arXiv:1710.00478v2 [cs.CV] 4 Oct 2017

Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification

Qiqi Xiao , Hao Luo , Chi Zhang
qiqix@andrew.cmu.edu, luohaocsc@zju.edu.cn, zhangchi@megvii.com

Abstract

Person re-identification (ReID) is an important task in computer vision. Recently, deep learning with a metric learning loss has become a common framework for ReID. In this paper, we propose a new metric learning loss with hard sample mining called margin smaple mining loss (MSML) which can achieve better accuracy compared with other metric learning losses, such as triplet loss. In experiments, our proposed methods outperforms most of the state-of-the-art algorithms on Market1501, MARS, CUHK03 and CUHK-SYSU.

1. Introduction

Person re-identification (ReID) is an important and challenging task in computer vision. It has many applications in surveilance video, such as person tracking across multiple cameras and person searching in a large gallery *etc*. However, some issues make the task difficult such as large variations in poses, viewpoints, illumination, background environments and occlusion. And the similarity of appearances among different persons also increases the difficulty.

Some traditional ReID approaches focus on low-level features such as colors, shapes and local descriptors [9, 11]. With the development of deep learning, the convolutional neural network (CNN) is commonly used for feature representation [27, 38, 6]. The CNN based methods can present high-level features and thus improve the performance of person ReID. In supervised learning, current methods can be divided into representation learning and metric learning in terms of the target loss. For the representation learning, ReID is considered as a verification or identification problem. For instance, in [57], the authors make the comparison between the verification baseline and the identification baseline: (1) For the former, two images are judged whether they belong to the same person. (2) For the latter, the method treats each identity as a category, and then minimizes the softmax loss. In some improved work, Lin et al. combined the verification loss with attributes loss [20], while Matsukawa et al. combined the identification loss with attributes loss [27]. Representation learning based methods have prominent advantages, having reasonable performance and being easily trained and reproducible. But those methods do not care about the similarity of different pairs, leading it difficult to distinguish between pairs of same persons and different persons. To mitigate that problem, different distance losses, such as contrastive loss [38], triplet loss [22], improved triplet loss [6], quadruplet loss [3], *etc*. are proposed. And [13] also proposes hard batch by sampling hard image pairs. These methods can directly evaluate the similarity of two input images according to their embedding features. Although these distance losses are sensitive to image pairs, which increases the training difficulty, they can generally get better performance than representation learning based methods.

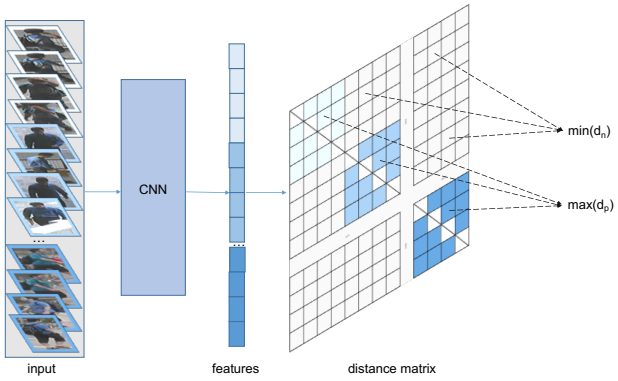


Figure 1. Framework of our method. Input data are designed to be groups of identities. Distace matrix of features extracted by CNN is calculated. The minimum of negative pair distances and the maximum of positive pair distances are sent to the loss function.

In this paper, we propose a novel metric learning loss with hard sample mining called margin smaple mining loss (MSML). It can minimize the distance of positive pairs while maximizing the distance of negative pairs in feature embedding space. For original triplet or quadruplet loss, the pairs are randomly sampled. In our method, we put each K images of P persons into a batch, and then calculate an $N \times N$ distance matrix where $N = K \times P$ denotes the

batch size. Then, we choose the maximum distance of positive pairs and the minimum distance of negative pairs to calculate the final loss. In this way, we sample the most dissimilar positive pair and the most similar negative pair, both of which are hardest to be distinguished in the batch. On Market1501, MARS, CUHK03 and CUHK-SYSU, our method outperforms most of state-of-the-art ones.

