Learning Cross Camera Invariant Features with CCSC Loss for Person Re-identification

Zhiwei Zhao, Bin Liu(✉), Weihai Li, and Nenghai Yu

Key Laboratory of Electromagnetic Space Information,
Chinese Academy of Sciences, School of Information Science and Technology,
University of Science and Technology of China, Hefei, China
zhaozhiwei1998@foxmail.com, {flowice,whli,ynh}@ustc.edu.cn

Abstract. Person re-identification (re-ID) is mainly deployed in the multi-camera surveillance scene, which means that learning cross camera invariant features is highly required. In this paper, we propose a novel loss named *Cross Camera Similarity Constraint loss* (CCSC loss), which makes full use of the camera ID information and the person ID information simultaneously to construct cross camera image pairs and performs cosine similarity constraint on them. The proposed CCSC loss effectively reduces the intra-class variance through forcing the whole network to extract cross camera invariant features, and it can be unified with identification loss in a multi-task manner. Extensive experiments implemented on the standard benchmark datasets including CUHK03, DukeMTMC-reid, Market-1501 and MSMT17 indicate that the proposed CCSC loss can bring a large performance boost on the strong baseline and it is also superior to other metric learning methods such as hard triplet loss and center loss. For instance, on the most challenging dataset CUHK03-Detect, Rank-1 accuracy and mAP are improved by **10.0%** and **10.2%** than the baseline respectively and simultaneously obtain a comparable performance with the state-of-the-art method.

Keywords: Person re-identification · CCSC loss · Cross camera

1 Introduction

Person re-identification (re-ID) has attracted close attention both in academic community and industry in recent years due to its great application prospects in many fields, such as video surveillance analysis, human-computer interaction, intelligent retail, etc. Given a query person-of-interest, person re-identification aims to retrieve all images that belong to the same person captured by multiple camera without view overlap at different time or scenarios.

Although the methods based on deep learning have brought great success to person re-ID [4,7,19,22], this field still faces many challenges. On the one hand, due to the extreme complexity of cross camera surveillance scenario, the image

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 429–441, 2019.
https://doi.org/10.1007/978-3-030-34120-6_35

pairs captured by different cameras for a specific person have dramatic variations in viewpoint, posture, background and illumination. As shown in Fig. 1, for this man, there is a large viewpoint variation between image pair captured by camera 2 and 4, an obvious posture change between image pair captured by camera 2 and 3, and a drastic change both in the background and illumination between image pair captured by camera 2 and 5. On the other hand, in a real large-scale surveillance scenario, the color of clothes and body shapes between different pedestrians may be very similar, which makes it difficult to distinguish even with the human eyes.

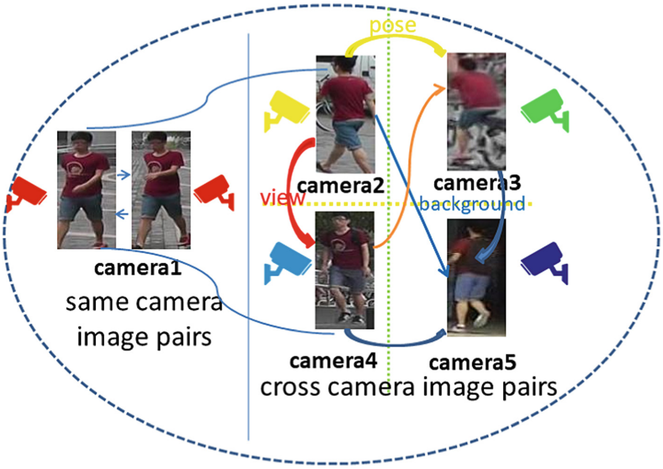

Fig. 1. In a multi-camera surveillance scenario, the image pairs of a pedestrian captured by different cameras have dramatic variations in viewpoint, posture, background and illumination, but the image pairs of a pedestrian captured by the same camera have very high similarity in appearance.

As mentioned above, many problems in person re-ID are caused by cross camera, so learn cross camera invariant feature representation is highly required. We notice that the few work in this field has utilized the camera ID information, which may play a very important role in the aspect of supervision. For instance, most existing re-ID methods based on deep metric learning such as constrastive loss [20] and triplet loss [3,15], which only consider the person ID information to construct positive and negative pairs but neglect useful camera ID information.

