MaskReID: A Mask Based Deep Ranking Neural Network for Person Re-identification

Lei Qi[1], Jing Huo[1], Lei Wang[2], Yinghuan Shi[1], Yang Gao[1⋆]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Computing and Information Technology, University of Wollongong, Australia

Abstract. In this paper, a novel mask based deep ranking neural network with skipped fusing layer (MaskReID) is proposed for person re-identification (Re-ID). For person Re-ID, there are multiple challenges co-exist throughout the re-identification process, including cluttered background, appearance variations (illumination, pose, occlusion, etc.) among different camera views and interference of samples of similar appearance. A compact framework is proposed to address these problems. Firstly, to address the problem of cluttered background, masked images which are the image segmentations of the original images are incorporated as input in the proposed neural network. Then, to remove the appearance variations so as to obtain more discriminative feature, a new network structure is proposed which fuses feature of different layers as the final feature. This makes the final feature a combination of all the low, middle and high level feature, which is more informative. Lastly, as person Re-ID is a special image retrieval task, a novel ranking loss is designed to optimize the whole network. The ranking loss relieved the interference problem of similar samples while producing ranking results. The experimental results demonstrate that the proposed method consistently outperforms the state-of-the-art methods on many person Re-ID datasets, especially large-scale datasets, such as, CUHK03, Market1501 and DukeMTMC-reID.

Keywords: Person Re-identification, Image Segmentation, Fine-grained Image Recognition, Ranking Loss.

1 Introduction

Person re-identification (Re-ID) is to match images of the same individual captured by different cameras with non-overlapping views. Broadly speaking, person Re-ID can also be treated as a special case of the image retrieval problem with the goal of finding a probe image from a large-scale gallery set quickly and accurately [1].

In recent years, person Re-ID has drawn an increasing interest in both academia and industry due to its great potential in surveillance applications,

⋆ Corresponding author.

such as airport security. For instance, when a person appears in the viewing range of a video camera, his/her track log can be recorded by Re-ID techniques. If networks of cameras are widely deployed in cities and smart buildings, the track logs can be used to infer when and where a person is located.

There are many challenging problems of person Re-ID, including cluttered background, appearance variations (illumination, pose, occlusion, resolution, etc.) among different camera views and interference of similar samples. Specifically, cluttered background makes the input image hard for subsequent processing. Appearance variations lead to noise in the extracted feature which requires more discriminative feature extraction method. The last is that as person Re-ID can be seen as an image retrieval problem, there is an interference problem of similar samples while producing ranking results. Many efforts [2][3][4][5] have been made by researchers to address the challenges of person Re-ID. However, most of these previous work focused on extracting discriminative feature. In the early years, the majority of researchers were working on designing robust feature descriptors [3][6]. Based on these hand-crafted feature, some learning methods were proposed to obtain a discriminative representation [7][8]. Recently, deep learning has been widely used in person Re-ID. Compared with the conventional hand-crafted feature extraction methods, there is a big performance improve on many publicly available Re-ID datasets. However, most of these existing work neglected the problem of background clutter and the interference of similar samples. If better input can be found, for example, person image with same plain background, then the feature extraction process can focus more on extracting person specific feature. Therefore, the segmentation result (masked image) is incorporated in the proposed framework for better feature extraction. In addition, as person Re-ID is also an image retrieval problem, most existing Re-ID frameworks are based on contrastive loss or triplet loss, which employ only one negative sample and one positive sample. As the ranking process in fact involves a set of samples, using the above loss may lead to results interfered by similarly looking person. A ranking loss is thus designed in this paper to address the problem.

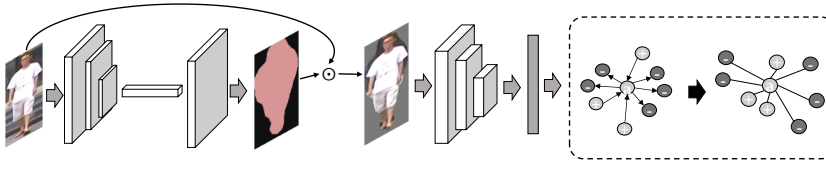


Fig. 1. Overview of the proposed method. The image segmentation network is firstly employed to obtain the segmentation results of a person. Then a novel neural network is proposed to deal with masked images for robust pedestrian feature extrcation. Finally, a novel ranking loss is designed to train the whole deep neural network.