In the following, we overview the main contents of our method and summarize the contributions:

- We propose a new loss with extremely hard sample mining named margin smaple mining loss, which outperforms other metric learning losses on person ReID task.

- Our method shows significant performance on those four datasets, being superior to most of state-of-the-art methods.

The paper is organized as follows: related works with more details are presented in section 2. In section 3, we introduce our MSML. Datasets and experiments are presented in section 4. Conclusions and outlook are presented in section 5.

2. Related Work

2.1. Deep convolutional networks

Including AlexNet(CaffeNet) [16], GoogleNet [36] and Resnet [12] etc. , several popular deep networks have been proposed in the past few years. A lots of works show that Resnet is better than other baseline models on person ReID task [60, 55, 59]. Most current paper choose Resnet50 pre-trained on the ImageNet LSVRC image classification datasets [32] as baseline networks. In this paper, we also choose Resnet50 as our baseline network but reconstruct it.

Resnet is the origin of deep residual networks, and there are some improved versions such as ResNeXt [45], DenseNet [14] and ShuffleNet [51] . All these works use efficient channel-wise convolutions or group convolutions into the building blocks to balance representation capability and computational cost. Different from traditional regular convolutions, group convolutions divide feature maps into several groups concatenated together after respective convolutions. The channel-wise convolutions in which the number of groups equal to the number of channels is a special case of group convolutions. Channel-wise convolutions can effectively reduce computational cost. GoogLeNet Xception [7] uses a large number of channel-wise convolutions. Using building blocks designed with group convolutions and channel-wise convolutions to replace regular convolutions in Resnet is popular and improves accuracy with less computational cost.

2.2. Deep metric learning

Before deep learning, most traditional metric learning methods focus on learning a Mahalanobis distance in Euclidean space. Cross-view Quadratic Discriminant Analysis (XQDA) [19] and Keep It Simple and Straightforward Metric Learning (KISSME) [15] were both classic metric learning methods in person ReID in the past. However, deep metric learning methods usually transform raw images into embedding features, and then compute the similarity scores or feature distances directly in Euclidean space.

In deep metric learning, two images of the same person are defined as a positive pair while two of different persons are a negative pair. The triplet loss is motivated by the threshold enforced between positive and negative pairs. In improved triplet loss, a distance loss of positive pairs is used to reinforce clustering of the same person images in the feature space. The positive pair and the negative one in a triplet share a common image. A triplet only has two identities. Quadruplet loss adds a new negative pair, and a quadruplet samples four images from three identities. For the quadruplet loss, a new loss enforces the distance between positive pairs of one identity and negative pairs of the other two identities. Deep metric learning methods is sensitive to the samples of pairs. Selecting suitable samples for training model by hard mining is shown to be effective [13, 3]. A common practice is to pick out dissimilar positive pairs and similar negative pairs according to similarity scores. Compared with identification or verification loss, distance loss for metric learning can lead to a margin between inter-class distance. But combining softmax loss with distance loss to speed up convergence is also popular.

2.3. Other proposed ReID methods

Some successful unsupervised or transfer learning methods have been proposed recently [8, 30, 29, 48]. One important concern is that there exists bias among datasets collected in different environments. Another problem is the lack of labeled data, which can cause overfitting easily. Despite that supervised learning methods based on CNN have been successful in some certain dataset, the network trained with that dataset could perform poorly on other datasets. There, one method of transfer learning is to train one task with one dataset, and then fine-tune from the trained model to train another task with another dataset. For example, the model trained on one dataset clusters the other dataset to predict the fake labels which are used to fine-tune the model [8]. In [29], an unsupervised multi-task dictionary learning is proposed to solve dataset-biased problem.