In this paper, we propose a novel loss named Cross Camera Similarity Constraint loss (CCSC loss), which takes full advantage of the camera ID information and person ID information to construct cross camera image pairs for every person and performs cosine similarity constraint on them. Specifically speaking, we first take each sample within a batch as the anchor, and then for each anchor, we select all the proper samples that have the same person ID but different camera

ID with the anchor to construct cross camera sample pair. Finally, we maximize the cosine similarity on all the cross camera sample pairs.

The proposed CCSC loss effectively alleviates a series of problems caused by cross camera, and it can be combined with identification loss in an multi-task learning framework. Compared with just using identification loss, the joint training of the CCSC loss and identification loss can bring significant performance improvements on the mainstream person re-ID benchmarks.

In summary, the contributions of this paper are two folds:

- We propose a novel loss named the CCSC loss, which explicitly utilizes the camera ID information and person ID information simultaneously to form cross camera sample pairs and performs similarity constraint on them. Through extensive experiments, we verify that the CCSC loss consistently improves the accuracy of the baseline over standard datasets, CUHK03, DukeMTMC- reid, Market-1501 and MSMT17.
- We fairly compare the proposed CCSC loss with hard triplet loss [7] and center loss [25]. Experiments show that when above losses are combined with identification loss respectively, the CCSC loss is not only superior to other losses in performance, but also makes it easier to train because it does not need to adjust additional hyperparameters. More importantly, it can achieve better performance with the help of hard triplet loss.

2 Related Work

With the tremendous success of deep learning in the field of computer vision, people abandoned hand-craft features for person re-identification [13,14], then deep learning based methods quickly dominate the person re-ID benchmarks. Recently deep re-ID methods mainly revolves around the following two lines.

One line is to find more discriminant and robust features to represent pedestrians, such as part-based methods [19,22,27], attention-based methods [12,17,21] and pose-guided methods [24,26,28]. Among them, the part-based methods achieve the state-of-the-art performance, which split a input feature map horizontally into a fixed number of strips and aggregate features from those strips. For example, Wang *et al.* [22] carefully design the multiple granularity network (MGN) to extract and utilize the global and local part features with multi-granularity for re-ID. However, MGN is very complex, and the computation cost of integrating all multi-branch feature vectors for testing is heavy, which is unrealistic for large-scale rapid person re-identification.

The other line is to find more effective metric learning method to make features of the same person more similar than those of different persons, such as constrastive loss [20], triplet loss [3,15], quadruplet loss [1] and center loss [25]. However, constrastive loss and triplet loss based methods suffer the common problem that they are prone to have a slow convergence speed and unstable performance in the circumstance of a large number of person identitiesnce and highly dependent on the sample's quality of the mini-batch in training. Although hard sample mining methods [7] effectively alleviates this problem, triplet loss still has extra margin hyperparameter to adjsut and it is not easy to train. Center loss, which simultaneously learns a center for deep features of each class and penalizes the euclidean distance between the deep features and their corresponding class centers. However, for center loss, each category center needs to be learned explicitly, and it does not fully take into account the rich variations of all the sample pairs. It is worth mentioning that none of the above metric learning methods have utilized the camera ID information. The proposed CCSC loss not only makes full use of the camera ID information to construct cross camera image pairs, but also takes full account of all the possible combinations of variations in appearance between image pairs. When above losses are all combined with identification loss respectively, compared with hard triplet loss and center loss, The proposed CCSC loss not only surpasses them in performance, but also make it easier to train.

3 Proposed Method

This section first introduces the structure of our model for person re-ID, and then explains the proposed CCSC loss in details.

