Learning to Purification for Unsupervised Person Re-Identification

Long Lan, *Member, IEEE*, Xiao Teng, Jing Zhang, *Senior Member, IEEE*, Xiang Zhang, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Unsupervised person re-identification is a challenging and promising task in computer vision. Nowadays unsupervised person re-identification methods have achieved great progress by training with pseudo labels. However, how to purify feature and label noise is less explicitly studied in the unsupervised manner. To purify the feature, we take into account two types of additional features from different local views to enrich the feature representation. The proposed multi-view features are carefully integrated into our cluster contrast learning to leverage more discriminative cues that the global feature easily ignored and biased. To purify the label noise, we propose to take advantage of the knowledge of teacher model in an offline scheme. Specifically, we first train a teacher model from noisy pseudo labels, and then use the teacher model to guide the learning of our student model. In our setting, the student model could converge fast with the supervision of the teacher model thus reduce the interference of noisy labels as the teacher model greatly suffered. After carefully handling the noise and bias in the feature learning, our purification modules are proven to be very effective for unsupervised person re-identification. Extensive experiments on two popular person re-identification datasets demonstrate the superiority of our method. Especially, our approach achieves a state-of-the-art accuracy 85.8% @mAP and 94.5% @Rank-1 on the challenging Market-1501 benchmark with ResNet-50 under the fully unsupervised setting. Code has been available at: https://github.com/tengxiao14/Purification_ReID.**

***Index Terms*— Clustering purification, knowledge distillation, unsupervised person ReID.**

I. Introduction

PERSON re-identification (ReID) aims to retrieve the same person under different camera views. It has attracted widespread attentions in the computer vision community due to its great potential in real world applications [1]. Although great performance has been achieved in the supervised person ReID setting, the demand of human annotation heavily limits the application. To make it more scalable in the real world, the task of unsupervised person ReID has been raised and attracted increasing more attention as it requires no human annotation.

Unsupervised person ReID mainly includes two categories, unsupervised domain adaptation (UDA) person ReID and purely unsupervised learning (USL) person ReID [2]. The UDA methods aim to learn from the annotated source dataset and transfer the knowledge to the unlabeled target dataset [3], [4], [5]. They usually adopt the two-stage training strategy. At first the model is pre-trained on the labeled source dataset, then the unlabeled target dataset is utilized to finetune the model. Unlike existing general unsupervised domain adaptation setting where the source domain and target domain share the same label space [6], [7], [8], UDA person ReID usually assumes there are no interactions between source and target domains in the label space, thus compared with existing unsupervised domain adaptation setting [9], [10], [11], UDA person ReID is more challenging. Similar to other general unsupervised domain adaptation methods, UDA ReID methods are also proposed based on the assumption that the discrepancy between source domain and target domain is not significant, and the performance of these methods will drop significantly when the gap between source and target domains is large.

To further relax the dependency on labeled source dataset, the USL methods directly learn from the unlabeled target dataset, which require no annotation information from other domains [12], [13], [14], [15]. Thus, compared with the UDA person ReID, the USL person ReID is more scalable. Nowadays state-of-the-art USL person ReID methods have achieved great progress by training the model with the pseudo labels generated by clustering algorithm [2], [16], [17]. These methods hold the assumption that the images of the same person share higher similarity in the feature space, thus will be more likely to be collected in the same cluster. Generally, these methods can be regarded as two-stage training schemes, firstly a clustering algorithm is applied to divide the features of images into different clusters, and assign pseudo labels to different clusters accordingly. Then the model is trained with generated pseudo labels. These two stages are conducted in an iterative scheme as they can promote each other in the whole training process. Based on this training framework, memory-based contrastive learning methods have achieved the state-of-the-art performance nowadays by taking advantage of

Manuscript received 22 June 2022; revised 21 February 2023 and 6 April 2023; accepted 13 May 2023. Date of publication 26 May 2023; date of current version 15 June 2023. This work was supported by the National Grand Research and Development Plan under Grant 2020AAA0103501. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jun Liu. *(Long Lan and Xiao Teng contributed equally to this work.) (Corresponding author: Xiao Teng.)*

Long Lan, Xiao Teng, and Xiang Zhang are with the Institute for Quantum Information, and the State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China (e-mail: long.lan@nudt.edu.cn; tengxiao14@nudt.edu.cn).

Jing Zhang is with the School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: jing.zhang1@sydney.edu.au).

Dacheng Tao is with the JD Explore Academy, Beijing 101111, China (e-mail: dacheng.tao@gmail.com).

Digital Object Identifier 10.1109/TIP.2023.3278860

contrastive learning with image features stored in the memory bank [2], [16], [17].

Although the above ReID methods have achieved great progress in recent years, the gap between supervised and unsupervised methods is still large. After carefully analysed the reason behind the phenomenon, we think the learning process of unsupervised person ReID is mainly influenced by the feature bias and label noise due to limited global feature representation power and lack of accurate predicted labels. As the aim of global feature learning is to capture the most salient clues of appearance to represent identities of different pedestrians, some non-salient but important detailed local cues can be easily ignored due to the limited scales and less diversities of the training dataset, which makes global features hard to distinguish from similar inter-class persons [18]. As a result, images with different identities but similar salient clues of appearance could be easily merged to the same cluster, which will make the learned feature representation biased. On the other hand, since the model is trained with pseudo labels generated by clustering algorithm, it will suffer from severe label noise during the whole convergence process as it is initialized with parameters pre-trained on ImageNet dataset, which has the significant discrepancy with person ReID datasets. To relieve the above problems, we propose the feature and label noise purification modules for unsupervised person ReID, as shown in Fig. 1.

Specifically, our method mainly includes two modules, the feature purification (FP) module and label noise purification (LP) module. The former takes into account the features from two local views to enrich the feature representation and purify the inherent feature bias of the global feature involved. As the global feature tends to capture the most salient clues of appearance while neglecting some non-salient but detailed local cues, we also introduce feature representations from two complementary local parts (upper and lower parts of the person). In this way, these feature representations can provide information from different views (i.e., global view and partial views) [19]. Meanwhile, the latter aims to purify the label noise by taking advantage of the knowledge of teacher model in an offline scheme. As the noise will be inevitably introduced in the clustering process, the model will suffer from the label noise in the whole training process. Based on the phenomenon that the trained model is more accurate than the initialized model, intuitively the knowledge of the trained model can be utilized as the guidance to help the student model relieve the influence of noise in the training process. Our contributions can be concluded in the following:

- We propose a feature purification module which carefully integrate the multi-view features in our cluster contrast learning framework and is proven to be effective in handling the bias of the global feature easily involved.
- We further propose a label noise purification module which aims to relieve the label noise introduced by clustering procedure by taking advantage of the knowledge of teacher model. To our knowledge, this is the first work to apply offline knowledge distillation for unsupervised person ReID, and we find it is very effective for such task.
- Extensive experiments are conducted on two popular person ReID benchmarks. The results show our method significantly outperforms existing state-of-the-art unsupervised person ReID methods. Specially, our method outperforms the state-of-the-art method [17] by 3.2% and 6.2% in terms of mAP on Market-1501 and MSMT17 datasets.