In addition, some paper focus on getting better global or local features. For instance, pose invariant embedding (PIE) aligns pedestrians to a standard pose to reduce the impact of pose [55] variation. Natural language description [17] and image data generated by generative adversarial networks

(GANs) [59] are respectively regarded as additional information input into networks. Inspite of image-based learning methods above, there are some video-based person ReID works, which take into account the sequence information such as motion or optical flow [41, 47, 46, 25, 28, 54, 23]. RNN architectures and attention model are also applied into embedding sequence features.

After getting image features, most current works choose L2 euclidean distance to compute similarity score for ranking or retrieval task. In [40, 60, 1], some re-ranking methods are proposed and obviously improve the ReID accuracy.

3. Our Method

Despite the deep network for feature extracting, our method includes a metric learning loss with hard sample mining called MSML.

3.1. Margin Sample Mining Loss for Metric Learning

The goal of metric embedding learning is to learn a function $g(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$ which maps semantically similar instances from the data manifold in $\mathbb{R}^F$ onto metrically close points in $\mathbb{R}^D$ [13]. The deep metric learning aims to find the function through minimizing the metric loss of training data. Then we should define a metric function $D(x, y) : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ to measure the distances in the embedding space. The distances are used to re-identity the person images.

3.1.1 Related Metric Learning Methods

One of the widely used metric learning loss is triplet loss [33] which helps generate features as discriminative as possible compared to softmax loss for classification. It is trained on groups of triplets. A triplet contains three different images$\{I_A, I_{A'}, I_B\}$, where I_A and $I_{A'}$ are images of the same identity while I_B is an image of a different identity. Each image would generate one extracted feature after a deep network. A triplet of ℓ_2-normalized features $\{f_A, f_{A'}, f_B\}$ would be used to calculate distances and the triplet loss is formulated as following:

$$L_{trp} = \frac{1}{N}\sum^{N}\left(\overbrace{\|f_A - f_{A'}\|_2}^{\text{to shorten}} - \overbrace{\|f_A - f_B\|_2}^{\text{to largen}} + \alpha\right)_+ \quad (1)$$

where $(z)_+ = max(z, 0)$ [31], and α is the value of the margin set to allow the network distinguish the positive samples with the negative ones. The first term shortens the distances of positive pairs, while the second term largens the distances of negative pairs. In triplet loss, each positive pair and negative pair share one same image, which makes it pay more attention to obtaining correct orders for pairs *w.r.t.* the same probe image. As a result, it suffers poor generalization, and is difficult to be applied for tracking tasks.

The quadruplet loss [3] extends the triplet loss by adding a different negative pair. A quadruplet contains four different images $\{I_A, I_{A'}, I_B, I_C\}$, where I_A and $I_{A'}$ are images of the same identity while I_B and I_C are images of another two identities respectively. Accordingly, a quadruplet of ℓ_2-normalized features $\{f_A, f_{A'}, f_B, f_C\}$ would be used to calculate distances. The quadruplet loss is formulated as following:

$$L_{quad} = \frac{1}{N}\sum^{N}\left(\overbrace{\|f_A - f_{A'}\|_2 - \|f_A - f_B\|_2 + \alpha}^{\text{relative distance}}\right)_+ + \frac{1}{N}\sum^{N}\left(\overbrace{\|f_A - f_{A'}\|_2 - \|f_C - f_B\|_2 + \beta}^{\text{absolute distance}}\right)_+ \quad (2)$$

where α and β are the values of the margins in two terms. The first term is the same as (1), which focuses on the distance between positive pairs and negative pairs containing one same probe image. The second term considers the distance between positive pairs and negative pairs which contain different probe images. With the second constraint, an inter-class distance is supposed to be larger than an intra-class distance. In[3], the margin β is set to be smaller than the margin α to achieve a relatively weeker constraint, so the second term does not play the leading role.

However, we can well combine these two terms into one and extend (2) to:

$$L_{quad'} = \frac{1}{N}\sum^{N}\left(\|f_A - f_{A'}\|_2 - \|f_C - f_B\|_2 + \alpha\right)_+ \quad (3)$$

where C can share the same identity with A or not.