3.1 Model Structure

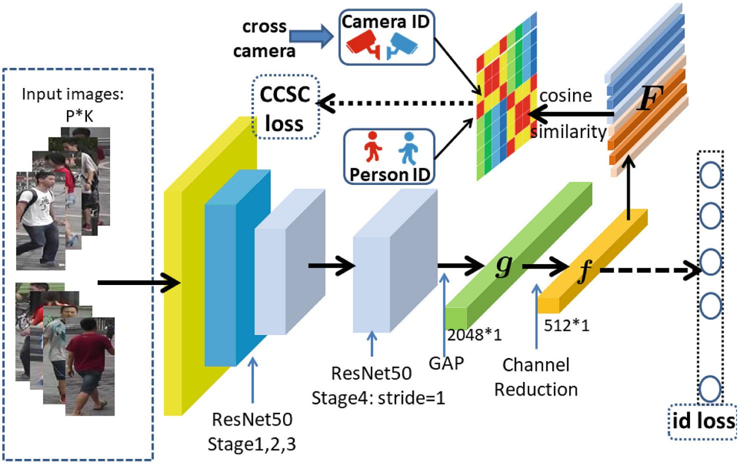

Fig. 2. The overall architecture of the proposed model. We adopt the modified ResNet-50 as the backbone for feature extraction, after that we use GAP (global average pooling) to get a 2048-dimensional feature vector $\boldsymbol{g}$, then we employ a 1×1 convolutional layer to get a 512-dimensional feature vector $\boldsymbol{f}$. Finally, the feature vectors of this batch $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N]$ are fed into two branches, one branch is used to calculate the proposed CCSC loss and the other branch is implemented by a classifier to calculate the identification loss. We combine the CCSC loss and identification loss in multi-task manner to train the entire network.

ResNet-50 Baseline. We use ResNet-50 [6], which is widely used in person re-identification, as backbone for feature extraction. Notice that we change the down-sampling stride of ResNet-50 stage4 from 2 to 1 to enlarge the the spatial size of the feature map, just like recent works [19,22] have done. After stage 4 of ResNet-50, we use GAP (global average pooling) to get a 2048-dimensional feature vector $\boldsymbol{g}$, then we employ a 1×1 convolutional layer, a batch normalization layer, and a ReLU layer to reduce the dimension of $\boldsymbol{g}$ to get a 512-dimensional feature vector $\boldsymbol{f}$. Finally, the dimension-reduced feature vector $\boldsymbol{f}$ is fed into the classifier. Classifier is implemented by a fully-connected layer and a softmax layer. We denote the structure described above as ResNet-50 baseline. For simplicity, we will use baseline to refer to ResNet-50 baseline in the rest of this paper. For baseline, we only use identification loss (softmax loss) to optimize the whole network. The baseline achieves 91.4% Rank-1 accuracy and 77.9% mAP on the Market-1501 dataset, we believe that the innovation on a stronger baseline can better demonstrates the effectiveness of our method.

Our Approach. As shown in Fig. 2, on the basis of the baseline, our method only adds an extra branch to compute the proposed CCSC loss. Specifically speaking, we first randomly sample P identities and K images of per person to constitute a training batch, thus the batch size $N = P \times K$. After that we feed a batch of images to network to extract the feature vectors of this batch $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N]$. Finally, $\boldsymbol{F}$ are fed into two branches, one branch is used to calculate the CCSC loss and the other branch is implemented by a classifier to calculate the identification loss. We combine these two types of losses in multi-task manner to train the entire network. Under this manner, our network not only has good property of distinguish different pedestrians, but also learns the cross camera invariant features, which greatly alleviates the problem of intra-class variation. In the test phase, the feature vector $\boldsymbol{f}$ is extracted for final distance metric.

3.2 Loss Function

The Proposed CCSC Loss. We follow the same batch sampling strategy with [7] to randomly sample P identities and K images of per person to constitute a training batch, thus in each mini-batch $\mathcal{N}$, we have $N = P \times K$ images. We denote the ith instance of pth person as s_i^p, denote the feature vector of s_i^p as $\boldsymbol{f}_i^p$. As Eq. (1) illustrates, we use indicator function $\boldsymbol{I}(s_i^p, s_j^p)$ to represent whether sample s_i^p and s_j^p come from different cameras, and camera(s_i^p) indicates the camera id of s_i^p.

