High-Confidence Sample Labelling for Unsupervised Person Re-identification

Lei Wang[1], Qingjie Zhao[1](✉), Shihao Wang[2], Jialin Lu[1], and Ying Zhao[3]

[1] Beijing Institute of Technology, Beijing 100081, China
zhaoqj@bit.edu.cn
[2] The Australian National University, Canberra ACT, Canberra 2600, Australia
[3] The University of Hong Kong, Hong Kong 999077, China

Abstract. Person re-identification (re-ID) is factually a topic of pedestrian retrieval across camera scenes. However, it is challenging due to those factors such as complex equipment modeling, light change and occlusion. Much of the previous research is based on supervised methods that require labeling large amounts of data, which is expensive and time-consuming. The unsupervised re-ID methods without manual annotation usually need to construct pseudo-labels through clustering. However, the pseudo-labels noise may seriously affect the model's performance. To deal with this issue, in this paper, we use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to assign pseudo-labels to samples and propose a model with the high-confidence samples' labels (HCSL), which is a fully unsupervised learning method and does not use any labeled data. The model constructs high-confidence triplets through cyclic consistency and random image transformation, which reduces noise and makes the model finely distinguish the differences between classes. Experimental results show that the performance of our method on both Market-1501 and DukeMTMC-reID performs better than the latest unsupervised re-ID methods and even surpasses some unsupervised domain adaptation methods.

Keywords: Re-identification · Unsupervised learning · Deep clustering · Pseudo-labels

1 Introduction

Person re-identification (re-ID) is a crucial task to retrieve the same person's identity across various devices. The challenge is how to alleviate the influence of different cameras, various postures, occlusion, and pedestrians' wear. In recent years, re-ID has been widely used in video surveillance systems and intelligent security, and has become the focus of academic research. Although deep learning approaches [32,42] exhibit superior performance, they typically rely on manually annotated datasets to train the model. Unsupervised re-ID approaches can avoid laborious data annotation with highly generalized models and they are more suitable for video surveillance and other cases. Therefore, it is more concerned by people at present.

F. Sun et al. (Eds.): ICCSIP 2021, CCIS 1515, pp. 61–75, 2022.
https://doi.org/10.1007/978-981-16-9247-5_5

Recently, unsupervised re-ID approaches has made good progress. Existing approaches mainly include cross-domain unsupervised re-ID and single-domain unsupervised re-ID. The cross-domain [14,23] unsupervised approaches usually need a manually annotated source dataset. They use the generative adversarial networks (GAN) to transfer the source domain's image style to fit the target domain's style. However, due to differences in background, equipment, pedestrian wear, and postures between different datasets, the target domain features may not be sufficiently distinguishable from the model pre-trained on the source domain dataset [39] as shown in Fig. 1. Performance of the cross-domain unsupervised models is still lagging behind supervised learning. In addition, it is challenging to select the appropriate source domain data for transfer learning in unsupervised re-ID because of domain differences [22,35]. The single-domain methods belong to fully unsupervised re-ID and do not require any manual labeled data. Their traditional methods [8,15,19] focus on hand-made features. However, the performance of these methods is lower than that of supervised methods. To relieve these problems, we choose to use self-supervised learning.

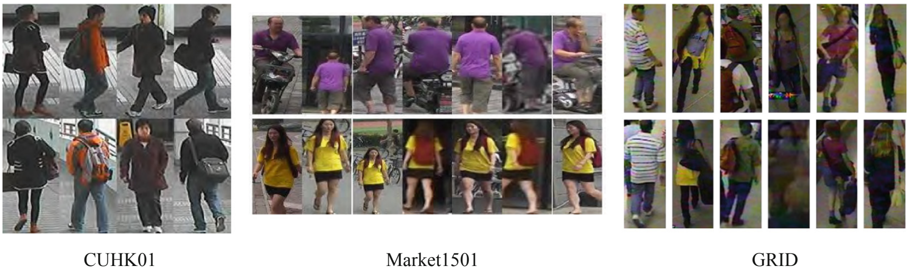


Fig. 1. Differences in background, equipment, pedestrian wear, and postures between different datasets