A direct application of the loss given in (3) does not achieve good performance. The reason is that the possible number of quadruplets grows rapidly as the dataset gets lager. The number of all the pairs generated from the quadruplets increases accordingly. Most of the samples are relatively easy, especially for the negative pairs, the number of which is squarely larger than that of positive ones. Although a margin is set to restrict the distance between positive and negative pairs, most samples are still too easy to the network, causing the "precious" hard samples overwhelmed and limiting the model performance. In order to relieve this, we apply hard sample mining as in [13]. Triplet loss with hard sample mining computes a batch of samples together. In each batch, it contains different identities, each of which have the same number of samples. For each sample, it picks the most dissimilar sample with the same identity and the most similar sample with a different identity to

get a triplet. In [13], the triplet loss with hard sample mining is formulated as following:

$$L_{trihard} = \frac{1}{N} \sum_{A \in batch} \left(\overbrace{\max_{A'}(\|f_A - f_{A'}\|_2)}^{\text{hard positive pair}} - \overbrace{\min_{B}(\|f_A - f_B\|_2)}^{\text{hard negative pair}} + \alpha \right)_+ \tag{4}$$

where N is the batch size. With hard sample mining, easy samples are filtered and thus improving the robustness of the model.

3.1.2 Margin sample mining loss

We apply a new hard example mining strategy for (3) named margin smaple mining loss (MSML). It picks the most dissimilar positive pairs and the most similar negative pair in the whole batch, as:

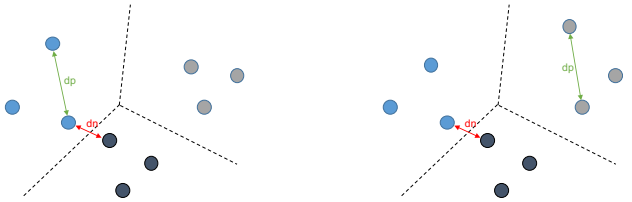
$$L_{eml} = \left(\overbrace{\max_{A,A'}(\|f_A - f_{A'}\|_2)}^{\text{hardest positive pair}} - \overbrace{\min_{C,B}(\|f_C - f_B\|_2)}^{\text{hardest negative pair}} + \alpha \right)_+ \tag{5}$$

where C and B can share the same identity with A or not.

As shown in Figure 2, the connections are extremely sparse, only two pairs in a batch participating in training phase. There are two examples in our margin smaple mining loss. In Figure 2(a), the positive pair and the negative pair have one common identity, which considers the relative distance. It covers the samples that the triplet loss (or with hard sample mining) can get. And in Figure 2(b), the positive pair and the negative pair do not have any common identities, which considers the absolute distance. Therefore, it can cover the second term of quadruplet loss. It seems that we waste a lot of training data. But the two chosen pairs are determined by all the data of one batch. With the loss reducing, not only the two chosen pairs, but the distances of most positive pairs and negative pairs will get larger. In addition, we randomly sample the training data in each batch, which allows the pairs diversity as training epoch grows.

In (5), the first term is the upper bound of the distance of all positive pairs, and the second term is the lower bound of the distance of all negative pairs in a batch. Different from other metric learning losses, which push away positive pairs and negative pairs by each sample, our MSML push away the bounds of two sets in a batch. With training epoch growing, there is a sharp demarcation between positive pairs and negative pairs in feature embedding space. We think it is a useful characteristic for some special tasks.

In summary, compared with other metric learning losses, our MSML has following advantages. First, MSML not only considers the relative distances between positive and negative pairs containing the same probe sample, but also considers absolute distances between positive and negative pairs from different probe samples. Second, it inherits the advantage of hard sample mining and other approaches. And we extend it to edge mining, which leads to a better performance. Finally, we think our MSML is easy to implement and combine with other methods.



(a) Relative distance (b) Absolute distance

Figure 2. Two examples of edge mining samples.

4. Experiments

We first conduct two sets of experiments: 1) to compare different networks on person ReID tasks; 2) to evaluate the performance of different losses. Then we compare the proposed approach with other state-of-the-art methods. Note that train a single model using all datasets as [42, 53].