Equation (2) shows the formulaic representation of the proposed CCSC loss, and diagram (Algorithm 1) shows a clear procedure to calculate the CCSC loss. Taking each sample $s \in \mathcal{N}$ as the anchor, we first select all the proper samples that have same person ID but different camera ID with anchor s, then we utilize them to construct cross camera sample pairs. At the same time, we compute the cosine similarity on all the cross camera sample pairs. For convenience of optimization, we transform the cosine similarity into the form of loss with the help of function: $T(x) = \frac{1}{1+x}$. Finally, we minimize the average loss on all the cross camera sample pairs.

$$\boldsymbol{I}(s_i^p, s_j^p) = \begin{cases} 1 & \text{if camera}(s_i^p) \neq \text{ camera}(s_j^p); \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$\mathcal{L}_{ccsc} = \frac{1}{M} \sum_{p=1}^{P} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\boldsymbol{I}(s_i^p, s_j^p)}{1 + \frac{(\boldsymbol{f}_i^p)^T \boldsymbol{f}_j^p}{\|\boldsymbol{f}_i^p\|_2 \|\boldsymbol{f}_j^p\|_2}} \tag{2}$$

$$M = \sum_{p=1}^{P} \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{I}(s_i^p, s_j^p) \tag{3}$$

Algorithm 1. The procedure of calculating the proposed CCSC loss.

```
Input:
    A batch of feature vectors: $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N]$, size: $N \times 512$;
    The corresponding person ID label vector: $\boldsymbol{p}$, size: $N \times 1$;
    The corresponding camera ID label vector: $\boldsymbol{c}$, size: $N \times 1$;
Output:
    The value of CCSC loss for this batch: $\mathcal{L}_{ccsc}$;
1: Calculating the cosine similarity matrix $\boldsymbol{S}$, $\boldsymbol{S}_{ij} = \frac{(\boldsymbol{f}_i)^T \boldsymbol{f}_j}{\|\boldsymbol{f}_i\|_2 \|\boldsymbol{f}_j\|_2}$, size:$N \times N$;
2: Converting similarity matrix $\boldsymbol{S}$ to loss matrix $\boldsymbol{D}$ with the help of function $T(\cdot)$
   mentioned above, thus $\boldsymbol{D} = T(\boldsymbol{S}) = \frac{1}{1+\boldsymbol{S}}$, element-wise operation;
3: Constructing the cross camera image pairs constraint flag matrix: $\boldsymbol{Mask}$;
       Expanding person ID label vector $\boldsymbol{p}$ to matrix $\boldsymbol{P} = [\boldsymbol{p}, \boldsymbol{p}, \cdots, \boldsymbol{p}]$, size: $N \times N$;
       Expanding camera ID label vector $\boldsymbol{c}$ to matrix $\boldsymbol{C} = [\boldsymbol{c}, \boldsymbol{c}, \cdots, \boldsymbol{c}]$, size: $N \times N$;
       Calculating person constraint flag matrix: Mask-P $\Leftarrow (\boldsymbol{P} == \boldsymbol{P}^T)$;
       Calculating camera constraint flag matrix: Mask-C $\Leftarrow (\boldsymbol{C} \neq \boldsymbol{C}^T)$;
       $\boldsymbol{Mask} \Leftarrow$ (Mask-P&Mask-C);
   Comment: Mask-P equals to 1 denotes that the corresponding image pair comes
   from the same person, Mask-C equals to 1 denotes that the corresponding image
   pair comes from different camera, so $\boldsymbol{Mask}$ equals to 1 is the flag of the cross
   camera image pair that we desire.
4: Calculating the average loss on all the selected cross camera sample pairs.
       $\mathcal{L}_{ccsc}$ = mean( $\boldsymbol{D}[\boldsymbol{Mask} == 1]$ );
5: return $\mathcal{L}_{ccsc}$;
```

Overall Loss. We regard the identification task as a multi-class classification problem, so identification loss actually refers to softmax loss in this paper. We combine identification loss with the proposed CCSC loss in multi-task manner like many works in this field [2,11] to train the entire network. The joint training of identification loss and the CCSC loss brings a significant improvement in performance. As illustrated in Eq. (4), the overall loss is the weighted sum of the proposed CCSC loss and identification loss, and λ is the balanced weight of the CCSC loss. In the experiments, we set $\lambda = 1.5$ for best performance.