In summary, to address the above mentioned problems, a mask based deep ranking neural network with skipped fusing layer (MaskReID) is proposed, which

is shown in Fig. 1. To reduce the influence of cluttered background, an image segmentation network is employed to obtain the segmentation results of a person. Then, the person image with removed background is adopted as an additional input. Then, for feature extraction, most existing work mainly focused on using the high level feature of a neural network for Re-ID. Considering that person Re-ID is a special fine-grained image recognition task and deep neural networks can extract low, middle and high level feature, the feature of different layers in the deep neural network is fused as the final feature. This strategy can extract variation invariant feature for person from different camera views, as it combined all the low level edge and shape information, middle level structure information and high level semantic information together. Besides, it can also delivery the loss to low layers of neural network for training, making the learning of low level and middle level feature more accurate. Lastly, as person Re-ID is a ranking task, *i.e.*, one person has multiple images from different camera views, a ranking loss function is developed to reduce the influence of similar samples while producing ranking results. As one ranking process can involve multiple positive and negative samples with similar features, it is thus important to take all the involved samples into consideration when designing the loss function. Therefore, in total, the proposed MaskReID framework comprehensively considers all the following three aspects of Re-ID, that are better input with cluttered background removed, a new deep neural network with skipped fusing layer for better feature learning and a ranking loss to optimize the whole network in an end-to-end way. Experimental results demonstrate that the proposed method consistently outperforms the state-of-the-art methods, especially on large-scale Re-ID datasets. The contributions mainly include:

- Masked images that are based on image segmentation are firstly incorporated as an additional input to solve person Re-ID, which illustrates by introducing better input, the performance of Re-ID can be improved.
- A novel deep learning framework is proposed, which accepts both the original and the masked image as input. Besides, there is a skipped fusing layer which combines the feature of different layers. Efficiency of such feature learning scheme is proved by experiments.
- For the person ranking task, a novel ranking loss function is proposed to optimize the weights of the proposed network. Thus, influence of similar samples with different labels are relieved in the ranking process.

2 Related Work

In recent years, many efforts have been made for person Re-ID. In the following, three aspects of related work are introduced.

Hand-crafted Feature: In the early years, the majority of researchers focused on designing robust hand-crafted feature descriptors to represent person images. Due to the appearance variations in pose, viewpoint, illumination, and occlusion in different camera views, feature representation for person Re-ID is

inherently difficult. To deal with these challenges, some color and textural information based methods were designed [9] [10][2][3]. In addition, vision salience is distinctive and reliable information in matching person across no-overlap camera views. Zhao *et al.* [11] exploited the pairwise salience distribution relationship between pedestrian images, and solving the person Re-ID problem. Based on the local information, Liao *et al.* [7] proposed an effective feature representation called Local Maximal Occurrence (LOMO) which analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Besides, some related methods [12][13] were designed based on the idea that person Re-ID is a special image retrieval problem. Another work [14] proposed that person image feature can be obtain by multi-level learning. Multi-level learning guarantees the property of saliency with a similarity constraint. Thus, the multi-level descriptors have a good balance between the robustness and distinctiveness. Since the extraction of the hand-crafted feature is interfered by human factors, it could not obtain the high-level feature of one image .

Deep Framework: Deep learning has been employed in many computer vision tasks. It also improves the performance of person Re-ID. To jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter, a few work [15][16][17][18] was proposed and they address the Re-ID problem using two classification tasks with different matching strategies. Pedestrian attribute information can help to identity one person. Therefore, some deep frameworks [19][20] were designed with attribute learning. Besides convolution neural network, Long Short-Term Memory (LSTM) can also be adopted when person Re-ID is modelled as a time series problem [21][22][22]. In addition, as part model can reduce the influence of person pose, some methods based on local regions were designed [23][4]. Currently, for the person Re-ID problem, there are only a small amount of publicly available large-scale datastes. In particular, Generative Adversarial Network (GAN) can generate samples to facilitate the training. To utilize these generated samples, Label Smoothing Regularization for Outliers (LSRO) was proposed, which assigns a uniform label distribution to the unlabeled generated images, and regularizes the supervised model, leading to an improvement of performance over the baselines [24]. Despite superior performance improvement was achieved by those deep framework, cluttered background among different camera views is still a challenge in person Re-ID. However, most of these existing work neglected the problem of background clutter.

Loss Function: For the research of deep neural network, except for focusing on the structure of the network, related work is also concerned about the design of the loss function [5]. In person Re-ID, most deep frameworks used softmax loss or triplet loss. As softmax loss does not consider compactness of the same class distances, Jin *et al.* [25] proposed to employ identification loss with center loss [26] to train a deep model for person re-identification. For triplet loss, it neglects the same class distances. An improved triplet loss function was designed [27], which pulls the instances of the same person closer and simultaneously pushes

the instances of different persons farther in the learned feature space. Hermans *et al.* [28] used a variant of the triplet loss and realized an end-to-end deep metric learning. The triplet loss pays more attention on obtaining correct orders on the training set. However, it still suffers from a weaker generalization capability on testing set compared with on training set, resulting in inferior performance. To address this problem, Chen *et al.* [29] designed a quadruplet loss, which can make the model output with a larger inter-class variation and a smaller intra-class variation compared with the triplet loss. In addition, sample selection is important for verification loss, Shi *et al.* [30] proposed a novel moderate positive sample mining method to train a robust convolutional neural network which deals with the problem of large variation. To combine the advantage of the classification loss and verification loss, Zheng *et al.* [31] presented a siamese network that simultaneously computes the identification loss and verification loss. In practice, person Re-ID is a ranking task. However, few loss functions are developed based on this task.