4.1. Datasets

We use public datasets including CUHK03 [18], CUHK-SYSU [43], Market1501 [56] and MARS [35] in our experiments.

CUHH03 contains 14,097 images of 1,467 identities. It provides the bounding boxes detected from deformable part models (DPM) and manually labeling. In this paper, we evaluate our method on the labeled set. Following the evaluation procedure in [18], we randomly pick 100 identities for testing. Since we train one single model for all benchmarks, it is a bit different from the standard procedure, which splits the dataset randomly for 20 times. We only split the dataset once for training and testing.

CUHK-SYSU is a large scale benchmark for person search, containing 18,184 images and 8,432 identities. The dataset is close to real world application scenarios for images are cropped from whole images. The training set contains 11,206 images of 5,532 query persons while the test set contains 6,978 images of 2,900 persons.

Market1501 contains more than 25,000 images of 1,501 labeled persons of 6 camera views. There are 751 identities in the training set and 750 identities in the testing set. In average, each identity contains 17.2 images with different appearances. All images are detected by the DPM detector and thus include 2,793 false alarms to mimic the real scenario.

MARS (Motion Analysis and Re-identification Set) dataset is an extenstion verion of the Market1501 dataset. It

Table 1. Comparison of different methods. Cls stands for classfication, Tri stands for triplet loss [33], TriHard stands for triplet loss with hard sample mining [13], Quad stands for quadruplet loss [3] and MSML stands for our margin smaple mining loss. We combine metric learning loss above with classification loss.

Base model	Methods	Market1501			MARS			CUHK-SYSU			CUHK03		
		mAP	r = 1	r = 5	mAP	r = 1	r = 5	mAP	r = 1	r = 5	r=1	r=5	r = 10
Resnet50	Cls	41.3	65.8	83.5	43.3	59.3	75.2	70.7	75.0	88.1	51.2	72.6	81.8
	Tri	54.8	75.9	89.6	62.1	76.1	89.6	82.6	85.1	94.1	73.0	92.0	96.0
	Quad	61.1	80.0	91.8	62.1	74.9	88.9	85.6	87.8	95.7	79.1	95.3	97.9
	TriHard	68.0	83.8	93.1	71.3	82.5	92.1	82.4	85.1	94.7	79.5	95.0	98.0
	MSML	**69.6**	**85.2**	**93.7**	**72.0**	**83.0**	**92.6**	**87.2**	**89.3**	**96.4**	**84.0**	**96.7**	**98.2**
Inception-v2	Cls	40.7	66.3	84.1	45.0	62.6	77.9	74.2	78.2	89.7	50.5	68.8	77.4
	Tri	57.9	78.3	91.8	55.5	70.7	85.2	87.7	89.7	96.6	76.9	93.7	97.2
	Quad	66.2	83.9	93.6	65.3	77.8	89.9	88.3	90.2	96.6	81.9	96.1	98.3
	TriHard	73.2	86.8	**95.4**	74.3	84.1	93.5	83.5	86.1	95.2	85.5	97.2	98.7
	MSML	**73.4**	**87.7**	95.2	**74.6**	**84.2**	**95.1**	**88.4**	**90.4**	**96.8**	**86.3**	**97.5**	**98.7**
Resnet50-X	Cls	46.5	70.8	87.0	48.0	63.8	80.2	74.2	78.2	89.7	57.2	77.7	85.6
	Tri	69.2	86.2	94.7	68.2	79.5	91.7	**89.6**	**91.4**	97.0	82.0	96.3	98.4
	Quad	64.8	83.3	93.8	63.6	77.7	89.4	87.3	89.6	96.2	80.7	94.9	97.9
	TriHard	71.6	86.9	94.7	69.9	82.5	92.4	86.4	88.8	96.3	82.8	96.1	98.1
	MSML	**76.7**	**88.9**	**95.6**	**72.0**	**83.4**	**93.3**	**89.6**	90.9	**97.4**	**87.5**	**97.7**	**98.9**

is a large scale video based person ReID dataset. Since all bounding boxes and tracklets are generated automatically, it contains distractors and each identity may have more than one tracklets. In total, MARS has 20,478 tracklets of 1,261 identities of 6 camera views.

