body scale determines the granularity of the part feature. We can also observe that the performance of person Re-ID can be further improved when multi-scale pooling is employed, as illustrated in Table 5.

Table 5 Effectiveness of number of body parts.

Number of body parts	Market1501		DukeMTMC-reID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
1	86.1	69.9	78.1	60.3	42.7	41.6
1+2	91.8	78.4	84.6	69.5	59.6	56.5
1+2+4	93.1	81.4	85.2	71.9	65.3	62.1
1+2+4+8	**93.6**	**82.3**	**85.7**	**72.5**	**67.1**	**64.1**

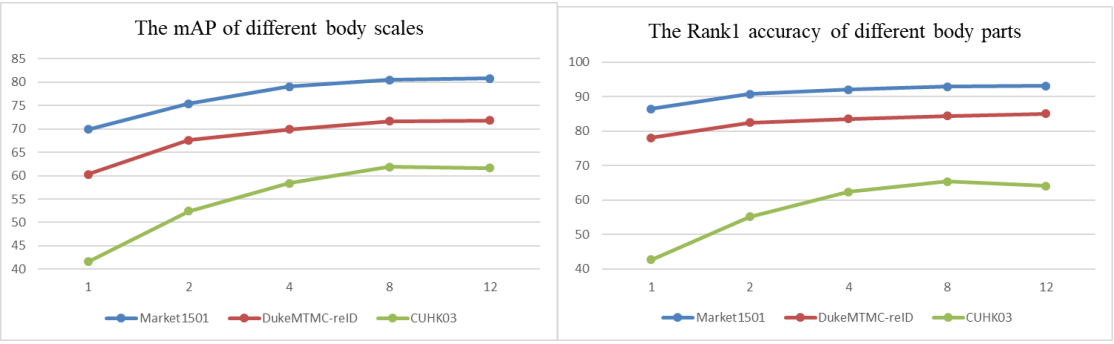

Fig. 5 The mAP and Rank1 accuracy of different body scales.

4.4.3 Effectiveness of Pooling Strategies

We evaluate the effect of different pooling strategies for MMHPN. As shown in Table 6, it can be seen that the performance of the GMP operation is better than the GAP operation in most situations. The reason is that the GAP exploits the full extent of a particular part and gives the same importance to all locations for final partial representation. Hence, when one discriminative partition of a pedestrian is surrounded by unrelated background patterns, it will have a low response and the discriminative information may be missed. In contrast, the GMP only focusses on the location with the largest response. These two pooling strategies are complementary in generating the final feature representation. Thus, it is essential to use both the GAP and GMP to maintain discriminative and robust information. Therefore we integrate GAP and GMP into a weighted sum to take advantage of the two strategies adaptively. From Table 5, we can observe that the Adaptive Pooling Strategy (APS) achieves better performance compared with using the GAP and GMP operations alone.

Table 6 Effectiveness of different pooling strategies on three datasets

Model	Market1501		DukeMTMC-reID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
MMHPN+GAP	93.6	82.3	85.7	72.5	67.1	64.1
MMHPN+GMP	93.9	82.8	**87.3**	74.3	67.9	**64.7**
MMHPN+APS	**94.2**	**83.1**	87.0	**74.6**	**68.3**	64.5

4.4.4 Effectiveness of Part Sensitive Loss

In order to verify the effectiveness of the Part Sensitive Loss (PSL), we also conduct comparison of MMHPN with and without applying PSL for training. As given in Table 7, it can be observed that PSL may not ensure the best performance using the Rank1 accuracy. However, we achieve consistent improvements using the three benchmark datasets in mAP, which is the most important metric to evaluate the effectiveness of person Re-ID methods. In fact, Rank 1 accuracy indicates the ability to match the easiest gallery in different cameras, whereas mAP characterizes the ability to retrieve all the galleries. Meanwhile, the PSL reduces the loss from an easy partition and encourages the network to focus on difficult ones, which enhances the robustness of the feature representation, especially when using small datasets such as CUHK03.

Table 7 Effectiveness of different pooling strategies on three datasets.

Model	Market1501		DukeMTMC-ReID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
MMHPN+softmax	94.2	83.1	87.4	74.7	**68.3**	64.5
MMHPN+PSL	**94.6**	**83.4**	**87.8**	**75.1**	68.2	**65.4**

5 Conclusion

In order to combine the appropriate global and local features to solve the key missing part cases in Re-ID, we propose a Multi-level and Multi-scale Horizontal Pooling Network (MMHPN). In addition to the traditional expansion of ResNet for person Re-ID, we extend the multi-level and

multi-scale features designed specifically for the Adaptive Pooling Strategy (APS). Local and global cues are stacked together in networks for joint optimization to make final predictions more reliable. We also build our Part Sensitive Loss (PSL) upon softmax loss to reduce the effect of easy partition and focus more on difficult ones, which decreases the risk of over-fitting to easy samples and facilitates training the person re-identification networks. Extensive experiments on three popular and challenging benchmark datasets thoroughly demonstrate the superiority of the proposed method over the state-of-the-art methods.

Acknowledgments

This work is supported by the Distinguished Creative Talent Program of Shenyang (RC170490), Fundamental Research Funds for the Central Universities (N172608005, N182608004), Natural Science Foundation of Liaoning (No.20180520040) and National Natural Science Foundation of China (No. 61471110, 61733003).

References

1. Y. Chen et al., "Person Re-Identification by Deep Learning Multi-Scale Representations," in 2017 *IEEE Int. Conf. on Computer Vision Workshop (ICCV Workshops)*, pp. 2590-2600, IEEE (2017).
2. C. Su et al., "Pose driven deep convolutional model for person re-identification," in *2017 IEEE Int. Conf. on Computer Vision (ICCV),* pp. 3980–3989, IEEE (2017).

3. W. Li et al., “Harmonious attention network for person re-identification,” in *2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2018).
4. D. Li et al., “Learning deep context-aware features over body and latent parts for person re-identification,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (pp. 384-393)
5. B. Ma et al., “Local descriptors encoded by fisher vectors for person re-identification,” in *European Conf. on Computer Vision,* Berlin, pp. 413-422, Springer (2012).
6. L. Bazzani et al., “Symmetry-driven accumulation of local features for human characterization and re-identification,” *Computer Vision and Image Understanding*, **117**(2), pp. 130–144 (2013).
7. S. Liao et al., “Person re-identification by local maximal occurrence representation and metric learning,” in *2015 IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*, pp. 2197-2206, IEEE (2015)

8. E. Insafutdinov et al., “A deeper, stronger, and faster multi person pose estimation model,” in *European Conf. on Computer Vision,* Cham, pp. 34-50, Springer (2016)
9. H. Zhao et al., “Spindle net: Person re-identification with human body region guided feature decomposition and fusion,” in 2017 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* pp. 1077-1085, IEEE (2017).
10. H. Liu et al., “End-to-end comparative attention networks for person re-identification,” *IEEE Trans. on Image Processing* **26**(7), pp. 3492-3506 (2017).
11. X. Liu et al., “Attentive deep features for pedestrian analysis,” in 2017 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp 350-359, IEEE (2017).
12. H. Yao et al., “Deep Representation Learning with Part Loss for Person Re-Identification,” *arXiv preprint arXiv:1707.00798* (2017).
13. E. Ahmed et al., “An improved deep learning architecture for person re-identification,” in 2015 *IEEE Conf. on*

Computer Vision and Pattern Recognition (CVPR), pp. 3908-3916, IEEE (2015)

14. R. R. Varior et al., "A siamese long short-term memory architecture for human re-identification," in *European Conf. on Computer Vision*, Cham, pp. 135-153, Springer (2016).
15. R. R. Varior et al., "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*, Cham, pp. 791-808, Springer (2016).
16. A. Hermans et al., "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737* (2017).
17. L. Zheng et al., "Scalable person re-identification: A benchmark," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1116–1124, IEEE (2015).
18. E. Ristani et al., "Performance measures and a dataset for multi-target, multi-camera tracking," in *European Conf. on Computer Vision*, Cham, pp 17–35, Springer (2016).