$$\mathcal{L} = \mathcal{L}_{id} + \lambda \mathcal{L}_{ccsc} \tag{4}$$

4 Experiments

4.1 Datasets and Protocols

Datasets. We conduct extensive experiments on four public person re-identification benchmarks, *i.e.*, CUHK03 [10], DukeMTMC-ReID [16,31], Market-1501 [29] and MSMT17 [23]. Note that for CUHK03, according to whether it is manual or DPM labeling, it is divided into CUHK03-Label and CUHK03-Detect and we use the recently proposed new protocol in [32] for CUHK03.

Protocols. In the experiments, to evaluate the performances of re-ID methods, we report the cumulative matching characteristics (CMC) at Rank-1 and mean average precision (mAP) on four datasets.

4.2 Implementation Details

Training. In the training phase, the input images are resized to 384×128, then random horizontal flip, normalization and random erasing [33] are applied as data augmentation. We set $P = 16$ and $K = 4$ to construct mini-batch, thus batchsize $N = 64$. The backbone ResNet-50 is initialized from the ImageNet pre-trained model [5]. We use the Adam optimizer [9] to train the whole network. We set $\lambda = 1.5$ for best performance. We train 100 epochs in total. The learning rate warms up from 3.5e−5 to 3.5e−4 linearly in the first 5 epochs, then it is decayed to 3.5e−5 and 3.5e−6 at 35th and 55th epoch respectively. Our network is trained using 1 NVIDIA TITAN XP GPU and adopted Pytorch as the platform. Note that all comparative experiments adopted the same settings to ensure fairness.

Testing. In the testing phase, the input images are resized to 384×128, and only augmented with normalization. We extract feature vector $\boldsymbol{f}$ for test and use euclidean distance to rank.

4.3 Comparative Experiments Analysis

In order to prove the effectiveness of the proposed CCSC loss, a series of comparative experiments are conducted on all datasets we mentioned above.

Effectiveness of the Proposed CCSC Loss. As shown in Table 1, the proposed CCSC Loss can bring significant improvement to baseline on all four benchmarks. For instance, with the help of the proposed CCSC Loss, Rank-1 accuracy and mAP are improved by **10.0%** and **10.2%** respectively on the most challenging datasets CUHK03-Detect. It fully demonstrates the effectiveness of joint training with softmax loss and the CCSC loss. We can also see that the proposed CCSC loss is very effective for some datasets that have rich viewpoint variations, such as CUHK03, MSMT17 and DukeMTMC-ReID.

Figure 3 shows top-5 ranking results for some given query pedestrian images on CUHK03 and DukeMTMC-ReID dataset. For the first given pedestrian, the top-3 ranking results of the baseline are all mismatched due to the great similarity of clothing color and body shape, but the top-3 ranking results of our method for this pedestrian are all correct, even if there is a large change in angle of view. For the second query, even with serious occlusion problems and viewpoint changes, our methods can still find all the right results. The above results illustrate that our method is very effective for person re-ID in real complex scenes.

Table 1. Comparison of the proposed method (baseline + CCSC loss) with the baseline on four datasets. Rank-1 accuracy (%) and mAP (%) are shown.

Method	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
baseline	63.8	60.8	60.4	56.6	82.8	66.9	91.4	77.9	69.0	40.7
baseline+ CCSC loss (**ours**)	73.6	70.8	70.4	66.8	85.0	69.8	92.1	81.8	72.7	45.4
increment ↑	**+9.8**	**+10.0**	**+10.0**	**+10.2**	+2.2	+2.9	+0.7	+3.9	+3.7	+4.7

Importance of Cross Camera Constraint. In this part, we illustrate the importance of cross camera constraint. As illustrated in Table 2, if we don't apply cross camera constraint when construct image pairs, i.e., we always set $\boldsymbol{I}(s_i^p, s_j^p) = 1$, the re-ID performance significantly degraded on all datasets. As Fig. 1 shows, the image pairs come from the same camera have very high similarity, so we don't need to consider these simple sample pairs, because they are harmful to the optimization process of the whole network. The restriction of cross camera condition also reflects the idea of hard sample mining [7].