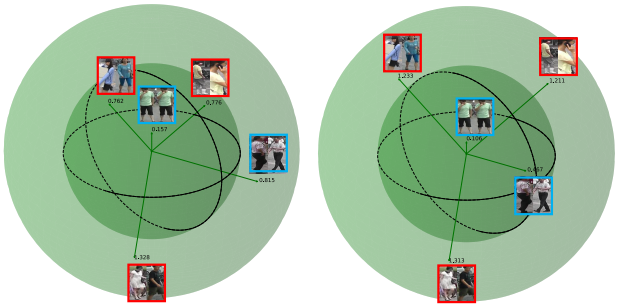
We evaluate our method with rank-1, 5, 10 accuracy and mean average precision (mAP), where the rank-*i* accuracy is the mean accuracy that images of the same identity appear in top-*i*. For each query, we calculate the average precision (AP). And the mean of the average precision (mAP) shows the performance in another dimension.

4.2. Implementation Details

Each image is resized into 224×224 pixels and conducted with data augmentaion. The augmentation includes randomly horizontal flipping, shifting, zooming and bluring. The base models (Resnet50, Inception-v2, Resnet50-Xception (Resnet50-X)) are pre-trained from ImageNet dataset. The final feature dimensions of Resnet50, Inception-v2 and Resnet50-X are transformed to 1024 through a fully-connected layer. The margin of triplet loss is set to $\alpha = 0.3$ and the margins of the quadruplet loss is set to $\alpha = 0.3$ and $\beta = 0.2$. The margin of triplet loss with hard mining and our loss with edge mining are also set to $\alpha = 0.3$. Adam optimizer is used and the inital learning rate is set to 10^{-3} in the first 50 epoches. Learning rate decreases to 10^{-4} in the next 150 epoches and 10^{-5} until convergence. And the batch size is set to 128.

We use Resnet50, Inception-v2 and Resnet50-X as base model respectively with different loss functions. There are several contrast experiments and the results are shown in Table 1.

4.3. Results analysis of Different Losses



(a) TriHard (b) MSML

Figure 3. Distance distributions of two different metric learning losses. Blue boxes are positive pairs while red boxes are negative pairs. Note that the direction arrows are only used for viewing.

We conduct experiments with different losses and provide the results to illustrate the effectiveness of our proposed MSML. They are shown in Table 1. Cls (classification loss) is the baseline experiment. Then, we combine different metric learning losses with classification loss. For Tri (triplet loss), the mAP and rank-1 accuracy increase by approximately 10.0% compared to baseline experiments. TriHard (triplet loss with hard sample mining) and Quad (quadruplet loss) both have better performance than triplet loss. TriHard is a little better on Market1501, MARS and CUHK-03 while Quad does better on CUHK-SYSU. Finally, our MSML gets best accuracy on most experiments datasets for all different base models.

In terms of accuracy, TriHard and MSML can both get high scores. We further visualize the distance distributions of some randomly chosen image pairs in Figure 3. The nu-

Table 2. Comparison on **Market1501** with single query

Methods	mAP	r=1
Temporal [26]	22.3	47.9
Learning [49]	35.7	61.0
Gated [38]	39.6	65.9
Person [5]	45.5	71.8
Pose [55]	56.0	79.3
Scalable [1]	68.8	82.2
Improving [20]	64.7	84.3
In [13]	69.1	84.9
Spindle[53]	-	76.9
Deep[52]*	68.8	87.7
Our	**76.7**	**88.9**

Table 3. Comparison on **MARS** with single query

Methods	mAP	r=1
Re-ranking [60]	68.5	73.9
Learning [50]*	-	55.5
Multi [37]*	-	68.2
Mars [35]	49.3	68.3
In [13]	67.7	79.8
Quality [24]*	51.7	73.7
See [61]	50.7	70.6
Our	**74.6**	**84.2**

meric values below the image pairs stand for the distances of their features in the embedding space. As we can see, the distances of negative pairs may be smaller than positive pairs, because TriHard does not focus on absolute distance. In contrast, our MSML can get a finer metric in feature embedding space.