19. Y. Fu et al., “Horizontal Pyramid Matching for Person Re-identification,” *arXiv preprint arXiv:1804.05275* (2018).
20. Q. Yu et al., “The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching,”, *arXiv preprint arXiv:1711.08106* (2017).
21. Wei. Li et al., "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*, pp. 152-159, IEEE (2014).
22. D. Yi et al., “Deep metric learning for person re-identification,” in *2014 22nd International Conference on Pattern Recognition (ICPR)*, pp. 34-39, IEEE (2014).
23. X. Zhang et al., “Alignedreid: Surpassing human-level performance in person re-identification,” *arXiv preprint arXiv:1711.08184* (2017).
24. Y. Sun et al., “Beyond Part Models: Person Retrieval with Refined Part Pooling,” *arXiv preprint arXiv:1711.09349* (2017).

25. R. Girshick, “Fast r-cnn,” in 2015 *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1440-1448, IEEE (2015).
26. M. Jaderberg et al., “Spatial transformer networks,” in *Advances in neural information processing systems*, pp 2017–2025 (2015).
27. L. Zhao et al., “Deeply-learned part aligned representations for person re-identification,” in 2017 IEEE *Int. Conf. Computer Vision (ICCV)*, pp. 3219–3228, IEEE (2017).
28. R. Hadsell et al., “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1735-1742, IEEE (2006).
29. X. Bai et al., “Deep person: Learning discriminative deep features for person re-identification,” *arXiv preprint arXiv:1711.10658*, (2017).
30. D. cheng et al., “Person re-identification by the asymmetric triplet and identification loss function,” *Multimedia Tools and Applications*, **77**(3), pp. 3533-3550 (2018).

31. K. He et al., “Deep residual learning for image recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE (2016).
32. J. Almazan et al., “Re-id done right: towards good practices for person re-identification,” *arXiv preprint arXiv:1801.05339* (2018).
33. T. Y. Lin et al., “Focal loss for dense object detection,” in *2017 IEEE Int. Conf. on Computer Vision (ICCV*), pp. 2999-3007, IEEE (2017).
34. Z. Zheng et al., “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3774-3782, IEEE (2017).
35. P. Felzenszwalb et al., “A discriminatively trained, multiscale, deformable part model,” in *2008 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, IEEE (2008).
36. Z. Zhong et al., “Re-ranking person re-identification with k-reciprocal encoding,” *in 2017 IEEE Conf. on Computer Vision*

and Pattern Recognition (CVPR), pp. 3652–3661, IEEE (2017).

Yunzhou Zhang is an assistant professor at the Northeastern University of China. He received his BS and MS degrees in Mechatronics Engineering from National University of Defense Technology of China in 1997 and 2000, respectively, and his PhD degree in Mode Identification and Artificial Intelligence from Northeastern University of China in 2009. He is the author of more than 40 journal papers and has written three book chapters. His current research interests include robotics and computer vision. He is an associate editor of International Journal of Advanced Robotic Systems.

Biographies and photographs for the other authors are not available.

Caption List

Fig. 1 Illustration of proposed Multi-level and Multi-scale Horizontal Pooling Network..

Fig. 2 Overview of Multi-level and Multi-scale Horizonal Pooling Network.

Fig. 3 Attention maps in different scales produced by the last convolutional output of different models.

Fig. 4 Top-10 ranking list and discriminative heatmap for some given query images on Market1501 dataset by MMPPN.

Table 1 Comparison of state-of-the-art results on Market1501 with Single Query setting and Multiple Query setting.

Table 2 Comparison of state-of-the-art results on DukeMTMC-reID.

Table 3 Comparison of state-of-the-art results on CUHK03 detected set.