The self-supervised approach [5,12] can be regarded as a particular unsupervised learning method. Its supervised information is self-mined from the unsupervised dataset, then the network is trained through this information. Deep clustering [2,25,41] is a self-supervised learning approach. It combines convolutional neural networks (CNN) [13] and unsupervised clustering to propose an end-to-end model. In re-ID, a fully unsupervised approach, Bottom-up Clustering (BUC) [17], is based on deep clustering. BUC uses bottom-up hierarchical clustering to merge samples, and after each step of merging, it uses the result clusters as pseudo-labels for deep neural network training. It then uses the trained network to get features and update clustering and the pseudo-labels continuously until the model achieves the best performance. However, BUC may not distinguish between complex samples in early model merging, leading to wrong merging and getting many wrong pseudo-labels. Simultaneously, these errors cause superimposition errors in subsequent merging, thereby severely degrade the model's performance.

To address these issues and reduce pseudo-labels noise impact in re-ID tasks that do not use any labeled data, we propose a method that trains the network with the high-confidence samples labels (HCSL). Moreover, HCSL is also a deep clustering method that does not require any manual labels. The iterative process of HCSL includes (1) training the network to extract features, (2) clustering pseudo-labels, (3) training classification tasks and updating the network's weights, (4) reusing the model to extract features. Specifically, HCSL assigns a pseudo-label to each sample and extracts image features through a pre-trained feature extractor. Then the model clusters the samples through Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [6] and updates the pseudo-labels. The last and most significant thing is that model fine-tunes the network using high-confidence triplet loss (HCTL) and saves the best performing model after several iterations.

In summary, the main contributions of this paper are: This paper aims to improve the accuracy of unsupervised re-ID without using any labeled data. To reduce pseudo-labels noise impact in fully unsupervised re-ID, the model uses high-confidence triplet loss to optimize the model. At the same time, the loss can balance noise-free pseudo-label samples and hard negative sample mining. Experimental results show that our performance on Market-1501 and DukeMTMC-reID is better than that of the latest unsupervised re-ID methods.

2 Related Work

2.1 Supervised Person Re-identification

Early research on re-ID focused on extracting robust and discriminative low-level visual features, such as color features [10,29], shape features [44], and texture features [3]. And traditional machine learning methods are used for metric learning [24] in the process of feature matching. Because re-ID faces severe challenges such as scenes, pedestrian postures, and occlusion, the above-mentioned traditional methods are difficult to achieve good results. Momentarily, deep learning has been introduced in re-ID, and significant progress has been made. Furthermore, in re-ID, deep learning is mainly used to extract more discriminative feature representations. By early 2021, the best performance on the re-ID general dataset Market-1501 reached Rank-1 = 96.2%, mAP = 91.7%, and reached Rank-1 = 91.6% and mAP = 84.5% on DukeMTMC-reID [47]. However, these supervised methods usually rely on labeled datasets to train the model. When the trained model is applied to other datasets, the performance is significantly reduced, and it may not be practical to label each new scene. Therefore, unsupervised re-ID will become a new research hot spot.

2.2 Unsupervised Person Re-identification

In recent years, cross-domain person re-ID has achieved encouraging results, and most studies use the style transfer theory. Zhong et al. [46] effectively utilize camera invariance in domain adaptation, and they use starGAN to generate a series of different camera-style images. Liu et al. [20] perform style transfer on ambient light, resolution, and camera field of the view separately and then integrate them. Deng et al. [4] use CycleGAN to convert the source domain's image style to that of the target domain without changing the image labels. And then, it trains the network on the generated images. Also, some cross-domain methods use the thought of clustering. Fu et al. [9] train the model on the source domain and then segment the target domain images, they cluster the patch and whole target domain images respectively to obtain pseudo-labels. However, these methods require the source dataset labels, and the style widely differs between the source domain and the target domain datasets. The generalization ability of unsupervised domain adaptation is insufficient.