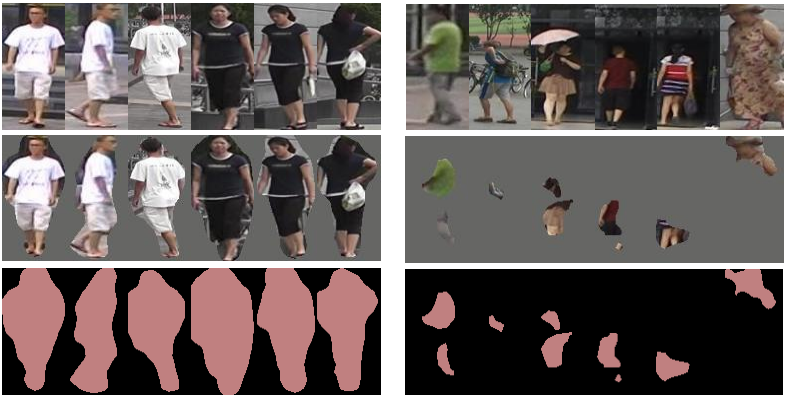
3 Mask Based Deep Ranking Neural Network

The framework of the proposed mask based deep ranking neural network is shown in Fig. 1. An image segmentation deep network is employed to extract person region and remove the background of the original image. Then the original image and the masked image (obtained from the segmentation results) are fed into a novel deep learning framework. The deep learning framework can accept two images as input. Besides, there is a feature fusing layer which combines the low, middle and high level feature as the final feature. Notice, the network structure can be easily extended to accept more inputs, if better inputs can be found besides the masked image. Finally, a ranking loss is developed to optimize the proposed network, which aims to rank person images with the same identity at top positions in a batch.

3.1 A New Mask Based Deep Neural Network Structure

Masked Input Pedestrian images captured under complex scenes (such as airport, station, etc.) can have very messy background. Person images under these situation make the subsequent processing steps hard. The feature extraction module need not only to extract person specific feature, but also try to distinguish the silhouette of the person so as to focus more on the person instead of the background. Thus it can affect the performance of person Re-ID methods. Image segmentation can separate the foreground from the background. Therefore, image segmentation is adopted in this paper to remove the image background. Fully Convolutional Networks (FCN) [32] is employed to obtain the masked images. Some segmentation results of pedestrian by FCN are shown in Fig. 2. For most of the pedestrian images, the segmentation results are good, as shown in Fig. 2 (a). However, there are also a few bad segmentation results, as shown in Fig. 2 (b). The bad segmentation results are partly resulted by low image resolution or

similar foreground and background. For these bad segmented images, if we use them to directly train the deep neural network, it may lead to wrong convergence directions. Therefore, to reduce the influence of bad image segmentations, the original person image is also used as an input. The network is designed to accept both the original and the masked image as input.



(a) Good segmentation results (b) Bad segmentation results

Fig. 2. Some images on Market1501 dataset. Top is the original images. Middle shows the masked images. Bottom displays the image segmentation results.

Skipped Fusing Layer Structure As for the network structure, the network of Xiao *et al.* [33] which has an inception structure is adopted as the basic network. The proposed network structure is shown in Fig. 3. The input of the network is the original images and the masked images. The masked images allow the network to focus on the person regions and the original images can supplement cases when there are bad segmentations. The two images are processed with three separated convolution layers. Then the two separated feature is concatenated and fed into a shared network with three levels, that are low level, middle level and high level. The feature of these three levels is fused by a skipped fusing layer to produce the final feature. The benefits of using different levels of feature have been proved in style transfer tasks [34]. In our experiments, We show in Re-ID, fusing different levels of feature is also beneficial. This is because person Re-ID is a special fine-grained image recognition task, the network should focus on details of the images. Fusing feature of three levels can bring more detailed feature into consideration and thus improves the performance. Another conducive aspect of this network structure is that it can delivery the loss to low layers in a better way, while previous networks experience vanishing gradient problems.