Table 2. The effect of whether to apply cross camera constraint or not.

Method	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
With constraint	73.6	70.8	70.4	66.8	85.0	69.8	92.1	81.8	72.7	45.4
Without constraint	70.7	68.1	67.1	64.2	83.3	69.2	91.4	80.4	71.3	43.1
Reduction ↓	**−2.9**	**−2.7**	−3.1	−2.6	−1.7	−0.6	−0.7	−1.4	−1.4	−2.3

Comparison of the Proposed CCSC Loss with Hard Triplet Loss and Center Loss. In order to further prove the superiority of the proposed CCSC loss, we fairly compare the CCSC loss with several commonly used metric learning losses on all benchmarks. Table 3 shows the performance comparison on all datasets when given different combinations of losses. We can see that the joint training with softmax loss and CCSC loss is the best, whereas the joint training with softmax loss and hard triplet loss is the second, and the joint training with softmax loss and center loss is the worst, but both are better than the baseline model which only use softmax loss. Meanwhile, we can see from Table 3 that our method can achieve better performance when combined with hard triplet loss. Futhermore, the proposed CCSC loss not only surpasses other losses in performance, but also make it easier to train, because it has very few parameters to adjust.

Table 3. Comparison of the proposed CCSC loss with the hard triplet loss and center loss when based on the same baseline.

Method	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Softmax + CCSC (**ours**)	**73.6**	70.8	70.4	66.8	85.0	69.8	92.1	81.8	72.7	45.4
Softmax + triplet	**69.4**	65.7	65.5	62.2	84.0	69.4	91.8	80.2	72.0	44.5
Softmax + center	**66.1**	62.2	62.3	59.0	83.4	69.0	91.3	78.9	72.5	44.0
Softmax only (**baseline**)	**63.8**	60.8	60.4	56.6	82.8	66.9	91.4	77.9	69.0	40.7
Softmax + CCSC + triplet	**74.2**	71.0	71.8	68.6	84.5	71.5	92.2	82.7	72.4	47.5

Comparison with State-of-the-Art Methods. We compare our proposed method with state-of-the-art methods on all candidate datasets in Table 4. It can be clearly see that although our method only utilizes global feature, it still achieves comparable performance with BFE [4], which achieves the strongest

Table 4. The comparison with state-or-the-art methods on CUHK03, DukeMTMC-reID, Market1501 and MSMT17 datasets.

Method	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE [30]	22.2	21.0	21.3	19.7	67.7	47.1	72.5	46.0	–	–
SVDNet [18]	–	–	41.5	37.3	76.7	56.8	82.3	62.1	–	–
AlignedReID [26]	–	–	–	–	81.2	67.4	90.6	77.7	–	–
HA-CNN [12]	–	–	–	–	80.5	63.8	91.2	75.7	–	–
SPReID [8]	–	–	–	–	84.4	71.0	92.5	81.3	–	–
PCB [19]	–	–	61.3	54.2	81.9	65.3	92.4	77.3	–	–
PCB + RPP [19]	–	–	62.8	56.7	83.3	69.2	93.8	81.6	–	–
MGN [22]	68.0	67.4	66.8	66.0	88.7	78.4	95.7	86.9	–	–
BFE [4]	**75.4**	**71.2**	74.4	70.8	86.8	71.5	93.5	82.8	–	–
baseline	63.8	60.8	60.4	56.6	82.8	66.9	91.4	77.9	69.0	40.7
baseline+CCSC (ours)	**73.6**	**70.8**	70.4	66.8	85.0	69.8	92.1	81.8	72.7	45.4

performance on CUHK03 dataset at present. On Market1501 and DukeMTMC-reID dataset, although our approach is not as good as MGN in performance, the model complexity and computational cost are much lower than MGN, which is more suitable for large-scale rapid person re-identification.