The self-supervised re-id methods usually use pseudo-labels generated by clustering for deep learning. Fan et al. [7] use source domain data to train the network, use Kmeans to cluster target domain samples to generate pseudo labels, and use pseudo labels to fine-tune the model. However, it is not a fully unsupervised method. Lin et al. propose BUC [17], using CNN to extract image features. Then BUC stipulates hierarchical clustering to merge a fixed number of classes at each step and uses the pseudo-labels generated in each step as supervision to fine-tune the model. Although this method achieves confident performance, this simple combination cannot solve pseudo-labels noise, and it is difficult for the deep networks to propose more discriminative features.

To solve the noise issue caused by pseudo-labels, Wang et al. [33] propose to use multi-label instead of single-label classification, they consider similarity and consistency of style to construct the feature bank, and then determine the soft multi-label. Lin et al. [18] propose a method that does not require clustering, but it uses a classification network with softened labels to reflect the similarity between images. Because these methods for constructing soft labels do not have hard-labels learning, errors caused by the hard classification are eliminated. Although some improvements have been made in these approaches, most approaches require domain adaptation or other auxiliary information to help estimate the similarity. After removing these aids that require much calculation, their performance is still not satisfying [18]. However, HCSL reduces the noise generated by pseudo-labels using a more efficient loss function and obtains better performance than previous methods.

3 Proposed Method

This paper proposes a self-supervised method with high-confidence samples' labels (HCSL). This model combines the deep learning network and unsupervised clustering, and it is optimized in an end-to-end manner.

3.1 HCSL Architecture

Figure 2 shows the overall framework of HCSL. This model mainly consists of three steps: pseudo-labels initialization, unsupervised clustering, and fine-tuning.

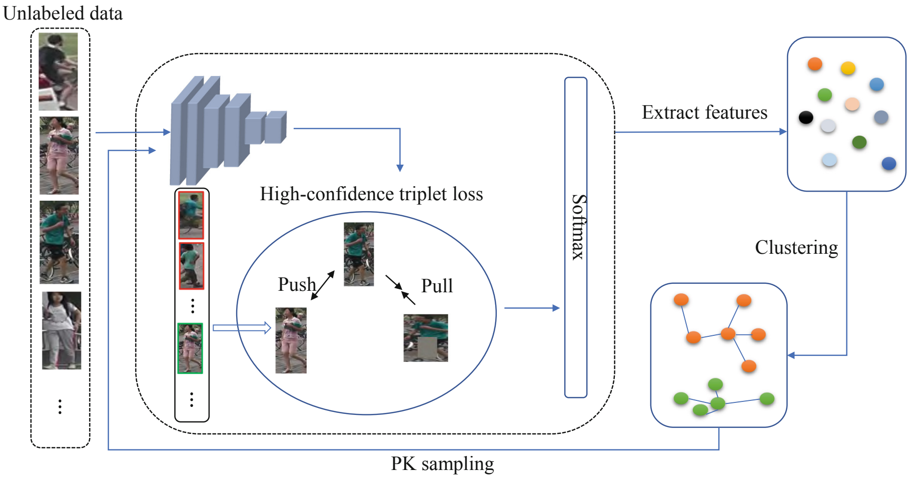


Fig. 2. The structure of HCSL. The framework does three steps alternatively: extract features, unsupervised clustering, and fine-tuning with high-confidence triplet loss.

Specifically, in the first stage, the model extracts image features through a pre-trained feature extractor and assigns pseudo-labels to each sample; in the second stage, the model uses unsupervised clustering methods: DBSCAN to cluster samples and assign new pseudo-labels; in the third stage, HCSL uses PK sampling [37] to obtain mini-batch from the dataset and fine-tune the network. Furthermore, we develop a high-confidence triplet loss (HCTL) to minimize the pseudo-labels noise caused by unsupervised clustering. HCTL, through cycle consistency, selects suitable *negative simples* to construct triples to calculate the loss. Compared with other methods, it can further improve the identification ability.

3.2 Iterative Pseudo Labeling

Inspired by Tang et al. [31], HCSL uses DBSCAN to cluster the initialized dataset. The advantage of DBSCAN is that there is no need to set the number of clustering. In addition, its clustering speed is fast, and it can effectively deal with noise points and find spatial clusters of arbitrary shapes.

Pseudo-Labels Initialization. To use the loss function to optimize the models, models need to generate pseudo-labels as supervision. For a dataset containing N image samples $X = \{x_1, x_2, \cdots, x_N\}$, HCSL treats each sample as a separate cluster to obtain pseudo-labels $Y = \{y_1, y_2, \cdots, y_N\}$. And then, HCSL uses ResNet [11] pre-trained by ImageNet as the extractor's backbone network and replaces the original fully connected (FC) layer with the new FC-1024 layer to output the feature vector. Through this feature extractor, we can get the feature embedding of each image for subsequent clustering.

DBSCAN. To obtain clustering results closer to the ground truth, HCSL uses DBSCAN to cluster the images' feature vectors after initializing the pseudo-labels. In HCSL, we use Euclidean distance to measure the similarity between samples, so DBSCAN is based on the Euclidean distance matrix *dist*. DBSCAN is different from the Kmeans [38] clustering because it does not need to give the number of clusters before the algorithm runs [6]. The DBSCAN generates k density-based clusters and assigns pseudo labels to each sample. However, the pseudo-labels obtained by unsupervised clustering have serious noise, we will get ineffective performance if we use them directly to train the network.

3.3 Loss Function

Triple loss [30] is a commonly used loss function in person re-ID model training. However, it is sensitive to abnormal samples. In previous deep clustering methods, the performance using triple loss was poor due to the huge error of pseudo-labels generated by unsupervised clustering. Therefore, based on the triple loss, we get inspiration from a common way of constructing supervision signals in self-supervised learning: cyclic consistency. And we propose another loss HCTL that can reduce the influence of pseudo-labels noise. Because the core thought of triple loss is to develop a triple of *anchor*, *positive sample* and *negative sample*, then it shortens the distance between *positive sample* and *anchor*, at the same time pushes *negative sample* away. In unsupervised learning, the choice of *positive* and *negative* samples has a decisive influence. The core thought of HCTL is to construct triples with more high-confidence samples through the random transformation of images and cyclic consistency. Thereby, the influence of abnormal samples is reduced to improve model performance. To meet the need of HCTL loss, we use PK sampling to generate mini-batches for training in each iteration. PK sampling means that we randomly select K instances from P identities to generate mini-batches according to clustering results, so it is easily to combine triples required by HCTL.

Positive Samples. Obviously, the image x_i^* generated after a series of random transformations of image x_i must be the *positive sample* of x_i. Therefore, when the HCSL performs PK sampling on the dataset X^* to construct a mini-batch, it stores a randomly transformed *positive sample* for each image.

Negative Samples. Self-supervised learning usually uses the cyclic consistency principle to construct self-supervised signals. Therefore, HCTL mines *negative sample* in a mini-batch for *anchor* images by cycle consistency. As shown in Fig. 3, the HCTL can search the image bidirectionally according to the cyclic consistency to mine *negative sample* for *anchor*. We set a non-negative threshold q and a pseudo-labelled mini-batch $\{x_i\}_{i=1}^{P \times K}$. This mini-batch selects P identities from all identities and selects K samples from each identity. For each sample x_i, HCTL calculates the Mahalanobis distance between x_i and other $(P-1) \times K$ samples of other clusters. Mahalanobis distance is different from Euclidean distance. It can consider the relationship between various attributes

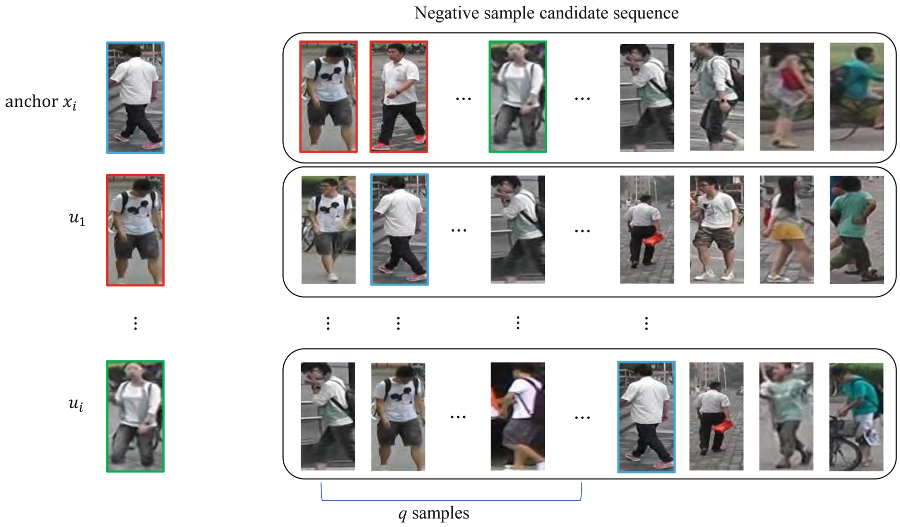


Fig. 3. Mining *negative sample*. We calculate the negative sample candidate sequence U of the *anchor*, and then calculate the negative sample candidate sequence of each sample in U starting from u_1. When the order of *anchor* in the sequence is greater than q, we find the *negative sample* of anchor x_i and stop the algorithm.

and pay more attention to the correlation between samples, while Euclidean distance treats the differences between multiple attributes equally. So Mahalanobis distance learning is a prominent and widely used approach for improving classification results by exploiting the structure of the data [28]. Given n data points $x_i \in R_m$, the goal is to estimate a matrix M such that:

$$d_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j)$$

Where R_m is a batch of the dataset with pseudo-labels. According to the distance, we sort them to obtain a negative sample candidate sequence $\{u_i\}_{i=1}^{(P-1)\times K}$ from small to large.The further back in the sequence, the more likely it is to be a negative sample, but more likely to be a simple negative sample. To balance noise-free pseudo-labels and hard negative samples mining, HCTL will not simply select the last sample of the sequence as the negative sample. For each sample u_i in the sequence, HCTL also calculates its *negative sample* candidate sequence. If x_i does not appear in the first q positions, then u_i is a *negative sample* of x_i. At the same time, we specify that the first u_i found is a hard *negative sample* of x_i.

HCTL can be expressed as:

$$L = \left\{ \|x_a - x_p\|_2 - \|x_a - x_n\|_2 + margin, 0 \right\}_+$$

Where x_a is the *anchor*, x_p is the *positive sample* generated after image transformation, and x_n is the *negative sample* mined according to the cycle consistency. The loss is calculated by the high-confidence triples composed of x_a, x_p, and x_n, which further reduces the influence of pseudo-labels noise, helps the model to shorten the distance within the class and pushes the distance between classes. While considering the reduction of sample pseudo-label noise, we also consider the importance of hard negative samples for model training. Thus the model performance is improved.

Algorithm 1. HCSL Algorithm

```
Require:
    unlabeled data $X = \{x_1, x_2, \cdots, x_N\}$
    non-negative sample threshold $q$
    iteration $t$
    similarity $s$
Ensure:
    best model $f(w, x_i)$
initialize:
    iteration $iter = 0$
    pseudo-labels: $\{y_i\}_i^N = 1$
train model with $X$ and $Y$, the model without high-confidence triplet loss
while $iter < t$ do
    initialize pseudo-labels:$\{y_i\}_i^N = 1$
    extract features
    calculate the Euclidean distance matrix $dist$
    calculate the minimum of the $dist$
    clustering with k-means clustering: $c$
    update $Y$ with new pseudo-labels
    fine-tune model with $X$ and $Y$, the model with high-confidence triplet loss
    evaluate model performance:$P$
    if $P > P_{best}$ then
        $P_{best} = P$
        best model:$f(w, x_i)$
    end if
    $iter = iter + 1$
end while
```

3.4 Model Updating

As shown in Algorithm 1, at the beginning of each iteration, HCSL assigns N image samples to N different clusters to obtain the initial pseudo-labels. The initialization, clustering, and fine-tuning processes make up an iteration. In every iteration, HCSL generates k high-quality clustering centers through DBSCAN clustering and allocates N image samples to k clusters to update the dataset X with new pseudo-labels. Then HCSL fine-tunes the model with the dataset

X according to HCTL. We iterate over the model and evaluate its performance until it stops improving.

4 Experiments

4.1 Datasets

Market-1501. Market-1501 [43] is collected in the campus of Tsinghua University. It consists of 32,668 images of 1,501 people and is shot by six cameras. The training set includes 12,936 images of 751 people, 17.2 training data per person on average. Moreover, the test set includes 19,732 images of 750 people, 26.3 test data per person on average.

DukeMTMC-reID. Dukemtmc-reID [45] is a person re-identification subset of the DukeMTMC dataset, which contains 36,411 images of 1,404 people and is shot by 8 cameras. The training set includes 16,522 images of 702 people, and the test set contains 17,661 images of 702 people.

4.2 Training Details

HCSL Training Setting. We use the ResNet-50 pre-trained by ImageNet as the backbone network and replace the original FC layer with a new FC-1024 layer to output the feature vectors. An image size of the model input is adjusted to 224 × 224. The *batchsize* is 64, and a mini-batch is generated by selecting $P = 16$ identities and $K = 4$ images randomly. In the model initialization phase, we use SGD [1] to optimize the model, and the momentum parameter is 0.9, *weight decay* is 5e−4. We train the model with *learning rate* 0.1 for 20 epochs. In the fine-tuning model stage, we use RAdam [21] to optimize the model, and the *learning rate* is 0.01 for 20 epochs, *weight decay* is 5e−4. Moreover, HCTL *margin* is 1, and the non-negative sample threshold q is 14 on Market-1501. The *positive sample* is obtained after random cropping, random flipping, and random erasure of the input image.

HCSL Evaluating Setting. We use the mean average precision (mAP) and the Cumulative Matching Characteristic (CMC) curve to evaluate the model performance. The mAP reflects the model's recall rate, and the CMC curve reflects the model's retrieval accuracy. We use Rank-1, Rank-5, and Rank-10 scores to represent the CMC curve.

4.3 Effectiveness of HCSL

Table 1 shows the performance comparison between the HCSL and the most advanced methods on the Market-1501 and DukeMTMC-reID. Our method achieved the best performance with Rank-1 = 73.6% and mAP = 51.3% on the Market-1501. Compared with our unsupervised baseline, BUC method, our model's accuracy on Rank-1 is improved by 7.4% and mAP by 13%. And our

Table 1. The performance comparison between the HCSL and several most advanced methods on the Market-1501 dataset and DukeMTMC-reID dataset. "None" means that these methods do not use labeled labels, and "Transfer" means that these methods need to be trained on the source domain and then applied to the target domain. "Weakly" means weakly supervised method.

Methods	Labels	Reference	Market-1501				DukeMTMC-reID			
			mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
BOW [43]	None	ICCV15	14.9	35.6	52.3	60.1	8.3	17.0	28.5	34.7
UMDL [26]	Transfer	CVPR16	12.3	34.5	52.4	59.7	7.2	18.7	31.5	37.4
PUL [7]	Transfer	TOMM18	20.3	44.9	59.4	65.7	16.3	30.1	46.2	50.7
SPGAN [4]	Transfer	CVPR18	26.6	57.9	75.8	81.4	26.4	46.7	62.3	68.4
HHL [46]	Transfer	ECCV18	31.7	62.3	78.4	84.6	27.4	46.7	61.1	66.7
TJ-AIDL [34]	Transfer	CVPR18	26.4	58.4	74.7	81.3	23.2	44.7	59.6	65.1
BUC [17]	None	AAAI19	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4
ATNet [20]	Transfer	CVPR19	25.7	55.9	73.7	79.8	25.2	45.3	59.9	64.8
UCDA [27]	Transfer	ICCV19	34.5	64.3	–	–	36.7	55.4	–	–
CSCL [36]	Transfer	ICCV19	35.6	64.7	80.2	85.6	30.5	51.5	66.7	71.1
WFDR [40]	Weakly	CVPR20	50.1	72.1	80.5	–	42.4	62.0	**75.1**	–
SSL [16]	None	CVPR20	37.8	71.1	83.8	87.4	28.6	52.5	63.5	68.9
HCSL (Ours)	None	This work	**51.3**	**73.6**	**87.5**	**91.2**	**47.9**	**62.7**	70.2	**75.7**

method also has superior performance with Rank-1 = 62.7%, mAP = 47.9% on DukeMTMC-reID. The improvement of HCSL performance is mainly due to high-confidence triplet loss (HCTL) that can distinguish image details better. Furthermore, it not only surpassing previous fully unsupervised methods but even surpasses some unsupervised domain adaptation (UDA) methods.

Table 2. The impacts of using high-confidence triplet loss (HCTL) on model performance.

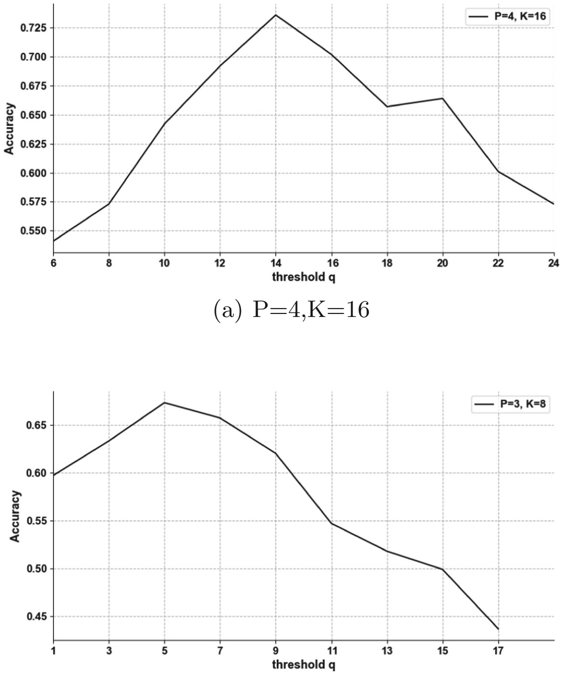
Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
HCSL (with triplet loss)	34.5	19.2	21.3	17.1
HCSL (with HCTL)	**73.6**	**51.3**	**62.7**	**47.9**

Comparison with Triple Loss. Table 2 shows the model's performance using high-confidence triple loss (HCTL) and using triple loss. Using HCTL has a remarkable performance improvement on both benchmarks. In Market-1501, using the high-confidence triple loss, compared with using the triple loss directly, the Rank-1 improves 39.1%, and the mAP improves 32.1%. Compared with BUC that does not use triple loss, triple loss makes the model performance worse because it is sensitive to abnormal samples. In unsupervised re-ID, the pseudo-labels noise generated by clustering can seriously affect triple loss calculation, but HCTL can achieve better results by reducing the impact. In addition, we also compared the performance of Mahalanobis distance and Euclidean distance when applied to HCTL. From Table 3, the Mahalanobis distance has a better effect. The experimental results prove our analysis in 3.3.

Table 3. Compare Euclidean distance and Mahalanobis distance. "*" means that HCSL uses Euclidean distance to obtain a negative sample candidate sequence

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
HCSL*	64.1	50.2	57.6	47.1
HCSL	**73.6**	**51.3**	**62.7**	**47.9**

Comparison with Different q in HCTL. In HCTL, the non-negative sample threshold q controls the selection of hard negative samples, and finally affects the confidence of triplet. To get the best performance, we set $P = 4$, $K = 16$ and evaluate the impact of different q on Market-1501. Our results are reported in Fig. 4(a). When we set $q = 14$, we get the best performance of HCSL. We believe



(a) P=4,K=16

(b) P=3,K=8

Fig. 4. Performance curve with different values of the non-negative sample threshold parameter q on Market-1501.

that in HCTL, the threshold q is affected by K in PK sampling. Although the re-ID dataset may not be ideal for calculating the negative sample candidate sequence due to pedestrian clothing and lighting factors, the difference between the classes still exists and can be HCTL obtained. At the same time, because of the limitations of unsupervised learning, HCSL does not work well at $q = 17$. To confirm the above conclusion, we set $P = 3$, $K = 8$ and evaluate the impact of different q on Market-1501. It can be seen in Fig. 4(b) that the performance when $q = 5$ is better than other cases, but the overall performance is worse than batchsize $= 64$.

5 Conclusion

In this paper, we propose a fully unsupervised re-ID method, HCSL. Different from previous works, this method does not require any labeled datasets. HCSL optimizes the following issue: in the previous deep clustering methods, the large amount of noise in the clustering pseudo-labels affects the model performance. Specifically, HCSL constructs high-confidence triplets through cyclic consistency and random image transformation, which reduces noise and makes the model finely distinguish differences between classes. With the model iteration, the pseudo-labels quality generated by DBSCAN is gradually improved, and the model performance is also steadily enhanced. The experiments prove that HCSL is not only surpassing previous fully unsupervised methods but even surpasses some unsupervised domain adaptation methods.

References

1. Bottou, L.: Stochastic gradient descent tricks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 421–436. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_25
2. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (2018)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2005)
4. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 994–1003 (2018)
5. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1422–1430. IEEE Computer Society, December 2015
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. AAAI Press, Palo Alto (1996)
7. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. ACM Trans. Multim. Comput. Commun. Appl. 83:1–83:18 (2018)

8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
9. Fu, Y., et al.: Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6111–6120 (2019)
10. Gou, M., Fei, X., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Computer Vision-ECCV 2014 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
12. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (ICLR), Vancouver, Canada, April 2018
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, pp. 1097–1105. NIPS 2012, Curran Associates Inc., Red Hook, NY, USA (2012)
14. Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
15. Liao, S., Yang, H., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
16. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8738–8745 (2019)
18. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3387–3396 (2020)
19. Lisanti, G., Masi, I., Bagdanov, A.D., Bimbo, A.D.: Person re-identification by iterative re-weighted sparse ranking. IEEE Trans. Patt. Anal. Mach. Intell. **37**, 1629–1642 (2015)
20. Liu, J., Zha, Z.J., Chen, D., Hong, R., Wang, M.: Adaptive transfer network for cross-domain person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
21. Liu, L., et al.: On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations (2020)
22. Long, M., Wang, J.: Learning transferable features with deep adaptation networks. JMLR.org (2015)
23. Lu, Y., et al.: Cross-modality person re-identification with shared-specific feature transfer. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Martinel, N., Micheloni, C., Foresti, G.L.: Saliency weighted features for person re-identification. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 191–208. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_14

25. Niu, C., Zhang, J., Wang, G., Liang, J.: GATCluster: self-supervised gaussian-attention network for image clustering. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12370, pp. 735–751. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58595-2_44
26. Peng, P., et al.: Unsupervised cross-dataset transfer learning for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1306–1315 (2016)
27. Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., Gao, Y.: A novel unsupervised camera-aware domain adaptation framework for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, 27 October–2 November 2019, Seoul, Korea (South), pp. 8079–8088. IEEE (2019)
28. Roth, P.M., Hirzer, M., Kstinger, M., Beleznai, C., Bischof, H.: Mahalanobis distance learning for person re-identification. Person Re-Identification (2014)
29. Rui, Z., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (2013)
30. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
31. Tang, H., Zhao, Y., Lu, H.: Unsupervised person re-identification with iterative self-supervised domain adaptation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
32. Tay, C.P., Roy, S., Yap, K.H.: AANet: attribute attention network for person re-identifications. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
33. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10978–10987 (2020)
34. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2275–2284 (2018)
35. Wei, L., Zhang, S., Wen, G., Qi, T.: Person transfer GAN to bridge domain gap for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
36. Wu, A., Zheng, W.S., Lai, J.H.: Unsupervised person re-identification by camera-aware similarity consistency learning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6921–6930 (2019)
37. Wu, C.Y., Manmatha, R., Smola, A.J., Krhenbühl, P.: Sampling matters in deep embedding learning. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
38. Wu, J., Xiong, H., Chen, J.: Adapting the right measures for k-means clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 877–886. Association for Computing Machinery (2009)
39. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
40. Yu, H.X., Zheng, W.S.: Weakly supervised discriminative feature learning with state information for person identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

41. Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C.: Online deep clustering for unsupervised representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
42. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4908–4917, June 2019
43. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124 (2015)
44. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Softw. Eng. **35**, 653–668 (2012)
45. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
46. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV), September 2018
47. Zhu, Z., Jiang, X., Zheng, F., Guo, X., Zheng, W.: Viewpoint-aware loss with angular regularization for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13114–13121 (2020)