3.2 Ranking Loss

Currently, for the person Re-ID task, most of the existing deep learning methods use classification loss or verification loss (such as triplet loss and contractive loss)

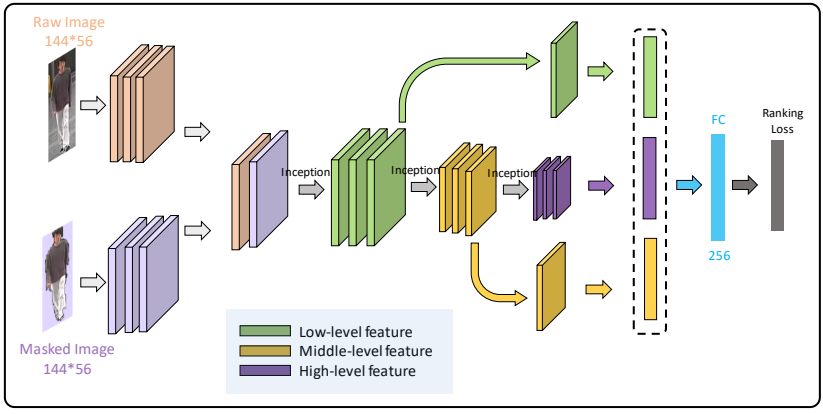


Fig. 3. The proposed framework with person masked images. The proposed network includes the raw image and the masked image, and it outputs 256-d feature. In addition, the scheme of multiple layers fusion is employed in MaskReID.

for training. However, the person Re-ID problem is in fact a ranking problem. One person has multiple images from different camera views and we want samples of the same identity be ranked at top positions. Usually, this task is evaluated via mean Average Precision (mAP).

The definition of mAP [35] is as follows. Let $p = \text{rank}(y)$ and $\hat{p} = \text{rank}(\hat{y})$ be two vector of ranking results with p the true ranking result and $\hat{p}$ the predicted result. y is the true score vector and $\hat{y}$ is the predicted score vector with each element stores the similarity of the sample and the query sample. For each entry in p and $\hat{p}$, there are two kinds of values, that are, the same class images have rank value 1 and irrelevant images have rank value 0. For example, for a corpus of three images ranked in the sequence of I_1, I_2, I_3, if I_1 and I_3 are of the same class with the anchor image (*i.e.*, query image), then $\text{rank}(\hat{y}) = \{1, 0, 1\}$. For one ranking, the mean average precision score is defined as:

$$\text{mAP}(p, \hat{p}) = \frac{1}{rel} \sum_{j:p_j=1} \text{Prec@}j, \tag{1}$$

where $rel = |i : p_i = 1|$ is the number of images with the same identity (ID) of the query image, and Prec@j is the percentage of same ID images in the top j images in the predicted ranking $\hat{y}$. Based on Eq. (1), the mAP loss can be formulated as:

$$\mathcal{L}_{\text{mAP}}(y, \hat{y}) = 1 - \text{mAP}(\text{rank}(y), \text{rank}(\hat{y})). \tag{2}$$

From the above analysis, for the evaluation of ranking tasks, an anchor (*i.e.*, a query image) will interact with multiple positive and negative samples. However, directly minimizing the above loss is time consuming. Inspired by the N-pair loss in [36], a novel ranking loss is proposed for the person Re-ID task. Define

a batch of samples for training a deep network, as $\mathcal{B} = \{I_1, ..., I_{|\mathcal{B}|}\}$. For any $I_k, k = 1, 2, ..., |\mathcal{B}|$, let $\mathcal{B}^+$ and $\mathcal{B}^-$ denote the set of relevant and irrelevant images in $\mathcal{B}$, respectively. $\{x_1, ..., x_{|\mathcal{B}|}\} = \{f(I_1), ..., f(I_{|\mathcal{B}|})\}$, $f(x_k)$ is feature from our neural network, which is normalized by $L2$-norm. The N-pair loss for an anchor sample x_k is defined as:

$$\mathcal{L}_{\text{N-pair}}(x_k) = \log\Big(1 + \sum_{j:x_j^- \in \mathcal{B}^-} \exp\big(\mathcal{S}(x_k, x_j^-) - \mathcal{S}(x_k, x_k^+)\big)\Big), \tag{3}$$

where $\mathcal{S}(x_i, x_j) = x_i^T x_j$ denotes the similarity between two images. x_k^+ is a sample has the same class with x_k. Compared with existing contrastive loss and triplet loss, the advantage of the N-pair loss is that it interacts with multiple negative samples simultaneously. However, an anchor point is only influenced by one positive sample.

For person Re-ID, the ranking process not only need to consider multiple negative samples, but also multiple positive samples. Therefore, based on the N-pair loss, we formulated a new ranking loss:

$$\mathcal{L}_{\text{rank}}(x_k) = \log\Big(1 + \sum_{i:x_i \in \mathcal{B}^+} \sum_{j:x_j \in \mathcal{B}^-} \exp\big(\mathcal{S}(x_k, x_j) - \mathcal{S}(x_k, x_i)\big)\Big). \tag{4}$$