For Quad and TriHard, some experiments were unable to reach its best accuracy with the same setting. And in Inception-v2 and Resnet-X experiments, they can be even worse than Tri. However, compared with them, our MSML can always have the best performance.

4.4. Comparison with state-of-the-art methods

We compare our method with representative ReID methods on several benchmark datasets (* means it is on ArXiv but not published). The results are shown in Table 2, 3, 4, 5. Methods which applied re-ranking[60] skills are not included.

5. Conclusion

In this paper, we propose a new metric learning loss with hard sample mining named MSML in person re-identification (ReID). For triplet and quadruplet loss, the positive pairs and negative pairs are randomly sampled. With hard sample mining, easy samples are filtered and thus improving the robustness of the model. In our method, we calculate a distance matrix and then choose the maximum distance of positive pairs and the minimum distance of negative pairs to calculate the final loss. In this way, MSML

Table 4. Comparison with existing methods on **CUHK03**

Methods	r=1	r=5	r=10
Person [19]	44.6	-	-
Learning [49]	62.6	90.0	94.8
Gated [38]	61.8	-	-
A [39]	57.3	80.1	88.3
In [13]	75.5	95.2	**99.2**
Joint [44]	77.5	-	-
Deep [10]*	84.1	-	-
Looking [2]*	72.4	95.2	95.8
Unlabeled [59]*	84.6	97.6	98.9
A [58]*	83.4	97.1	98.7
Spindle[53]	**88.5**	**97.8**	98.6
Our	87.5	97.7	98.9

Table 5. Comparison with existing methods on **CUHK-SYSU**

Methods	mAP	r=1
End[43]	55.7	62.7
Neural [21]*	77.9	81.2
Deep [34]*	74.0	76.7
Our	**89.6**	**90.9**

uses the most dissimilar positive pair and most similar negative pair to train the model.

We use Resnet50, Inception-v2 and Resnet50-X as base models to do some contrast experiments with different metric learning losses. The results show our MSML gets best performance and learns a finer metric in feature embedding space. Then, we compare our method with some state-of-the art methods. On several benchmark datasets, including Market1501, MARS, CUHK-SYSU and CUHK-03, our method shows better performance than most of other methods.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*, 2017.

[2] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *arXiv preprint arXiv:1701.03153*, 2017.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*, 2017.

[4] Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv preprint arXiv:1707.01220*, 2017.

[5] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016.

[8] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[10] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[11] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–6. IEEE, 2008.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[15] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. *arXiv preprint arXiv:1702.05729*, 2017.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. pages 152–159, 2014.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

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

[21] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. *arXiv preprint arXiv:1707.06777*, 2017.

[22] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017.

[23] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017.

[24] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. *arXiv preprint arXiv:1704.03373*, 2017.

[25] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[26] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, pages 858–877. Springer, 2016.

[27] T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2428–2433. IEEE, 2016.

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

[29] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[30] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20. Springer, 2016.

[31] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[34] A. Schumann, S. Gong, and T. Schuchert. Deep learning prototype domains for person re-identification. *arXiv preprint arXiv:1610.05047*, 2016.

[35] Springer. *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[37] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017.

[38] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[39] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153, 2016.

[40] J. Wang, S. Zhou, J. Wang, and Q. Hou. Deep ranking model by large adaptive margin learning for person re-identification. *arXiv preprint arXiv:1707.00409*, 2017.

[41] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[42] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

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

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017.

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[46] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016.

[47] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai. Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[48] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[49] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[50] W. Zhang, S. Hu, and K. Liu. Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*, 2017.

[51] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*, 2017.

[52] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv preprint arXiv:1706.00384*, 2017.

[53] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. CVPR, 2017.

[54] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):356–370, 2017.

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

[56] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference*, 2015.

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

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

[59] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.

[60] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017.

[61] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification.