Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification

Cheng Wang[1], Qian Zhang[2], Chang Huang[2], Wenyu Liu[1],
and Xinggang Wang[1](✉)

[1] School of EIC, Huazhong University of Science and Technology, Wuhan, China
{wangcheng,liuwy,xgwang}@hust.edu.cn
[2] Horizon Robotics Inc., Beijing, China
{qian01.zhang,chang.huang}@hobot.cc

Abstract. We propose a novel deep network called Mancs that solves the person re-identification problem from the following aspects: fully utilizing the attention mechanism for the person misalignment problem and properly sampling for the ranking loss to obtain more stable person representation. Technically, we contribute a novel fully attentional block which is deeply supervised and can be plugged into any CNN, and a novel curriculum sampling method which is effective for training ranking losses. The learning tasks are integrated into a unified framework and jointly optimized. Experiments have been carried out on Market1501, CUHK03 and DukeMTMC. All the results show that Mancs can significantly outperform the previous state-of-the-arts. In addition, the effectiveness of the newly proposed ideas has been confirmed by extensive ablation studies.

Keywords: Person re-ID · Attention · Curriculum sampling
Multi-task learning

1 Introduction

Person re-identification (re-ID), aims at spotting a person of interest in a camera network, is a well-established research problem in computer vision [39]. Due to its great impact in the application of video surveillance [16], and the public available large-scale re-ID datasets and the encouraging re-ID results of deep learning systems, person re-ID has become increasingly popular in computer vision.

However, the person re-ID problem is quite challenging in the situations of large viewpoint variation, large misalignment, and occlusion, etc. Thus, lots of works have been proposed to learn an effective person representation based on the training images with person identities are given. The learning problem is naturally formulated as a distance metric learning problem [6,40]. It aims to find a

C. Wang—The work was done when Cheng Wang was an intern in Horizon Robotics.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11208, pp. 384–400, 2018.
https://doi.org/10.1007/978-3-030-01225-0_23

new distance metric to transform the original person feature (such as HOG [9] and SIFT [24]) into a new space in which the examples have the same identity are closer and otherwise have large distances. In deep learning person re-ID systems, the idea of distance metric learning is usually formulated as a ranking loss and has been proven to be effective. A typical ranking loss is triplet loss, such as [27]. Given an anchor example with a positive example that has the same identity as the anchor and a negative example that has a different identity, triplet loss enforces the anchor-positive distance is smaller than the anchor-negative distance. Besides the triplet loss, there are other types of metric learning losses have been proposed, such as the histogram loss [30] and the quadruplet loss [6]. Due to the unbalanced number of positive and negative sample pairs, when training with the metric learning losses, the strategy of example sampling is an essential issue. Recent studies show that mining hard negative is beneficial to learn a robust deep person representation [11,27]. In addition, another loss function, the classification loss function which directly classifies the person images into its own identity class, is still very useful [19]. The deep re-ID networks can provide great global deep person representation. However, aligning and matching discriminative local features for person re-ID is still very necessary due to the inaccurate person detection, person pose variation etc. To achieve this goal, there are different ways, such as in-explicitly feature aligning and matching using spatial attention [37] and explicitly feature aligning using LSTM [4] or aligning by finding the shortest path [35].

By reviewing the current person re-ID research works, we can find that, due to challenges in the problem, there exists at least the following issues need to be handled: (1) the choices of loss functions; (2) the misalignment problems; (3) finding discriminative local features; and (4) how to sample training examples during the optimization of the ranking loss functions. In current person re-ID research works, few of them have addressed all these issues in the same framework. Therefore, in this paper, we propose Mancs, a unified person re-ID deep network, to deal with the mentioned issues at the same time.

Mancs has the following building blocks. It has a backbone network, such as ResNet-50, to extract deep feature hierarchies for the input person image. The backbone network is supervised with a ranking loss and a classification loss. The ranking loss is a triplet loss; we propose a novel curriculum sampling strategy to train with the triplet loss; the curriculum sampling method is motivated by curriculum learning [5] that helps to train the network by sampling examples from easy to hard. The classification loss is a focal loss which has been proven to be helpful for dense object detection [21]. To deal with the misalignment problem and localize discriminative local features, we propose a new fully attentional block (FAB) which creates both channel-wise and spatial-wise attention information to mine the useful features for re-ID. To better learn the FABs in our network, we further propose to use the deep supervision idea [14] by adding a classification loss function for each FAB; thus, the classification loss function is termed as attention loss. In the end, the triplet loss, the focal loss, and the attention loss are combined for training our person re-ID network in a multi-task manner.

In the experiments, we have studied the Mancs on three large-scale person re-ID datasets, which are Market-1501 [38], CUHK03 [17], and DukeMTMC-reID [42]. The results clearly demonstrate the contribution of newly proposed triplet loss with curriculum sampling, the deeply supervised fully attentional block, the focal loss, and the unified multi-task learning framework. Besides, Mancs obtains significant better accuracies than the previous state-of-the-arts on all the datasets.

2 Related Work

Attentional Network. Recently, lots of works have adopted attentional deep learning approaches to tackle the misalignment problem in person re-ID. Usually, they use an additional subnet to obtain a region of interest and extract features from those attention areas. MSCAN [15] uses a spatial transformation network (STN) [13] to get several attention regions and then extracts local feature from the regions. HA-CNN [19] combines both soft attention methods and hard attention methods. Apart from acquiring hard attention region, they also rely on channel-wise attention and spatial-wise attention, which are complementary to previous hard attention. CAN [23] combines attention methods with LSTM to obtain discriminative attention feature of the whole image. The proposed Mancs adopts a 1×1 convolution to acquire an attention mask with the same shape of the feature map.

Metric Learning. It is widely used for learning image embeddings, such as [3,4,6,27,35,40]. In face recognition, [27] uses triplet loss to push the negative pair further and pull the positive pair closer. Except triplet loss, contrastive loss [40] and quadruplet loss [6] are also used in person re-ID task. For triplet loss, online hard examples mining (OHEM) is important, namely selecting the furthest positive examples and closet negative examples for training. In the proposed Mancs framework, we sample training examples in a curriculum way.

Multi-task Learning. Since metric learning and representation learning can both be applied to person re-identification task, [4,10] combine softmax loss with triplet loss to train model for a robust performance. [1] adopts two losses but divides them into two stages. The proposed Mancs combines focal loss with triplet loss and can be trained in an end-to-end way.

3 Method

In this section, we present the proposed Mancs person re-ID framework by first describing the training framework and its building blocks, then describing the multi-task learning strategy and finally describing the inference network.

3.1 Training Architecture

The network architecture for training is shown in Fig. 1. Basically, it has three major components. The backbone network, the attention module, and the loss functions, which are described as follows.

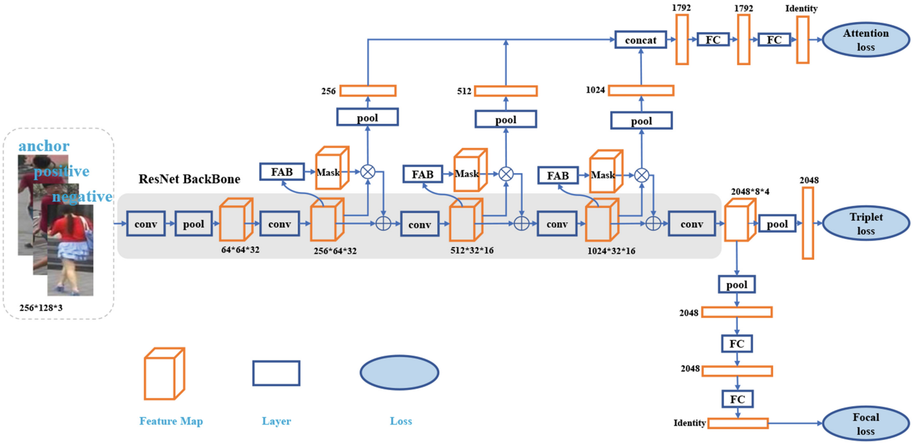


Fig. 1. The Mancs training architecture: its backbone network is ResNet-50; the pooling layers are all spatially average pooling; the FAB block is an attention module which is described in Fig. 2; and it has three loss functions: attention loss, triplet loss and focal loss

The backbone network is served as a multi-scale feature extractor. Without loss of generality, here we apply the popular ResNet-50. As shown in Fig. 1, we take the conv-2, conv-3 and conv-4 feature maps to generate attention masks which are added back into the mainstream. The last conv-5 feature map is used for generating the final person identity feature.

3.2 Fully Attentional Block

Attention is very useful in person re-ID, which has been proved in the previous studies [15,15,19]. In our understanding, attention can localize the most discriminative local regions for person re-ID. To fully illustrate the usage of attention, we propose a fully attentional block (FAB). FAB is motivated by the recent Squeeze-and-Excitation Network (SENet) [12] method, which illustrates that different channels of a feature map play different roles in specifying objects. In consideration of that, the SE block (Fig. 2(a)) in SENet utilizes the preference of channels and gives a weighting coefficient to each channel of the feature map. However, the initial SE block only re-calibrates feature response on channel-wise while ignores the spatial-wise response on the account of using global pooling which leads to losing the spatial structure information. To remedy this problem,

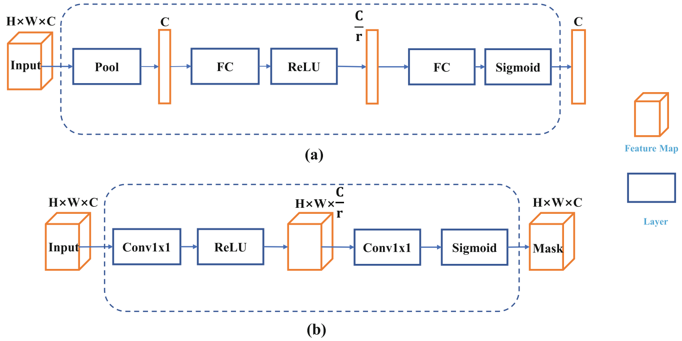


Fig. 2. (a) is a SE Block and the reduction factor r is set to 16; (b) is our Fully Attentional Block where $r = 16$

the proposed FAB discards the pooling layer and employs 1×1 convolutional layers instead of fully-connected layers to regain spatial information. Therefore we can get an attention mask with the same size of input feature map and this attention model is called fully attentional block. FAB is illustrated in Fig. 2(b) and formulated as follows.

Given a convolutional feature map F_i, its attention map is computed as:

$$M = \text{Sigmoid}\ (\text{Conv}(\text{ReLU}(\text{Conv}(F_i)))), \tag{1}$$

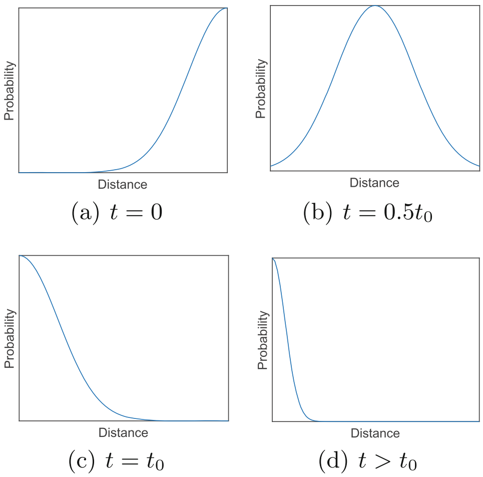
where the two Conv operators are 1×1 convolution. The inner Conv is used for squeeze and the outer Conv is used for excitation. After obtaining the attention map M the output feature map of F_i is calculated as:

$$F_o = F_i * M + F_i, \tag{2}$$

where the operator $*$ and $+$ are performed in an element-wise manner. This means that the attention induced feature map is added into the original feature map to emphasize discriminative features. It is worth to note that the proposed FAB is pluggable and can be applied to any existing CNN, since FAB does not change the size of the convolutional feature map.

3.3 ReID Task #1: Triplet Loss with Curriculum Sampling

A ranking loss is essential for a person re-ID deep network since it has better generalization ability than the contractive/classification loss especially when the training dataset is not large enough. Thus, we firstly introduce a ranking branch with triplet loss to our model. To clearly describe the proposed triplet loss method, we denote the feature of image I_i for the triplet loss as $f_{\text{rank}}(I_i)$, where $f_{\text{rank}}(\cdot)$ means the feature extraction network for ranking features. As shown in Fig. 1, $f_{\text{rank}}(\cdot)$ shares the backbone network with other branches and has a pooling layer and a FC layer owned by itself. When applying a triplet loss, its sampling algorithm matters (Fig. 3).



(a) $t = 0$

(b) $t = 0.5t_0$

(c) $t = t_0$

(d) $t > t_0$

Fig. 3. Different selecting probabilities on negative examples under a given t. X-axis represents the distance between negative examples and anchor, while Y-axis represents the probability of a negative example being selected

Most person re-ID works [4,11,35] adopt the triplet loss proposed by [27]. The main idea of [27] is to do online hard triplets sampling through the so-called PK sampling method, which randomly samples P identities and then randomly K images for each identity to form a mini-batch with the size of $P \times K$. In a mini-batch $\mathcal{I} = \{I_i\}_{i=1}^{PK}$, for each image, it is considered as an anchor image denoted as I_i^a, and the hardest positive image and the hardest negative of the anchor are found in $\mathcal{I}$ which are denoted as I_i^p and I_i^n respectively. Thus, $T_i = \{I_i^a, I_i^p, I_i^n\}$ is a triplet and PK triplets can be obtained. The above sampling procedure is also called online hard examples mining (OHEM). It is widely used in many visual application problems. However, it is easy to collapse according to [25]. Inspired by curriculum learning [5], we propose a new sampling way named curriculum sampling. The idea is to train a person re-ID network in a progress from easy triplets to hard triplets.

More specifically, we discard the method of sampling hardest instances in the beginning of training and start from easy instances. Given an anchor instance I_i^a, firstly, we randomly select one of its positive instances as I_i^p; secondly, we sort negative instances according to their distances to the anchor from small to large, which means that the negatives are sorted from hard to easy; thirdly, we give each negative instance a probability of being selected. These probabilities obey a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ and σ are defined as below:

$$\mu = [N_n - \frac{N_n}{t_0} t]_+, \tag{3}$$

$$\sigma = a \times b^{[\frac{t-t0}{t_1-t_0}]_+}, \tag{4}$$

where $[\cdot]_+ = max(\cdot, 0)$, N_n is the numbers of negative instances. a is initial std value and b is the decay exponent when $t > t_0$. t_0 and t_1 are hyper-parameters that control the speed of learning process from easy to hard. The above procedure selects an anchor, a positive instance and a negative instance to form a triplet. Next, the aim is still the same; we randomly select another different positive instance as the second procedure does. Then, we select another negative example based on the previous probability distribution (since the anchor is still the same). Now, we have selected our second triplet. When all positive instances of this anchor are selected, we move to the next anchor. The process described above totally gives us $PK(K-1)$ triplets. PK is the number of anchors. $K-1$ is the number of positive instances of each anchor.

Based on the curriculum sampling method, the final loss for ranking branch can be defined as:

$$L_{\text{rank}} = \frac{1}{P(K-1)K} \sum_{i=1}^{P(K-1)K} [m + D(f_{\text{rank}}(I_i^a), f_{\text{rank}}(I_i^p)) - D(f_{\text{rank}}(I_i^a), f_{\text{rank}}(I_i^n))]_+, \tag{5}$$

where $D(\cdot,\cdot)$ is the Euclidean distance between two feature vectors. The probability of I_i^n being chose is defined as below:

$$Pr(I_i^{n*} = I_i^n \mid I_i^a) \propto \mathcal{N}(\mu, \sigma) \tag{6}$$

3.4 ReID Task #2: Person Classification with Focal Loss

Recent studies show that combining both ranking loss and classification loss is helpful for person re-ID [4]. In Mancs, we also have a classification branch. Since hard examples mining is essential in the ranking loss, we think it can also be applied in the classification task. Now that hard examples are more important than easy examples in learning, we decide to increase the ratio that negative examples take up in the total loss. Apparently, the newly proposed focal loss [21] for dense object detection is an appropriate option, since it is able to let hard examples have a higher weight than easy examples.

We denote the feature extractor of the classification branch as $f_{\text{cls}(\cdot)}$. Given an image I_i and its ground-truth identity c_i, the probability of I_i belonging to the c_i-th class is denoted as follows:

$$p_i = \text{Sigmoid}_{c_i}\left(\text{FC}\left(f_{\text{cls}}(I_i)\right)\right), \tag{7}$$

where the subscript c_i of Sigmoid means taking the output value in its c_i-th dimension. Then, the focal loss for classification can be defined as follows:

$$L_{cls} = -\frac{1}{PK}\sum_{i=1}^{PK}(1-p_i)^{\gamma}\log(p_i). \tag{8}$$

3.5 ReID Task #3: Deep Supervision for Better Attention

As shown in Fig. 1, we can acquire different scales of attention responses based on different level of intermediate features. Besides, in order to acquire accurate attention maps, we use person identity information to deeply supervised them. The idea is inspired by the deeply-supervised nets work [14]. The deep supervision is helpful in alleviating the problem of gradient vanishing.

To implement this goal, the multi-scale attention maps are spatially and averagely pooling into a one-dimensional feature vector; then, the feature vectors are concatenated into an attention feature vector. We denote the attention feature extractor as $f_{\text{att}}(\cdot)$. Similar to the setting in the Sect. 3.4, the probability of I_i belonging to the c-th class is given as:

$$q_i^c = \text{Sigmoid}_c\left(\text{FC}\left(f_{\text{att}}(I_i)\right)\right). \tag{9}$$

Then, we define the loss function of the attention branch as:

$$L_{\text{att}} = -\frac{1}{PKC}\sum_{i=1}^{PK}\sum_{c=1}^{C} y_i^c \log(q_i^c) + (1-y_i^c)\log(1-q_i^c) \tag{10}$$

where $y_i^c = 1$ if I_i belongs to the c-class and otherwise $y_i^c = 0$.

3.6 Multi-task Learning

As shown in Fig. 1, the three tasks share the same backbone network. In training, the corresponding three loss functions are optimized jointly. The final loss is given by:

$$\mathcal{L} = \lambda_{\text{rank}} L_{\text{rank}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{att}} L_{\text{att}}, \tag{11}$$

where λ_{rank}, λ_{cls}, and λ_{att} weight factors for the loss functions.

3.7 Inference

In testing the inference network is quite simple which is shown in Fig. 4. We choose the deep feature for ranking loss, i.e., f_{rank}, as the final re-ID feature for each instance. This is mainly because the proposed triplet loss with curriculum sampling can produce deep feature with better generalization ability. The choice of using ranking features has been confirmed in many other research works, such as [4,32].

4 Experiments

4.1 Datasets

We mainly focus on three large-scale reID datasets, which are Market1501, CUHK03 and DukeMTMC-reID. The details of the three datasets are given as follows.

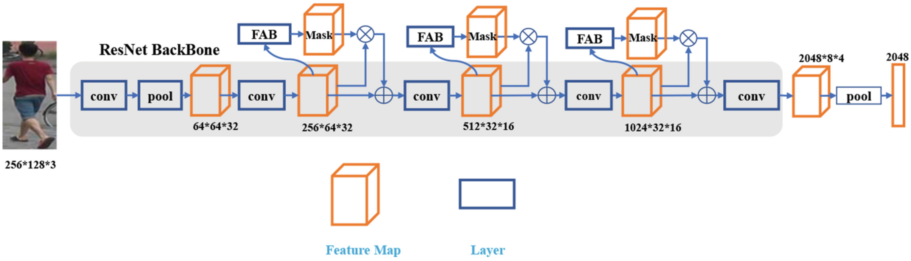


Fig. 4. The architecture of the inference network in Mancs.

Market-1501 [38]**:** It contains 32,668 images of 1,501 identities captured by six camera views. The whole dataset is divided into a training set containing 12,936 images of 751 identities and a testing set containing 19,732 images of 750 identities. For each identity in testing set, we select one image from each camera as a query image, forming 3,368 queries following the standard setting in [38].

CUHK03 [17]**:** It contains 14,097 images of 1,467 identities. It provides person bounding boxes detected both from the deformable part model detector and from manual labeling. We conduct experiments both on the labeled dataset and detected dataset. The dataset offers a 20-splits dividing, resulting in a training set with 1,367 identities and a testing set with 100 identities. The average performance of 20 splits is adopted as the final result of this dataset. Similar with [44], we also evaluate a division way with the training set of 767 identities and the testing set of 700 identities.

DukeMTMC-reID [42]**:** Similar to Market-1501, DukeMTMC-reID contains 36,411 images of 1,812 identities taken by 8 cameras, where only 1,404 identities appeared in more than 2 cameras. The other 408 identities are regarded as distractors. The training set contains 16,522 images of 702 identities while the testing set contains 2,228 query images of 702 identities and 17,661 gallery images.

4.2 Evaluation Protocol

We follow the official training and evaluation protocols in Market-1501, CUHK03 and DukeMTMC-reID. We use the cumulative matching characteristics (CMC) and mean Average Precision (mAP) metrics. We conduct experiments on Market-1501 under both single query and multi-query mode. While on CUHK03 and DukeMTMC-reID, we conduct experiments only in single query mode. Especially in CUHK03, there are 2 different ways of dividing the training set and testing set. One is dividing to 1,367/100 split, the other is dividing to 767/700 split. The former needs to run 20 rounds and get an averaged result which we use rank1, rank5 and rank10 matching rate to evaluate. The later is similar to

Market-1501 and DukeMTMC-reID and only need to run once which is evaluated by rank1 matching rate and mAP. We perform experiments for both splits.

4.3 Implementation Details

We implement Mancs based on Pytorch [26]. We take the ResNet-50 model pretrained on ImageNet as the backbone. As described above, we insert a fully-connected layer with channel numbers of 2048 before the last classification layer.

Data Augmentation. We first resize training images to 256 × 128. Then we randomly crop each image with scale in the interval $[0.64, 1.0]$ and aspect ratio in $[2, 3]$. Third, we resize these cropped images back to the size of 256 × 128 and randomly horizontally flip them with the probability of 0.5. Finally, we add a random erasing data augmentation method as described in [45]. Before sent to the network, each image is subtracted the mean value and divided by the standard deviation according to standard normalization procedure when using the pretrained model on ImageNet.

Training Configurations. As described in Sect. 3.3, we adopt *PK Sampling* strategy to form every mini-batch. The values of both P and K is set distinguished among different datasets. For Market1501, P and K is set to 16 and 16, respectively. For CUHK03, P is set to 32 and K is set to 8. DukeMTMC-ReID shares the same configurations with Market1501. Each epochs includes $[N_c/P]$ mini-batches. We train our models for 160 epochs. t_0, t_1, a and b described in Eqs. (3) and (4) are set to $30, 60, 15$ and 0.001, respectively. $\lambda_{rank}, \lambda_{cls}$ and λ_{att} are set to $1, 1$ and 0.2, respectively. The margin m in Eq. (5) is set to 0.5. γ in Eq. (11) is set to 2. We adopt Adam optimizer with an initial learning rate of 3×10^{-4} in our experiments to minimize the three losses. In addition, we add gradient clipping to prevent model collision. The activation function of the last convolutional layer is changed from ReLU to PReLU, which can enrich the expressiveness of the final feature. All the experiments run on a server with 4 TITAN XP GPUs.

4.4 Comparisons with the State-of-art Methods

Evaluation On Market-1501. We evaluated our proposed Mancs against 13 existing methods on Market-1501. As showed in Table 1, our model outperforms HA-CNN which also uses an attention subnetwork by 6.6% on mAP and 1.9% on rank1 matching rate under single query mode, respectively. Compared with Deep-Person which also adopts multi-task learning, our Mancs outperforms it by 2.7% in mAP and 0.8% in rank1 matching rate under the single query model, respectively. Under multiple query mode, Mancs outperforms Deep-Person by 2.4% on mAP and 0.9% on rank1 matching rate, respectively. With the combination of re-ranking approach, the performance can be further improved. Under the single query mode, mAP and rank1 can be boosted to 92.3% and 94.9%, respectively. While under multiple query mode, it can reach 94.5% and 95.8% (Tables 2 and 3).

Table 1. Comparisons on Market-1501 with state-of-art methods. SQ: single query, MQ: multiple queries. Mancs obtains the best results.

Methods	SQ		MQ	
	rank1	mAP	rank1	mAP
CAN [23]	60.3	35.9	72.1	47.9
DNS [34]	61.0	35.6	71.5	46.0
Gated S-CNN [31]	65.9	39.6	76.0	48.4
CRAFT [8]	68.7	42.3	77.0	50.3
Spindle [36]	76.9	-	-	-
MSCAN [15]	80.3	57.5	86.8	66.7
SVDNet [29]	82.3	62.1	-	-
PDC [28]	84.1	63.4	-	-
TriNet [11]	84.9	69.1	90.5	76.4
JLML [18]	85.1	65.5	89.7	74.5
HA-CNN [19]	91.2	75.7	93.8	82.8
Deep-Person [4]	92.3	79.6	94.5	85.1
AlignedReID [35]	92.6	82.3	-	-
Mancs(Ours)	**93.1**	**82.3**	**95.4**	**87.5**

Table 2. Comparisons on the CUHK03 dataset in terms of mAP and rank1 matching rate, using both manually labeled person bounding boxes and automatic detections by DPM, under the setting of 767/700 split. Mancs gets the best results.

Settings	767/700 split			
	Labeled		Detected	
Methods	rank1	mAP	rank1	mAP
BoW+XQDA [33]	7.9	7.3	6.4	6.4
LOMO+XQDA [20]	14.8	13.6	12.8	11.5
IDE(C) [44]	15.6	14.9	15.1	14.2
IDE(C)+XQDA [44]	21.9	20.0	21.1	19.0
IDE(R) [44]	22.2	21.0	21.3	19.7
IDE(R)+XQDA [44]	32.0	29.6	31.1	28.2
DPFL [7]	43.0	40.5	40.7	37.0
SVDNet-ResNet50 [29]	-	-	41.5	37.6
HA-CNN [19]	44.4	41.0	41.7	38.6
TriNet+Random Erasing [11,45]	58.1	53.8	55.5	50.7
Mancs(Ours)	**69.0**	**63.9**	**65.5**	**60.5**

Table 3. Comparisons on CUHK03 in terms of rank1, rank5, rank10 matching rate, using both manually labeled person bounding boxes and automatic detections by DPM, under the setting of 1367/100 split. Mancs obtains the best results.

Settings	1367/100 split					
Models	Labeled			Detected		
	r1	r5	r10	r1	r5	r10
DNS [34]	62.5	90.0	94.8	54.7	84.7	94.8
Gated-SCNN [31]	-	-	-	68.1	88.1	94.6
MSCAN [15]	74.2	94.3	97.5	68.0	91.0	95.4
Quadruplet [6]	75.5	95.2	99.2	-	-	-
SSM [2]	76.6	94.6	98.0	72.7	92.4	96.1
SVDNet [29]	-	-	-	81.8	95.2	97.2
CRAFT [8]	-	-	-	84.3	97.1	98.3
JLML [18]	83.2	98.0	99.4	80.6	96.9	98.7
DPFL [7]	86.7	-	-	82.0	-	-
PDC [28]	88.7	98.6	99.2	78.3	94.8	97.2
Deep-Person [4]	91.5	99.0	99.5	89.4	98.2	99.1
AlignedReID [35]	91.9	98.7	99.4	-	-	-
Mancs(Ours)	**93.8**	**99.3**	**99.8**	**92.4**	**98.8**	**99.4**

Evaluation On CUHK03. As mentioned in Sect. 4.1, there are two ways of dividing the CUHK03 dataset into training and testing set. Typically, the 767/700 split setting is harder than the 1367/100 setting. Because the former split has less training images and more testing images than the later. We evaluate Mancs in both settings. Without the help of re-ranking, in the detected split, Mancs can reach to 92.4% under the 1367/100 split and 65.5% under the 767/700 split on rank1 target, respectively. Especially under the 767/700 split, Mancs is 23.8% higher than HA-CNN and 10.0% higher than TriNet with Random Erasing.

Evaluation On DukeMTMC-reID. Similar to Market-1501, the comparisons with related methods is depicted in Table 4. Compared with the state-of-art method Deep-Person [4], Mancs achieves an improvement of 7.0% on mAP and 4.0% on rank1 performance.

From the above experiment results, we can observe that Mancs obtains excellent person re-ID performance. However, to future discovery the limitations, we visualize some randomly selected failure cases of Mancs in Fig. 5, in which the results of 4 probes in DukeMTMC-reID are listed. From the results in the second and the third row, we can observe that Mancs may be affected by some unusual situations, such as multi-person in one image and a car occupies the image, which is very unusual in the training set. So, when applying Mancs in real applications, it is better to have an accurate person detector. From the results in the first and

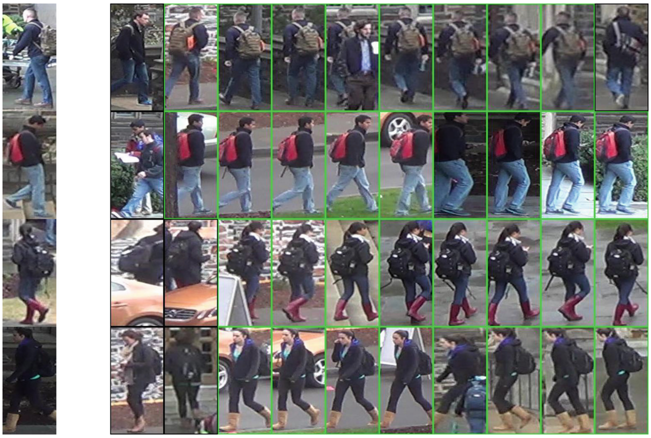


Fig. 5. Some failure cases (missed in the rank1 matching) on DukeMTMC-reID. Left are probes, right are the ranking results. Persons surrounded by green box have the same identities as their probes. (Color figure online)

the fourth row, we can observe that there are still some very similar distractors may affect Mancs, which will be deeply investigated in future research. However, these failure cases can be remedied by a re-ranking post process.

4.5 Ablation Study

We further perform several extra experiments to verify the effectiveness of each individual component of our proposed model. Market-1501 and CUHK03 are used in experiments of ablation study. Specifically, we perform all experiments under single query mode. In addition, we use the 767/700 split of CUHK03 with

Table 4. Comparisons with state-of-art results on DukeMTMC-reID.

Methods	rank1	mAP
BoW+XQDA [33]	25.1	12.2
LOMO+XQDA [20]	30.8	17.0
LSRO [43]	67.7	47.1
AttIDNet [22]	70.7	51.9
PAN [41]	71.6	51.5
SVDNet [29]	76.7	56.8
DPFL [7]	79.2	60.6
HA-CNN [19]	80.5	63.8
Deep-Person [4]	80.9	64.8
Mancs(Ours)	**84.9**	**71.8**

Table 5. Ablation studies of the modules of Mancs, based on both Market-1501 and CUHK03 datasets. Specifically, the results below are under single query mode and the detected part and the 767/700 split are used in CUHK03. f_{cls}: global branch, RE: random erasing, f_{rank}: ranking branch, FL: using Focal Loss instead of cross-entropy loss, f_{att}: fully attentional block, OHEM: online hard example mining, CS: curriculum sampling.

Components				baseline			
f_{cls}		√	√	√	√	√	√
RE			√	√	√	√	√
f_{rank}				√	√	√	√
OHEM				√	√	√	
FL					√	√	√
f_{att}						√	√
CS							√
rank1/mAP	Market-1501	69.5/46.1	71.6/47.6	92.4/80.4	92.7/81.0	92.9/81.7	**93.1/82.3**
	CUHK03	33.9/30.8	42.2/38.9	63.8/58.4	63.9/59.2	64.4/60.1	**65.5/60.5**

bounding boxes extracted by DPM. Table 5 shows the results and effectiveness of each component.

Effectiveness of Curriculum Sampling. We further evaluate the effect of CS by comparing with the popular OHEM sampling way. As can be seen in the Table 5, with Market-1501, CS outperforms OHEM by 0.6% on mAP and 0.2% in rank1 matching rate. The improvement can even reach 0.4% and 1.1% in CUHK03, respectively. This shows that the proposed curriculum sampling can help model learn a better representation.

Effectiveness of Full Attentional Block. We verify the effectiveness of attention branch in Table 5. mAP/rank1 are improved 0.7%/0.2% and 0.9%/0.5% on Market-1501 and CUHK03, respectively. FAB provides a fine-grained attention to emphasize the irregular discriminative part of the pedestrian object in an end-to-end way. It is also pluggable and can be added to any existing models.

Effectiveness of Focal Loss. As Table 5 depicts, on Market-1501, focal loss exceeds cross-entropy loss by 0.6%/0.3% in mAP/rank1, respectively. And in CUHK03, the benefit reaches 0.8%/0.1% in mAP/rank1, respectively. Similar to OHEM in triplet loss, focal loss can also mine more information from examples that are hard to classify, which is essential in improving the generalization of the model.

Effectiveness of Random Erasing. Random erasing is not only a way of data augmentation but also helps alleviate occlusion problem by artificially adding occlusion patch to initial image. It makes our model more robust to occlusion situation. Figure 5 also shows that, when combined with a simple classification branch, random erasing can still obtain an obvious improvement.

5 Conclusions

In this paper, we introduce a novel deep network called Mancs to learn stable features for person re-ID. The experiment results on three popular datasets show that Mancs is superior to the previous state-of-art methods. In addition, the effectiveness of the proposed fully attentional block with deep supervision and curriculum sampling have been confirmed in the ablation studies. In the future, we would like jointly investigate the sampling problem for ranking loss and data augmentation methods to obtain more generalizable person re-ID features.

Acknowledgements. This work was partly supported by HUST-Horizon Computer Vision Research Center and NSFC (No. 61733007, 61503145 and 61572207). Xinggang Wang was sponsored by the Program for HUST Academic Frontier Youth Team.

References

1. Almazan, J., Gajic, B., Murray, N., Larlus, D.: Re-ID done right: towards good practices for person re-identification. ArXiv e-prints, January 2018
2. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: CVPR, vol. 6, p. 7 (2017)
3. Bai, S., Bai, X., Tian, Q., Latecki, L.J.: Regularized diffusion process on bidirectional context for object retrieval. TPAMI **37**, 803–815 (2018)
4. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. ArXiv e-prints, November 2017
5. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48. ACM (2009)
6. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of CVPR, vol. 2 (2017)
7. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2590–2600 (2017)
8. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE Trans. Patt. Anal. Mach. Intell. **40**(2), 392–408 (2018)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
10. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep Transfer Learning for Person Re-identification. ArXiv e-prints, November 2016
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
13. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
14. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570 (2015)

15. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 384–393 (2017)
16. Li, J., Zhang, S., Wang, J., Gao, W., Tian, Q.: LVreID: Person Re-Identification with Long Sequence Videos. ArXiv e-prints, December 2017
17. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
18. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 2194–2200 (2017)
19. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR, vol. 1, p. 2 (2018)
20. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
21. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV, pp. 2999–3007 (2017)
22. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 (2017)
23. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017)
24. Lowe, D.G.: Object recognition from local scale-invariant features. In: 1999 The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
25. Manmatha, R., Wu, C., Smola, A.J., Krähenbühl, P.: Sampling matters in deep embedding learning. In: IEEE International Conference on Computer Vision, ICCV, pp. 2859–2867 (2017)
26. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
27. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
28. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3980–3989. IEEE (2017)
29. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 3820–3828 (2017)
30. Ustinova, E., Lempitsky, V.S.: Learning deep embeddings with histogram loss. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, pp. 4170–4178 (2016)
31. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 791–808. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_48
32. Vo, N., Hays, J.: Generalization in Metric Learning: Should the Embedding Layer be the Embedding Layer? ArXiv e-prints, March 2018

33. Wang, H., Gong, S., Xiang, T.: Highly efficient regression for scalable person re-identification. In: Proceedings of the British Machine Vision Conference 2016, BMVC (2016)
34. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016)
35. Zhang, X., et al.: AlignedReID: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
36. Zhao, H., et al.: 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)
37. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: IEEE International Conference on Computer Vision, ICCV, pp. 3239–3248 (2017)
38. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
39. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
40. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person reidentification, vol. 14, p. 13. ACM (2017)
41. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. arXiv preprint arXiv:1707.00408 (2017)
42. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
43. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 3774–3782 (2017)
44. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3652–3661. IEEE (2017)
45. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)