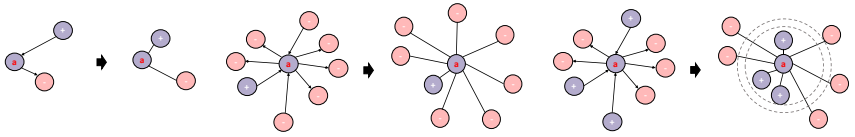
$\mathcal{L}_{\text{rank}}(x_k)$ defines the ranking loss while querying a sample x_k. However, the calculation of Eq. (4) needs to compute $|\mathcal{B}^+| \times |\mathcal{B}^-|$ pairs of samples in one batch while training. To reduce the computation, the most dissimilar positive sample of the query sample is chosen as a reference sample. Meanwhile, to prevent overfitting (*i.e.*, too much attention is paid on samples that have a correct position in the ranking), we only select some negative samples which have small distances with the query sample. Moreover, considering that the same class distances should be closer, the similarity values of all positive samples and the query samples are forced to be close to one (the closest, as features are normalized by $L2$-norm). Thus, by rewriting Eq. (4), we get the final ranking loss used in the proposed neural network as follows:

$$\begin{aligned}\mathcal{L}_{\text{rank}}(x_k) =& \log\Big(1 + \sum_{j:x_j \in \mathcal{B}^-} \exp\big([\mathcal{S}(x_k, x_j) - \min_{i:x_i \in \mathcal{B}^+} \mathcal{S}(x_k, x_i) + \alpha]_+\big)\Big) \\ &+ \frac{\lambda}{2|\mathcal{B}^+|} \sum_{i:x_i \in \mathcal{B}^+} \big(\mathcal{S}(x_k, x_i) - 1\big)^2,\end{aligned} \tag{5}$$

where $[t]_+$ denotes $\max(0, t)$ which is the hinge loss. α is the margin as defined in [37]. In Eq. (5), the first term makes the negative samples and the most dissimilar positive sample has a gap. The objective of the second term is to make all positive samples similar to the query image. λ is a parameter to balance the two terms.

Compared with the conventional verification loss, the advantage of the proposed rank loss is that it simultaneously considers multiple positive and negative

samples in each update. Thus, it can make the same ID images from different views more similar to query images, and the distances of the different ID images become large, as shown in Fig. 4.



(a) Triplet loss (b) N-pair loss (c) The proposed ranking loss

Fig. 4. Comparison between triplet loss, N-pair loss and the proposed ranking loss. For triplet loss, the anchor only interacts with one positive sample and one negative sample. For N-Pair loss, the anchor interacts with one positive sample and mulitiple negative samples. However, in the proposed ranking loss, the anchor interacts with multiple positive and negative samples in each update.

4 Experiments

4.1 Datasets and Evaluation Protocol

In this paper, we employed multiple datasets including small-scale datasets and large-scale datasets to validate the effectiveness of the proposed method.

Small-Scale Datasets: In the early years, there are only some small-scale datasets consists of a small number of identities from few cameras for Person Re-ID. VIPeR [38] is one of the most challenging datasets among this kind. It has 632 person with various poses, viewpoints, image resolutions and lighting conditions. 3DPeS [39] has 193 identities but the number of images for each person is not fixed. iLIDS [40] contains images of 119 persons. The images are captured by surveillance cameras in an airport which leads to large occlusions due to luggages and heavy crowd. PRID [41] extracts pedestrian images from recorded trajectories of video frames. It has two camera views, each contains 385 and 749 identities, respectively. But only 200 of them appeared in both views. Shinpuhkan [42] is another dataset with more than 22,000 images. The highlight of this dataset is that it contains only 24 individuals, and all of them are employed to train our proposed network. CUHK01 [43] has two camera views and 971 person in total.

Large-Scale Datasets: In recent years, with the extensive application of deep learning in Re-ID, several large-scale datasets have been published. CUHK03 [15] is one of the largest, it consists of five different camera views and has more than $14,000$ images of $1,467$ person. The Market1501 dataset contains bounding boxes from a person detector which was selected based on their intersection-over-union overlap with manually annotated bounding boxes. It contains 32,668 images of 1,501 person. As defined by [12], the dataset is split into training/testing

sets of 12,936/19,732 images respectively. The DukeMTMC-reID dataset [44] contains 1,404 identities, 16,522 training images, 2,228 queries and 17,661 gallery images. With the large-scale images captured by 8 cameras, DukeMTMC-reID manifests itself as one of the most challenging Re-ID datasets up to now.

Evaluation Protocol: We followed the standard evaluation protocol. Concretely, the cumulative matching characteristics (CMC) at rank-1 and mean average precision (mAP) are adopted for performance evaluation on Market1501 and DukeMTMC-reID. The precision of rank-1 is reported for CUHK03 and all the small-scale datasets. Following most of the related work [23][45], experiments on Market1501 are performed under single query and multiple query settings.