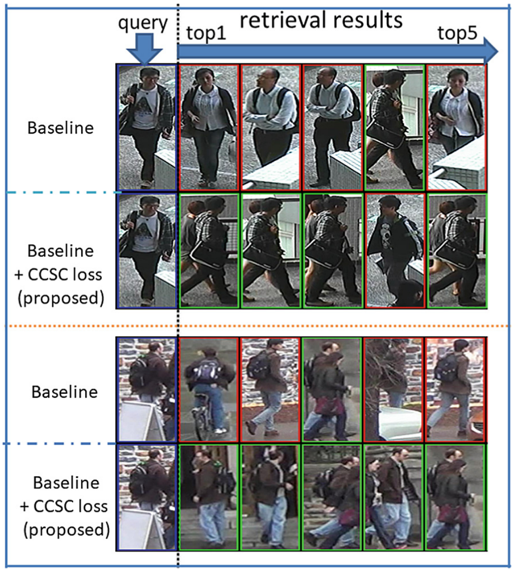

Fig. 3. Comparison of top-5 ranking results between our proposed method (baseline + CCSC loss) and baseline. The images with green borders belong to the same identity as the given query, and that with red borders do not. The images with blue borders represent query, best viewed in color. (Color figure online)

Visualization of the Feature Response Map. We believe that the proposed method did learns the cross camera invariant features. Figure 4 shows some feature response maps for some input pedestrian images, extracted from the last feature map before GAP. The brighter the area is, the more concentrated it is. We can clearly see that for query and gallery, which have a large change of view, our network is more concerned about some body areas that are keep unchanged cross cameras, such as the collar and sleeves of jacket, thighs, shoes and so on, which remain visible during the change of view.

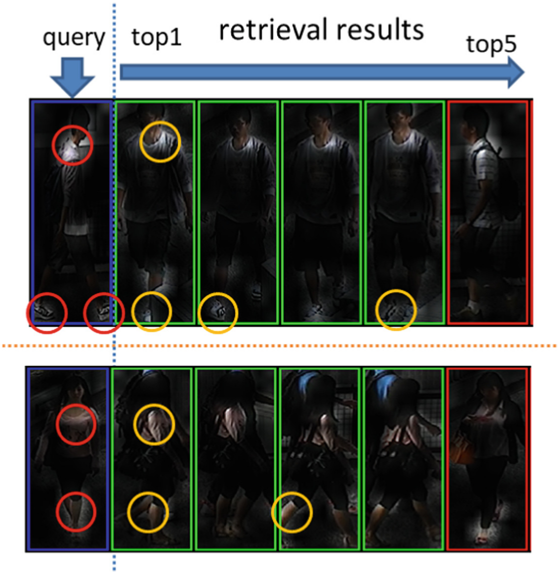

Fig. 4. Visualization is done by overlapping the intensity of corresponding last feature map onto the original images. The brighter the area is, the more concentrated it is, best viewed in color. The red circle and the yellow circle respectively point out the corresponding camera invariant regions for query and gallery. (Color figure online)

5 Conclusion

In this paper, we propose a novel loss named the CCSC loss to learn the cross camera invariant features for person re-identification. The proposed CCSC loss simultaneously utilizes the camera ID information and person ID information to construct cross camera sample pairs and performs cosine similarity constraint on them. The CCSC loss largely boost the performance of re-ID through the joint training with identification loss, and it is also superior than other metric learning losses in performance. Extensive experiments implemented on the standard benchmark datasets confirm the effectiveness of the proposed CCSC loss.

References

1. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 403–412 (2017)
2. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
3. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)

4. Dai, Z., Chen, M., Zhu, S., Tan, P.: Batch feature erasing for person re-identification and beyond. arXiv preprint arXiv:1811.07130 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
8. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. 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)
11. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724 (2017)
12. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
13. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3610–3617 (2013)
14. 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)
15. 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)
16. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
17. Si, J., et al.: Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5372 (2018)
18. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
19. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 501–518. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_30
20. 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

21. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 384–400. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_23
22. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 274–282. ACM (2018)
23. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
24. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 420–428. ACM (2017)
25. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
26. Zhang, X., et al.: Alignedreid: surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
27. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)
28. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. IEEE Trans. Image Process. (2019)
29. 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)
30. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv preprint arXiv:1610.02984 (2016)
31. 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, pp. 3754–3762 (2017)
32. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)
33. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)