4.2 Implementation Details

For the small-scale datasets, all datasets and CUHK03 are joined together to train a model. The split scheme to divide the dataset into the training and testing set is the same as the work of Xiao *et al.*[33]. Since many identities have only two images on these datasets, such as VIPeR, softmax loss is employed for training the proposed deep model. For Market1501 and DukeMTMC-reID datasets, based on the pre-trained model which is trained on small-scale datasets, two networks are trained separately with the proposed ranking loss on the two datasets.

In some existing work [28], selecting images to form batches is done by randomly sampling P classes (*i.e.*, person identities), and then randomly sampling K images from each class (person), resulting in a batch of $P \times K$ images. In this paper, we randomly select one person, then P images with the same identity are chosen randomly. Afterwards, N negative samples are sampled from N person randomly (*i.e.*, one person has only a image). Therefore, in one batch, we can have more negative samples than conventional sampling schemes, which is more reasonable for a ranking task. In the experiment, we set P/N to 10/54.

4.3 Comparison with Related Methods

In this section, various experiments are conducted on multiple datasets to validate the effectiveness of the proposed method.

Results on Small-Scale Datasets: For the small-scale datasets, all datasets are combined to train a model. The settings are the same as those in [33]. Since several person only have two images, softmax loss is used for training our network and we evaluate the effectiveness of the masked image as input and the feature fusion structure. The proposed method is compared with hand-crafted feature methods [7][2][14], different deep learning methods [18][21][33][4] and different loss functions based methods [27][27][29]. Experimental results are reported in Table 1. In the table, MaskReID-M and MaskReID-MF are our methods. MaskReID-M denotes the proposed framework which employs the masked image. MaskReID-MF represents the proposed framework employs both the masked image and the feature fusion strategy.

Table 1. Comparison of rank-1 results. Note that $1^{st}/2^{nd}$ best in red/blue. MaskReID-M denotes the proposed framework which only employs the masked image. MaskReID-MF denotes the proposed framework which employs both the masked image and the feature fusion strategy.

Method	VIPeR	PRID	3DPeS	iLIDS	CUHK01	CUHK03	Reference
LOMO [7]	40.00	–	–	–	–	52.20	CVPR2015
GOG [2]	49.70	–	–	–	57.80	67.30	CVPR2016
WLC [14]	51.40	–	–	–	65.80		AAAI2017
MGCNN [18]	37.80	–	–	–	–	68.10	ECCV2016
SLSTM [21]	42.40	–	–	–	–	57.30	ECCV2016
DGD [33]	38.60	64.00	56.00	64.60	66.60	75.30	CVPR2016
Spindle [4]	53.80	67.00	62.10	66.30	79.90	88.50	CVPR2017
TCP [27]	47.80	–	–	–	53.70	–	CVPR2016
P2S [27]	–	–	71.16	–	77.34	–	CVPR2017
Quadruplet [29]	49.05	–	–	–	81.00	75.53	CVPR2017
MaskReID-M (Our)	44.62	65.00	66.12	69.57	84.26	88.75	This paper
MaskReID-MF (Our)	45.57	70.00	68.60	70.43	84.05	92.25	This paper

From the table, some observations can be made which are as follows: (1) **Deep learning based methods have better performance than the hand-crafted feature based methods.** This is mainly due to the strong learning ability of deep learning. Besides, the extraction of feature is not interfered by human factors. Thus, it can automatically extract key information from the original image and transform the information to better feature representation. (2) **Using the masked image as additional input and the multiple layer feature fusion scheme are both effective.** The proposed method adopts the structure of DGD [33] as the basic network structure. Compared with DGD network, MaskReID-M has a big performance improve. This illustrates using masked image as additional input is effective. Moreover, MaskReID-MF has better performance than MaskReID-M. This indicates the multiple layer feature fusion scheme is also effective. (3) **Performance varies on different datasets with better performance achieved on large-scale dataset.** The proposed network is trained on merged multiple datasets. However the performance varies on different datasets. The testing results on big datasets tend to have better performance. In particular, the rank-1 is up to 92.25% on the CUHK03. This may because it has more images on the training set. Besides, results on small datasets are not competitive as those on large datasets, such as results on VIPeR. The relatively poor performance on VIPeR is partly because the images of VIPeR are of low resolution which makes the segmentation results poor. (4) **The proposed method is competitive compared with other state-of-the-art deep learning methods for person Re-ID.** Employing the masked image without background information can facilitate the feature learning process. Besides, the multiple layer feature fusion structure can combine all the low, middle and high level features together which is more detailed and informative. These two schemes make the proposed method effective for person Re-ID.

Results on Large-Scale Datasets: The proposed method is also validated on large-scale datasets with a large number of person from multiple cameras. On DukeMTMC-reID and Market1501 datasets, the deep model is trained with different loss functions including softmax loss, triplet loss and the proposed ranking loss. Specifically, we trained our model with triplet loss by online hard sample mining and used the same parameters (including the number of iterations, learning rate, batch size, etc.) with our ranking loss based network. In this experiment, the proposed method is also compared with a set of the state-of-the-art methods. Results are given in Table 2 and 3.

From the two tables, we can observe that: (1) **Compared with softmax loss, employing the proposed ranking loss can improve the performance, especially for mAP.** The reason is that the proposed ranking loss pulls the intra-class samples closer and push the inter-class samples farther away. The ranking loss minimization process takes multiple positive and negative samples into consideration, leading to better ranking results and thus better mAP. (2) **The proposed ranking loss is more effective than triplet loss.** As person Re-ID is a special retrieval task, the proposed ranking loss is more effective for this task. Experimental results validated the superiority of our ranking loss compared with triplet loss. (3) **The proposed method consistently outperformed the state-of-the-art methods on large-scale datasets.** Especially, on Market1501 datasets, rank-1/mAP of the multiple queries are up to 93.32%/82.29%. (4) **Employing the re-ranking method [46] can further improve the performance of our proposed method.** Re-ranking method [46] is commonly used for Re-ID. We also tried to combine re-ranking techniques with our method. As can be seen in both Table 2 and 3, performance can be further improved. On Market1501 dataset, mAP is now up to 91.94%.

Table 2. Single and multiple query results are reported on Market1501. Note that $1^{st}/2^{nd}$ best in red/blue. The subscripts as *soft*, *trip* and *rank* denote that training the network employ softmax loss, tiplet loss and our proposed ranking loss, respectively.

	Single query		Multiple query		
Method	Rank-1	mAP	Rank-1	mAP	Reference
MGCNN [18]	65.88	39.55	76.04	48.45	ECCV2016
SLSTM [21]	61.60	35.31	–	–	ECCV2016
DLPAR [47]	81.00	63.40	–	–	ICCV2017
SVDNet [48]	82.30	62.10	–	–	ICCV2017
Spindle [4]	76.90	–	–	–	CVPR2017
MSCAN [23]	80.31	57.53	86.79	66.70	CVPR2017
P2S [27]	70.72	44.27	85.78	55.73	CVPR2017
SSM [45]	82.21	68.80	88.18	76.18	CVPR2017
JLML [49]	83.90	64.40	89.70	74.50	IJCAI2017
MaskReID$_{soft}$	88.18	70.57	91.48	78.06	This paper
MaskReID$_{trip}$	85.30	65.84	90.71	74.71	This paper
MaskReID$_{rank}$	90.02	75.30	93.32	82.29	This paper
MaskReID$_{rank+re\text{-}ranking}$	92.04	88.03	94.18	91.94	This paper

Table 3. Experimental results on DukeMTMC-reID. Note that $1^{st}/2^{nd}$ best in red/blue. The subscripts as *soft* and *rank* denote that training the network employ softmax loss and our proposed ranking loss, respectively.

Method	Rank-1	mAP	Reference
LSRO [24]	67.68	47.13	ICCV2017
SVDNet [48]	76.70	56.80	ICCV2017
OIM [50]	68.10	–	CVPR2017
ACRN [51]	72.58	51.96	CVPRW2017
MaskReID$_{\text{soft}}$	78.01	59.08	This paper
MaskReID$_{\text{rank}}$	78.86	61.89	This paper
MaskReID$_{\text{rank+re-ranking}}$	84.07	79.73	This paper

4.4 Parameter Analysis of the Loss Function

In this part, parameters of the proposed ranking loss function are analysed, *i.e.*, parameters in Eq. (5). Eq. (5) has two terms, where first term is to force a margin between positive and negative samples and the second term is designed to make all the positive samples similar to the query image. λ is a parameter to trade off the two terms. α is the margin which controls the distances between positive and negative samples. Several experiments are carried out with different λ and α. Results are reported in Table 4. In particular, we want to validate the effectiveness of each term. For α, it is set to 0.1 and 0.15. λ is set to $0, 1, 2, 5, 10$.

From the table, we can observe that: (1) when λ is set to 0, *i.e.*, removing the second item in Eq. (5), the performance is the worst. Especially, there is a large decrease on mAP. This demonstrates that enhancing the similarity between all the positive images and the query image in the optimization process is useful. (2) In addition, when λ increases to values larger than 2, the results have slight reduction. The illustrates focusing to much on the second term of Eq. (5) is also inadvisable. We need a reasonable λ to balance the two terms in Eq. (5). (3) The results of setting α to 0.1 and 0.15 are similar, with the results of 0.15 slightly better. α is margin parameter which should not be too large as larger α may easily lead to overfitting problems. In conclusion, we set λ/α to 2/0.15 in all experiments.

Table 4. Performance on Market1501 with different α and λ. Note that red/green denote best/worst results.

α	0.1		0.15	
λ	Rank-1	mAP	Rank-1	mAP
0	89.79	72.78	89.76	72.71
1	89.85	74.97	89.93	75.22
2	90.05	75.12	90.02	75.30
5	89.90	75.23	89.88	75.04
10	89.88	74.90	89.22	74.42

5 Conclusion

In this paper, a mask based deep ranking neural network is proposed to deal with person Re-ID. First, to reduce the influence of messy background in different camera views, the masked images together with the original images are used as input in the proposed deep neural network. Second, considering that person re-identification is a special fine-grained image recognition task and deep neural network can extract low, middle and high level feature with different layers, the feature of different layers in the deep neural network is fused as the final feature. Third, based on the specific ranking task, we proposed a novel ranking loss function to optimize the weights of the network to reduce the discrepancy of data-distribution in different camera views. The proposed method outperforms the state-of-the-art methods on many large-scale datasets.

Since the part model can reduce the interference of variance of human pose in different views, several part models are proposed to deal with the person Re-ID problem. MaskReID is a global based method, *i.e.*, employing the whole image as input, without considering body part information. In the future, we will exploit some strategies to combine the masked image and the body part information to further improve the performance of the person Re-ID task.

References

1. Zhang, R., Lin, L., Zhang, R., Zuo, W., Zhang, L.: Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Trans. Image Processing **24**(12) (2015) 4766–4779
2. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2016) 1363–1372
3. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7) (2013) 1622–1634
4. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 907–915
5. Zhou, S., Wang, J., Wang, J., Gong, Y., Zheng, N.: Point to set similarity based deep feature learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 5028–5037
6. An, L., Kafai, M., Yang, S., Bhanu, B.: Person reidentification with reference descriptor. IEEE Transactions on Circuits and Systems for Video Technology **26**(4) (2016) 776–787
7. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2015) 2197–2206
8. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition,CVPR. (2012) 2288–2295

9. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2010) 2360–2367
10. Varior, R.R., Wang, G., Lu, J., Liu, T.: Learning invariant color features for person reidentification. IEEE Trans. Image Processing **25**(7) (2016) 3395–3410
11. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: IEEE International Conference on Computer Vision, ICCV. (2013) 2528–2535
12. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision, ICCV. (2015) 1116–1124
13. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2015) 1741–1750
14. Yang, Y., Wen, L., Lyu, S., Li, S.Z.: Unsupervised learning of multi-level descriptors for person re-identification. In: Proceedings of the Thirty-First Conference on Artificial Intelligence, AAAI. (2017) 4306–4312
15. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2014) 152–159
16. Ahmed, E., Jones, M.J., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2015) 3908–3916
17. Subramaniam, A., Chatterjee, M., Mittal, A.: Deep neural networks with inexact matching for person re-identification. In: Annual Conference on Neural Information Processing Systems, NIPS. (2016) 2667–2675
18. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: Computer Vision European Conference, ECCV. (2016) 791–808
19. Matsukawa, T., Suzuki, E.: Person re-identification using CNN features learned from combination of attributes. In: International Conference on Pattern Recognition, ICPR. (2016) 2428–2433
20. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: Computer Vision European Conference, ECCV. (2016) 475–491
21. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: Computer Vision European Conference, ECCV. (2016) 135–153
22. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Processing **26**(7) (2017) 3492–3506
23. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 7398–7407
24. 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) 3774–3782
25. Jin, H., Wang, X., Liao, S., Li, S.Z.: Deep person re-identification with improved embedding. arXiv (2017)
26. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer Vision European Conference, ECCV. (2016) 499–515

27. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2016) 1335–1344
28. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv (2017)
29. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 1320–1329
30. Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identification: A study against large variations. In: Computer Vision European Conference, ECCV. (2016) 732–748
31. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person re-identification. arXiv (2016)
32. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2015) 3431–3440
33. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2016) 1249–1258
34. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2016) 2414–2423
35. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of the 30th Annual International Conference on Research and Development in Information Retrieval SIGIR. (2007) 271–278
36. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems,NIPS. (2016) 1849–1857
37. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2015) 815–823
38. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS). Volume 3., Citeseer (2007) 1–7
39. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 2011 joint workshop on Human gesture and behavior understanding. (2011) 59–64
40. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference, BMVC. (2009) 1–11
41. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Image Analysis Conference, SCIA. (2011) 91–102
42. Kawanishi, Y., Wu, Y., Mukunoki, M., Minoh, M.: Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: Joint Workshop on Frontiers of Computer Vision. Volume 5. (2014)
43. Li, W., Wang, X.: Locally aligned feature transforms across views. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3594–3601
44. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, ECCV, Springer (2016) 17–35

45. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 3356–3365
46. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 3652–3661
47. 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. (2017) 3239–3248
48. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: IEEE International Conference on Computer Vision, ICCV. (2017) 3820–3828
49. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI. (2017) 2194–2200
50. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2017) 3376–3385
51. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW. (2017) 1435–1443