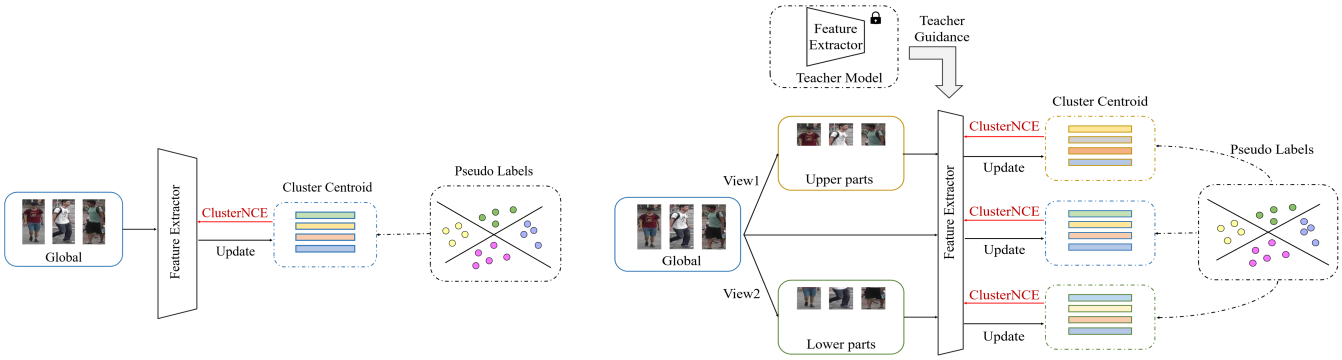
II. Related Work

In this section, we introduce the most related work from three perspectives: 1) Unsupervised person ReID, which includes unsupervised domain adaptation (UDA) person ReID and purely unsupervised learning (USL) person ReID; 2) Part-based person ReID, which takes advantage of local parts of the person to get more discriminative feature representations; and 3) Knowledge distillation, which includes some techniques of knowledge distillation in different areas.

A. Unsupervised Person ReID

Unsupervised person ReID can be summarized into two categories, unsupervised domain adaptation (UDA) person ReID and purely unsupervised learning (USL) person ReID. The former aims to learn from the annotated source dataset and transfer the knowledge from the source domain to the unlabeled target domain [2], [3], [4], [5], [20], [21]. While the latter directly trains on the unlabeled target dataset without any labeled data [2], [16], [17]. To make full use of unlabeled target dataset, unsupervised person ReID methods usually apply existing clustering algorithms, such as Kmeans [22] and DBSCAN [23] to generate pseudo labels for each sample in the target domain. Then the generated pseudo labels and the unlabeled dataset are used together to train the model in an iterative scheme [16], [17]. To further improve the quality of pseudo labels, many variants of pseudo label generation methods have been proposed. BUC [12] proposed a bottom-up clustering framework by exploiting the intrinsic diversity among identities and similarity within each identity to learn more discriminative feature representations. GLT [24] proposed a Group-aware Label Transfer algorithm that facilitates the online interaction and mutual promotion of the pseudo labels prediction and feature learning. To avoid the label noise accumulation in a single model setting, MMT [5] refines the noisy pseudo labels by optimizing two neural networks under the joint supervisions of off-line refined hard pseudo labels and on-line refined soft pseudo labels.

The above methods can be applied to both UDA ReID and USL ReID. However, different from USL ReID, UDA ReID also has the auxiliary labeled source dataset. Thus the key of UDA ReID methods is how to take advantage of labeled source dataset to improve the performance of the model on unlabeled target dataset. These methods usually work based on the assumption that the discrepancy between the source domain and the target domain is not significant and apply transfer learning techniques to tackle such problem. To further mitigate the gap between source and target domains, some domain-translation-based methods are proposed, which aim to take advantage of generative adversarial networks



(a) Cluster contrast learning (b) Cluster contrast learning with purification modules

Fig. 1. Comparison of cluster contrast learning with/without purification modules. (a) Cluster contrast learning trains the model solely with cluster centers of global feature representations. (b) Besides the global level cluster contrast learning, feature and label noise purification modules are also applied. The former takes into account the features from two local views to purify the bias of the global feature involved. Meanwhile, the latter aims to purify the label noise by taking advantage of the knowledge of teacher model in an offline scheme.

(GANs) [25] to translate the source-domain images to have target-domain styles while preserving their original IDs to augment the dataset [26], [27], [28]. In addition, IDM [29] and IDM++ [30] propose to align the source and target domains against their intermediate domains so as to facilitate a smooth knowledge transfer.

In this work we focus on USL ReID and our work is established on memory-based contrastive learning frameworks, which are the state-of-the-arts for unsupervised person ReID. SPCL [16] proposed a self paced method which gradually create more reliable clusters to refine the hybrid memory and learning targets. To solve the problem of inconsistency in the memory updating process, CCL [17] proposed a novel cluster contrast learning framework which was built on a cluster-level cluster memory dictionary and achieved great performance. Recently, HDCPD [31] uses a unified local-to-global dynamic learning and self-supervised probability regression framework that leverages all clustered and unclustered instances. MCL [32] introduces a new paradigm for unsupervised person ReID, where a subset of the entire unlabeled data is pseudo-labeled through clustering, and the learned cluster centroids are used as a proxy annotator to softly annotate the remaining unlabeled data. ISE [33] generates support samples from actual samples and neighboring clusters in the embedding space through a progressive linear interpolation strategy to reveal underlying information for accurate cluster representation. In [34], a self-guided hard negative generation method is proposed by adversarially training a hard negative generation network and a re-ID network to improve each other. HCM [35] proposes a framework for identity-level and image-level contrastive learning to explore feature similarities among hard sample pairs for unsupervised person ReID. In [36], a group sampling strategy is proposed to address the over-fitting problem in unsupervised person ReID by alleviating the negative impact of individual samples on statistical stability and exploiting the potential of the contrastive baseline.

In recent years, various selection or refinement modules have been proposed to enhance the performance of unsupervised person ReID. For example, Dual-Refinement [37] adopts a hierarchical clustering scheme and an instant memory spread-out regularization to jointly refine pseudo labels and features. PPLR [38] refines pseudo labels by leveraging the complementary relationship between global and local features. CACL [39] proposes a cluster refinement module to remove noisy samples in larger clusters. RLCC [40] introduces a pseudo label refinement module by utilizing temporally propagated and ensembled pseudo labels. Although these refinement methods have shown effectiveness in unsupervised person ReID, they all rely on the model itself to refine the pseudo labels or features. Hence, in the early stages of training, when the model is not yet trained on the target dataset, it may suffer from severe label noise due to the significant discrepancy between the pre-trained parameters and the target dataset. Although some methods utilize peer model or EMA updated teacher model [5], [31], they still suffer from this problem as their peer model or teacher model are also initialized with ImageNet pretrained parameters. To tackle this challenge, we propose to utilize a well-trained teacher model to guide the student model in the fully unsupervised person ReID setting. While utilizing global and local cluster centers as prototypes in the training process to purify the bias of the global feature, we also purify the label noise for these branches by learning from the teacher model from global and local views, and we find it is effective for unsupervised person ReID task. Additionally, while some refinement methods have been proposed for person ReID with label noise, such as PurifyNet [41] and CORE [42], they require noisy labels annotated by human experts at the beginning and degenerate when the noise ratio increases, making them not scalable for unsupervised person ReID.

In this work, to purify the feature and label noise for unsupervised person ReID, we also take into account local views and the knowledge of the teacher model. Thus we also discuss some works related to these techniques in the below.

B. Part-Based Person ReID

Most deep learning-based person ReID approaches take advantage of only the global feature of the person, which turns out to be sensitive to the missing key parts. To relieve the issue, recently many works focused on leveraging part discriminative feature representations. These works aim to make use of local

parts to make more accurate retrieval. Part-based person ReID can be divided into three categories. In the first category, the prior knowledge like poses or body landmarks are required to be estimated to locate the accurate parts of the person. However, the performance of such approaches heavily rely on the accuracy of the pose or landmarks estimation models [43], [44]. The second category utilized the attention mechanism to adaptively locate the high activation in the feature map. But the selected regions lack of semantic interpretation [45]. The third category directly utilizes the predefined strips as it assumes the person is vertically aligned. Compared with the first category it is more scalable as it requires no extra pre-trained models, thus it is widely used in the person ReID and achieved great improvements in recent years.

Specifically, MGN [18], PCB [46], SSG [47] and PPLR [38] also combine local features and global features to improve ReID models. Different from our method, PCB [46] and MGN [18] are proposed for supervised person ReID, with the given labels, these methods could learn more discriminative feature representation easily. Our work is more similar to SSG [47] and PPLR [38], which are also proposed for unsupervised person ReID and utilizes local and global features in both pseudo labels generation and model optimization processes. However, SSG [47] generates separated pseudo labels for each set of global and local feature branches, then trains the model with triplet loss and instance-level features, which may neglect high-level semantic and consistent meanings for different branches. PPLR [38] aims to refine pseudo labels by exploiting the complementary relationship between global and local features, then cross-entropy loss and triplet loss are combined to optimize the parameters of the model. However, pseudo labels generated by global and local features in their method are unreliable in the early period, which may mislead the model. Compared with these methods, global and local branches in our method share the same pseudo labels, and we utilize global and local cluster centers as prototypes to optimize the model to obtain more semantic and consistent feature representations for these branches. In addition, we further relieve the label noise for these branches by learning from the teacher model from both global and local views.

C. Knowledge Distillation

The aim of knowledge distillation is to transfer the knowledge from the network to another. The original idea of knowledge distillation is to compress the knowledge from the teacher network to a smaller student network. Recently, more works have focused on self-knowledge distillation, which keeps the structure of the teacher and student network the same [10], [48], [49]. These methods usually directly use the outputs of the teacher whose structure is the same as the student. Specifically, a simple but effective baseline was proposed for few shot learning in [50] by minimizing the loss where the target is the distribution of class probabilities induced by the teacher model. CS-KD [48] proposed a new regularization technique, which matches the distribution predicted between different samples of the same class. SSD [49] proposed a effective multi-stage training scheme for long-tailed recognition, which utilized the output of the teacher to generate soft label for the student. Recently, a collaborative ensemble learning scheme is proposed in [51] to utilize the relationship among different classifiers for cross-modality person ReID. It aims to enhances the discriminability with the ensemble outputs and their consistency.

Similar to our method, the knowledge distillation is also applied in [31] and [42] to relieve label noise. Specifically, CORE [42] optimizes networks and label predictions collaboratively by distilling the knowledge from other peer networks, and limited and inaccurate annotations are required in their work. HDCPD [31] aims to align the probability distribution between the network and the teacher network updated by Exponential Moving Average (EMA) method. Although these methods have achieved great improvements for relieving label noise, they may suffer from severe label noise in the early period of the training stage as the teacher model and the student model are both initialized with ImageNet-pretrained parameters. Different from these methods, we aim to resort to the well-trained teacher model for help in the unsupervised person ReID setting. As the trained teacher model is more accurate than the initialized student model, the teacher model can guide the student model to relieve the label noise in the early period of the training phase, and we find it is effective for unsupervised person ReID task.

III. Method

A. Cluster Contrast Learning Framework

Let $X = \{x_1, x_2, \ldots, x_N\}$ denotes the unlabeled training set which contains N instances. $F = \{f_1, f_2, \ldots, f_N\}$ denotes the corresponding feature maps extracted from the training set with the encoder f_θ, which can be described as $f_i = f_\theta(x_i)$. $U = \{u_1, u_2, \ldots, u_N\}$ denotes the feature vectors got from the feature maps after the pooling operation. u_q is the corresponding feature vector of the query instance q extracted with encoder f_θ. $\Phi = \{\phi_1, \phi_2, \ldots, \phi_C\}$ denotes C cluster representations in the training. Note that the number of the cluster C can vary according to clustering results.

Memory-based cluster contrast learning frameworks have achieved the state-of-the-art performance by taking advantage of memory mechanism and contrastive learning [16], [17]. Specifically, these methods utilize Kmeans [22] or DBSCAN [23] to generate pseudo labels for unlabeled samples. Thus a pseudo labeled dataset $X' = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_{N'})\}$ can be obtained, where $y_i \in \{1, \ldots, C\}$ is the pseudo label generated for the $i-th$ sample and N' is the number of the labeled samples in the dataset, note that N' is usually smaller than N as some sparse samples which are far from their neighbors are regarded as outliers in the clustering process. Then contrastive learning and memory mechanism can be applied on the pseudo labeled dataset. Among existing memory-based cluster contrast learning frameworks, cluster contrast learning [17] has achieved impressive performance by implementing contrastive learning on the cluster-level cluster memory dictionaries as following:

$$L = -\log \frac{\exp\left(u_q \cdot \phi_+ / \tau\right)}{\sum_{k=0}^{C} \exp\left(u_q \cdot \phi_k / \tau\right)}, \tag{1}$$

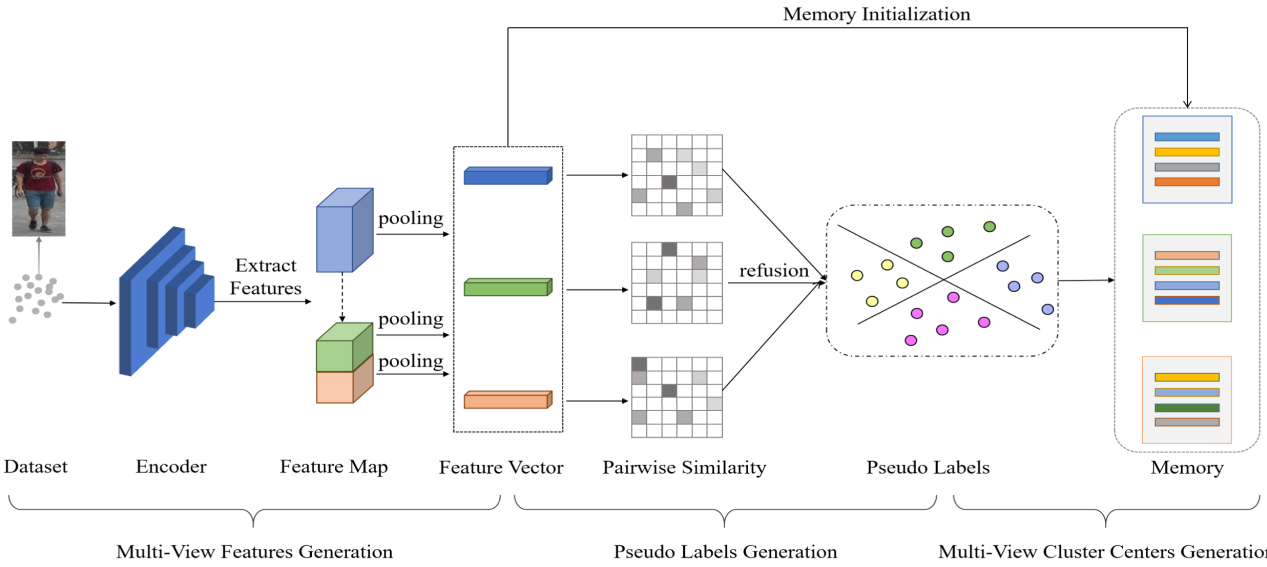


Fig. 2. Feature purification module. As the global feature tends to capture the most salient clues of appearance to represent identities of different pedestrians, some non-salient but important detailed local cues, however, may be easily ignored due to the limited scale and diversity of the training dataset. Therefore, besides the global feature, centers of the local features are also maintained as independent cluster memory dictionaries to enhance the feature representation and purify the inherent bias of the global feature involved. Specifically, given the unlabeled data, the model is firstly utilized to generate multi-view features. Then the DBSCAN clustering algorithm is applied on the fused pairwise similarity matrix to generate the consistent pseudo labels. Finally, the shared pseudo labels are utilized to guide the local/global features to initialize their respective memory cluster representations.

where u_q is the feature vector of the query sample. ϕ_k is the centroid feature vector representing the $k-th$ cluster stored in the memory, which is initialized by the average feature vector of samples in the $k-th$ cluster and ϕ_+ is the centroid feature vector of the cluster the query sample belongs to. C is the number of the cluster. τ is the temperature hyper-parameter. Then the centroid feature vector stored in the memory dictionary sets can be updated in the following way:

$$\phi_k = m\phi_k + (1-m)u_q, \tag{2}$$

where m is the momentum updating factor and k is the index of the cluster query sample belongs to.

Our unsupervised person re-identification is implemented in the framework of the cluster contrast learning. To achieve the goal of purification in the unsupervised person re-identification. We design two functional components, namely FP module and LP module, as shown in Fig. 2 and Fig. 4, respectively. The former takes into account the features from two local views to enrich the feature representation and purify the inherent feature bias of the global feature confront. Meanwhile, the latter aims to purify the label noise by taking advantage of the knowledge of teacher model in an offline scheme.

B. Feature Purification Module

Although most works only utilize the global feature map for the unsupervised person ReID [3], [4], [5], [17], the inherent feature bias of global feature may hinder the learning of the model for differing different persons as they tend to capture the most salient clues of the appearance while ignoring some detailed local cues. From this view, we propose the feature purification module, which aims to take advantage of extra local views to enhance the feature learning process by encouraging the model to discover more discriminative

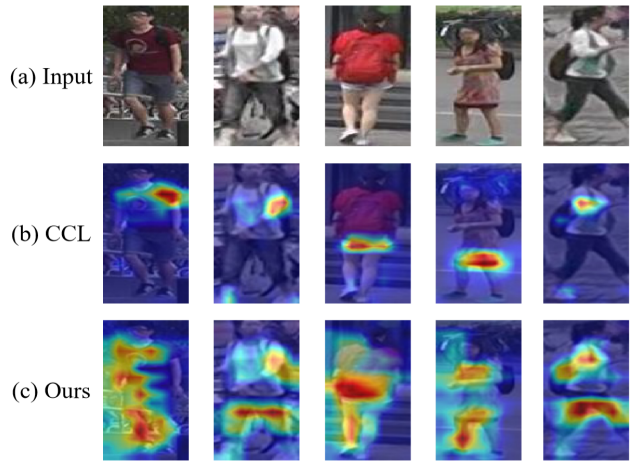


Fig. 3. The class activation maps (CAMs) [52] for some samples from Market-1501 extracted by CCL [17] and our method. The maps highlight the discriminative image regions used for retrieval. (a) Some original images from Market-1501. (b) CAMs extracted by CCL. (c) CAMs extracted by our method.

local cues in the feature representation. As shown in Fig. 3, CCL tends to pay attention to the most salient part while our method tends to capture more detailed local cues. The reason is probably that in CCL only global feature vectors are involved in the training process, which is limited for encoding the characteristics of the person. While our method forces the model to learn more detailed local cues by introducing two local views. The procedure of the feature purification module is shown in Fig. 2. To clearly describe the proposed FP module, we divide this module into three sub-modules as follows.

1) Multi-View Features Generation Process: Given an unlabeled training set $X = \{x_1, x_2, \ldots, x_N\}$, where N is the number of the samples in the dataset. We can get the corresponding feature maps $F = \{f_1, f_2, \ldots, f_N\}$ with the encoder f_θ. Then we split feature maps in F into two parts horizontally,

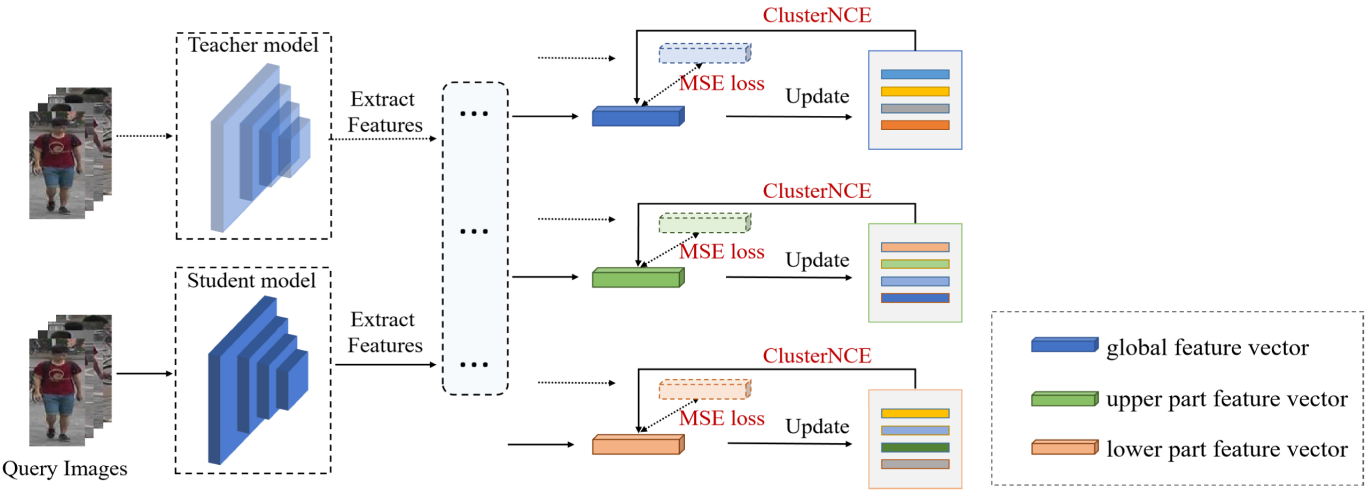


Fig. 4. Label noise purification module. We fixed the teacher model to learn the student model. The ClusterNCE loss and L2 loss are applied to update the student model.

which are denoted as $F^{up} = \{f_1^{up}, f_2^{up}, \ldots, f_N^{up}\}$ and $F^{dw} = \{f_1^{dw}, f_2^{dw}, \ldots, f_N^{dw}\}$ respectively. To get the feature vectors from them, Generalized-Mean (GEM) pooling operations [53] are applied on these feature branches independently. As a result, we can get three sets of feature vectors respectively.

$$\begin{cases} U^{gb} = \{u_1^{gb}, u_2^{gb}, \ldots, u_N^{gb}\} \\ U^{up} = \{u_1^{up}, u_2^{up}, \ldots, u_N^{up}\} \\ U^{dw} = \{u_1^{dw}, u_2^{dw}, \ldots, u_N^{dw}\}, \end{cases} \tag{3}$$

where U^{gb}, U^{up} and U^{dw} are three sets of feature vectors respectively. Compared with the global feature representations, feature vectors from these local views can introduce more detailed and complementary information about the person. Note that feature maps of introduced two local views are directly split from the global feature maps, thus the generation of these local feature vectors bring no extra computation burden to the model.

2) Pseudo Labels Generation Process: After getting the three sets of feature vectors in equation 3, following SPCL [16] and CCL [17], we also apply DBSCAN [23] clustering algorithm on these feature vectors to generate pseudo labels. DBSCAN [23] is an efficient clustering algorithm which can discover clusters of arbitrary shape. Nowadays, DBSCAN clustering algorithm has been widely used in recent unsupervised person ReID methods. These methods have proven that DBSCAN is more suitable for generating pseudo labels for person ReID datasets compared with other clustering algorithms, and thus we also adopt DBDCAN clustering algorithm in our work. Compared with Kmeans clustering algorithm, DBSCAN doesn't require the number of clusters in advance, thus it is more applicable for unsupervised person ReID tasks. Unlike these works which only utilize global features, we aim to generate pseudo labels by taking advantage of both global and local features. Our motivation is that as global features tend to capture the most salient cues, some non-salient but important detailed local cues can be easily ignored due to limited scales and less diversities of the training dataset. Thus images with different identities but similar appearance could be easily merged to the same cluster if we only utilize global features in the pseudo labels generation process. Specifically, with global and local feature vector sets U^{gb}, U^{up} and U^{dw}, the Jaccard distance matrix of the dataset can be calculated independently, which are denoted as D^{gb}, D^{up} and D^{dw}. Then a re-weighted pairwise distance matrix can be achieved using the following function:

$$D = (1 - 2\lambda_1)\, D^{gb} + \lambda_1 D^{up} + \lambda_1 D^{dw}, \tag{4}$$

where D is the re-weighted pairwise distance matrix, λ_1 is the balancing factor. Following many existing unsupervised person ReID methods [17], [31], [32], [33], [47], [54], we also calculate k-reciprocal Jaccard Distance [55] to generate unique distance matrices for each of global and local branches, where k is set to 30, then these matrices are re-weighted to form the final distance matrix. Similar to the re-ranking [55] technique which calculates the distance as the weighted aggregation of the original distance and the Jaccard distance, we calculate the distance as the weighted aggregation of global and local Jaccard distance to relieve the inherent bias of global feature representations. Then the pseudo labels $\tilde{Y}$ can be generated by DBSCAN clustering algorithm with matrix D. In this way, we can get a pseudo labeled dataset $X' = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_{N'})\}$, where N' is the number of the pseudo labeled dataset. Note that N' is smaller than the number of samples in the original dataset N due to the existence of outliers in the clustering process.

3) Multi-View Cluster Centers Generation Process: Following CCL [17], we also implement contrastive learning on the cluster-level memory dictionaries to avoid the problem of inconsistency in the memory updating process. Specifically, as the pseudo labeled dataset is obtained, then cluster centroids in the memory are initialized by the corresponding mean feature vectors and the pseudo labels as following:

$$\phi_k = \frac{1}{|C_k|} \sum_{i \in C_k} u_i, \tag{5}$$

where C_k denotes the $k-th$ cluster $|\cdot|$ denotes the number of the instances in the corresponding cluster and u_i is the feature vector of the $i-th$ sample. To mitigate the limited representation of global features, we also maintain features from

local views to promote the model to discover more detailed information in the learning process. As shown in Fig. 2, these three branches calculate the cluster center according to Eq. (5) independently with their own feature vectors and the shared pseudo label set $\tilde{Y}$. Thus, we can get three sets of cluster centroid representations as following:

$$\begin{cases} \Phi^{gb} = \left\{\phi_1^{gb}, \phi_2^{gb}, \ldots, \phi_C^{gb}\right\} \\ \Phi^{up} = \left\{\phi_1^{up}, \phi_2^{up}, \ldots, \phi_C^{up}\right\} \\ \Phi^{dw} = \left\{\phi_1^{dw}, \phi_2^{dw}, \ldots, \phi_C^{dw}\right\}, \end{cases} \tag{6}$$

where C is the number of the clusters, as these three branches share the same pseudo labels, thus the number of clusters in these three branches are the same. ϕ_i^{gb}, ϕ_i^{up} and ϕ_i^{dw} are the $i-th$ cluster centers in these three branches. In our method, local parts can be roughly aligned for most images as we only split images horizontally into two parts and most images are cropped accurately by detection algorithm. Furthermore, in the pseudo label generation stage, as we also take similarities from local views into consideration, some un-aligned persons can be regarded as outiers by clustering algorithm as they have larger gap with other well-aligned persons. In this way, our method can relieve the negative influence of the not well-aligned problem.

C. Label Noise Purification Module

The training process of state-of-the-art unsupervised person ReID methods can be regarded as two stages. First, pseudo labels are generated by dividing the dataset into diverse clusters, then the model is trained with the pseudo labels. These two stages are conducted in an iterative scheme [16], [17]. However, the noise will be inevitably introduced in the convergence process as the model initialized with ImageNet pre-trained ResNet-50 [56] performs poorly on these person ReID datasets at the beginning, which may accumulate label noise during the training process. To relieve the issue, we propose the LP module, which aims to utilize the knowledge of the teacher to help the student model relieve the influence of the label noise. For a fair comparison, we take the model trained on the same dataset as the teacher model and the new ImageNet pre-trained initialized model as the student model, thus the structures of these two models are the same and it requires no extra information.

Specifically, the teacher model is trained with cluster contrast learning and the proposed FP module with the following objective:

$$\begin{cases} L_q^{gb} = -\log \dfrac{\exp\left(u_q^{gb} \cdot \phi_+^{gb}/\tau\right)}{\sum_{k=0}^{C} \exp\left(u_q^{gb} \cdot \phi_k^{gb}/\tau\right)} \\ L_q^{up} = -\log \dfrac{\exp\left(u_q^{up} \cdot \phi_+^{up}/\tau\right)}{\sum_{k=0}^{C} \exp\left(u_q^{up} \cdot \phi_k^{up}/\tau\right)} \\ L_q^{dw} = -\log \dfrac{\exp\left(u_q^{dw} \cdot \phi_+^{dw}/\tau\right)}{\sum_{k=0}^{C} \exp\left(u_q^{dw} \cdot \phi_k^{dw}/\tau\right)}, \end{cases} \tag{7}$$

where u_q^* is the feature vector of the query instance q from the corresponding view. ϕ_k^* is the centroid feature vector representing the $k-th$ cluster stored in the memory. ϕ_+^* is the centroid feature vector representing the cluster query instance q belongs to stored in the memory. τ is the temperature hyper-parameter and C is the number of the cluster. Different from SPCL [16] and CCL [17] which only use global features in the training process, we also maintain feature representations from local views in the training phase to promote the model to discover more detailed information as following:

$$L_{stage1} = (1-\lambda_2)\, L_q^{gb} + \lambda_2 (L_q^{up} + L_q^{dw}), \tag{8}$$

where λ_2 is the loss weight to balance the importance between global and local features and more details about the training process of teacher model can refer to Sec. III-D.1.

When the trained teacher model is prepared, then LP module can be applied on the student model. This module includes two parts, the warm up part and the knowledge distillation part. More details about these parts can refer to Sec. III-D.2. As the initialized student model performs poorly in the person ReID, the generated pseudo labels will contain numerous label noise in the early training period, thus may cause the feature representation biased. To tackle the issue, in the warm up part, we directly utilize the trained teacher model to generate pseudo labels and use its feature vectors to initialize the cluster center representations as in Eq. (6). Then the student is directly trained with the pseudo labels and fixed cluster center representations generated by the teacher model for a short period. Due to the significant discrepancy between ImageNet and person ReID datasets the student model suffers from more severe label noise at the beginning of the training. Thus, in the proposed warm-up period, the student model can learn the knowledge directly from the teacher model in a fast way to generate more accurate pseudo labels in the early period of the training phase.

In the remaining training phase, given the pseudo labeled dataset and memory center dictionaries as described in Sec. III-B, the student model computes the objective function with multi-view knowledge distillation as following:

$$\begin{cases} L_{Stu}^{gb} = L_q^{gb} + \mu \left| \dfrac{u_q^{gb}}{\left\|u_q^{gb}\right\|} - \dfrac{\tilde{u}_q^{gb}}{\left\|\tilde{u}_q^{gb}\right\|} \right|_2^2 \\ L_{Stu}^{up} = L_q^{up} + \mu \left| \dfrac{u_q^{up}}{\left\|u_q^{up}\right\|} - \dfrac{\tilde{u}_q^{up}}{\left\|\tilde{u}_q^{up}\right\|} \right|_2^2 \\ L_{Stu}^{dw} = L_q^{dw} + \mu \left| \dfrac{u_q^{dw}}{\left\|u_q^{dw}\right\|} - \dfrac{\tilde{u}_q^{dw}}{\left\|\tilde{u}_q^{dw}\right\|} \right|_2^2, \end{cases} \tag{9}$$

where L_{Stu}^{gb}, L_{Stu}^{up} and L_{Stu}^{dw} are the objective functions of three branches of the student model. L_q^{gb}, L_q^{up} and L_q^{dw} are the ClusterNCE loss presented in Eq. (7). μ is the balancing factor. $\{u_q^{gb}, u_q^{up}, u_q^{dw}\}$ and $\{\tilde{u}_q^{gb}, \tilde{u}_q^{up}, \tilde{u}_q^{dw}\}$ are the three feature vectors of query q in the student model and teacher model respectively. Therefore, the final objective function of the student model is as following:

$$L_{stage2} = (1-\lambda_2)\, L_{Stu}^{gb} + \lambda_2 (L_{Stu}^{up} + L_{Stu}^{dw}), \tag{10}$$

Algorithm 1 Training Process of the Teacher Model

Require: Unlabeled training data X
Require: Initialize the encoder f_θ with ImageNet-pretrained ResNet-50
Require: Temperature hyper-parameter τ for Eq. (7)
Require: Balancing factors λ_1 and λ_2 for Eq. (4) and Eq. (8)
Require: Momentum updating factor m for Eq. (11)

for n in [1,*num_epochs*] **do**
  Extract feature vector sets $\{U^{gb}, U^{up}, U^{dw}\}$ from X by f_θ;
  Clustering $\{U^{gb}, U^{up}, U^{dw}\}$ into C clusters with Eq. (4) and DBSCAN;
  Initialize three memory dictionaries individually with Eq. (5) ;
  for i in [1,*num_iterations*] **do**
    Sample $P \times K$ query images from X;
    Compute objective function with Eq. (8) ;
    Update cluster feature representations with Eq. (11);
  end
end

where λ_2 is the balancing factor, which is the same as in Eq. (8). Then the cluster feature representations stored in the memory dictionary sets are updated similar with the teacher model in the following way:

$$\begin{cases} \phi_k^{gb} = m\phi_k^{gb} + (1-m)u_q^{gb} \\ \phi_k^{up} = m\phi_{k_2}^{up} + (1-m)u_q^{up} \\ \phi_k^{dw} = m\phi_{k_3}^{dw} + (1-m)u_q^{dw}, \end{cases} \tag{11}$$

where m is the momentum updating factor. k is the index of the cluster query belongs to, which is the same in these three branches as they share the same pseudo label set. The details of the training procedure of the student model are described in Sec. III-D.2. Note that the pseudo label generation process and training process are conducted iteratively until the model converges. In the test phase, we only adopt the global feature branch for computation efficiency.

D. Training Process

1) Training Process of the Teacher Model: The detailed training process of the teacher model is shown in Algorithm 1. Given the unlabeled dataset X and the encoder f_θ initialized with parameters of ResNet-50 pretrained on ImageNet [56]. For each epoch, we can get the corresponding feature map set F^{gb} with the encoder f_θ. Then we split feature maps F^{gb} into two parts horizontally, which are denoted as F^{up} and F^{dw} respectively. Then GEM pooling is applied to get the corresponding feature vector sets U^{gb}, U^{up} and U^{dw}, respectively. To get the pseudo labels $\tilde{Y}$ for samples in dataset X, Eq. (4) and DBSCAN algorithm are applied. Then, Eq. (5) is used to initialize memory dictionaries individually for these three branches. When the pseudo labels and memory dictionaries are prepared, we start to train the model. Specifically, in each iteration, we firstly sample $P \times K$ query images from X to update parameters of the model according to Eq. (8), where P denotes the number of identities included in each mini-batch while K represents the number of images for each identity. Then features stored in memory dictionaries are updated with Eq. (11).

Algorithm 2 Training Process of the Student Model

Require: Unlabeled training data X
Require: Initialize the encoder f_θ with ImageNet-pretrained ResNet-50
Require: The teacher encoder $\tilde{f}_\theta$ trained on the unlabeled training data X using Algorithm 1
Require: Balancing factors μ for Eq. (9)

`// warm up period`
Extract feature vector sets $\{\tilde{U}^{gb}, \tilde{U}^{up}, \tilde{U}^{dw}\}$ from X by $\tilde{f}_\theta$;
Clustering $\{\tilde{U}^{gb}, \tilde{U}^{up}, \tilde{U}^{dw}\}$ into C clusters with Eq. (4) and DBSCAN;
Initialize three memory dictionaries individually with Eq. (5) ;
for i in $[1, num\_iterations \times 2]$ **do**
  Sample $P \times K$ query images from X;
  Compute objective function with Eq. (8) ;
end
`// knowledge distillation period`
for n in [1,*num_epochs*] **do**
  Extract feature vector sets $\{U^{gb}, U^{up}, U^{dw}\}$ from X by f_θ;
  Clustering $\{U^{gb}, U^{up}, U^{dw}\}$ into C clusters with Eq. (4) and DBSCAN;
  Initialize three memory dictionaries individually with Eq. (5) ;
  for i in [1,*num_iterations*] **do**
    Sample $P \times K$ query images from X;
    Compute objective function with Eq. (10) ;
    Update cluster feature representations with m and Eq. (11);
  end
end

2) Training Process of the Student Model: The detailed training process of the student model is shown in Algorithm 2. Besides the unlabeled dataset X and the encoder f_θ initialized with parameters of ResNet-50 pretrained on ImageNet, the teacher model trained following Algorithm 1 is also required. The training process of the student model includes two parts, the warm-up part and the knowledge distillation part.

In the warm-up part, we directly utilize the trained teacher model to encode the dataset X into feature map set F^{gb}, and then use horizontally split operation and GEM pooling to obtain corresponding feature vector sets U^{gb}, U^{up} and U^{dw}. Then Eq. (4) and DBSCAN algorithm are applied to get the pseudo labels and initialized memory dictionaries for each branch as described in the training process of the teacher model. Then $P \times K$ query images are sampled from X to

update parameters of the model according to Eq. (8) without updating the features stored in memory dictionaries. Note that in this part we aim to use fixed cluster centers stored in memory dictionaries from the teacher model to help the student model directly learn in a fast way to avoid label noise accumulation in the early period. Then in the knowledge distillation part, for each epoch the training procedure of the student model is the same as the teacher model except that we update parameters of the student model according to Eq. (10) which contains the regularization of knowledge distillation in a global-local manner.

IV. Experiment

A. Datasets and Evaluation Protocol

We conduct experiments on two public person Re-ID benchmarks, including Market-1501 [57] and MSMT17 [27]. Market-1501 dataset contains 32,668 images of 1,501 IDs captured by 6 different cameras. MSMT17 dataset contains 126,441 images of 1,041 IDs captured by 15 different cameras.

Following existing person ReID works [16], [17], [31], [57], we also adopt mean average precision (mAP) and Cumulated Matching Characteristics (CMC) as the evaluation metrics, and we report Top-1, Top-5, and Top-10 of the CMC evaluation metric in the paper. For fair comparisons, we don't adopt any post processing techniques in the evaluation period. As the setting of other purely unsupervised ReID works, we don't use any labeled data or other source domain datasets in the training process.

B. Implementations Details

We use the Resnet-50 [56] initialized with the parameters pre-trained on the ImageNet [67] as the backbone encoder. Following existing cluster contrast framework [17], we remove all sub-module layers after layer-4 and add GEM pooling followed by batch normalization layer [68] and L2-normalization layer. During training, we use the DBSCAN [23] as clustering algorithm to generate pseudo labels at the beginning of each epoch. During test phase, we only adopt the feature vector of the first global feature branch for computation efficiency.

For training, each mini-batch contains 256 images of 16 pseudo person identities, which are resized as 256×128. For input images, random horizontal flipping, padding, random cropping, and random erasing [69] are applied. To train our model, Adam optimizer with weight decay 5e-4 is adopted. We set the initial learning rate as 3.5e-4, and reduce it every 20 epochs for a total of 50 epochs. The balancing factor λ_1 in Eq. (4) is set to 0.15 while the balancing factor λ_2 in Eq. (8) is set to 0.2. The balancing factor μ in Eq. (9) is set to 1. For DBSCAN clustering algorithm, the minimal number of neighbours is set to 4, which is the same as other unsupervised person ReID methods, such as CCL [17], ICE [54] and HDCPD [31], etc. The maximum distance d is set to 0.6 for Market1501 and 0.7 for MSMT17.

C. Comparison With State-of-the-Arts

We compare our proposed method with the state-of-the-art unsupervised person ReID methods, including UDA person ReID and fully unsupervised person ReID. The result is shown in Table I. Although these methods leverage the knowledge of the source domain, our proposed method outperforms all of them on these two datasets. The reason is probably that the gap between source and target domains is large and it is hard to transfer the knowledge from source domain to the target domain.

Compared with the state-of-the-art fully unsupervised person ReID methods, our proposed method also achieves better performance. As our proposed purification modules are established on the framework of CCL [17], our method outperforms CCL by 3.2% and 6.2% in terms of mAP on Market-1501 and MSMT17 datasets, respectively. Compared with Market-1501 dataset, more gains can be achieved on the MSMT17 dataset. The reason is probably that more noise exist in MSMT17 dataset as it is more challenging compared with Market-1501. Although the pseudo labels generated by the ImageNet pretrained model are low-quality, the model initialized with parameters pre-trained on the large-scale diverse ImageNet dataset can discover general patterns on the downstream person ReID datasets. Then pseudo label generation stage and training stage are conducted iteratively to promote each other. And these are critical factors why existing unsupervised person ReID methods can achieve great performance on these challenging person ReID datasets. In addition, our proposed label noise purification module can further relieve the label noise.

D. Ablation Studies

In this section, we study the effectiveness of different components and hyper-parameters in our proposed method. As our work is implemented based on the CCL [17], the hyper-parameters introduced in our method include the balancing factors λ_1, λ_2 and μ. The other hyper-parameters follow the setting of CCL.

1) Different Combinations of the Components: As our method is combined with two different purification modules, we conduct experiments on these two person ReID datasets described in Sec. IV-A versus different combinations of different modules. As our work is implemented based on CCL, we take CCL as baseline and our proposed modules include FP module and LP module. As shown in Table II, the first line means the result of CCL on different person ReID datasets. Compared with previous methods, CCL can achieve a good performance by taking advantage of contrastive learning and cluster center memory, but it is still limited by the feature bias and label noise as mentioned in the paper. The second line is the result of the combination of CCL and our proposed FP module, compared with the first line we can find that our proposed FP module can improve the baseline by 1.6% and 1.0% in terms of mAP on Market-1501 and MSMT17 datasets. The third line is the result of the combination of CCL and our proposed LP module, compared with the first line, the improvement of 1.9% and 3.4% in terms of mAP can be achieved on these datasets. The last line denotes the result of the combination of CCL and our proposed two purification modules. Compared with the first line, the improvement of 3.0% and 5.3% in terms of mAP can be

TABLE I

EXPERIMENTAL RESULTS OF OUR PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON MARKET-1501 AND MSMT17. THE TOP THREE RESULTS ARE MARKED AS RED, BLUE AND GREEN, RESPECTIVELY

Method	Reference	Market-1501				MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10
Unsupervised Domain Adaption									
ECN [58]	CVPR'19	43.0	75.1	87.6	91.6	10.2	30.2	41.5	46.8
MMCL [59]	CVPR'20	60.4	84.4	92.8	95.0	16.2	43.6	54.3	58.9
JVTC [60]	ECCV'20	61.1	83.8	93.0	95.2	20.3	45.4	58.4	64.3
DG-Net++ [61]	ECCV'20	61.7	82.1	90.2	92.7	22.1	48.8	60.9	65.9
MMT [5]	ICLR'20	71.2	87.7	94.9	96.9	23.3	50.1	63.9	69.8
DCML [62]	ECCV'20	72.6	87.9	95.0	96.7	-	-	-	-
MEB [63]	ECCV'20	76.0	89.9	96.0	97.5	-	-	-	-
SPCL [16]	NeurIPS'20	76.7	90.3	96.2	97.7	26.8	53.7	65.0	69.8
HCD [64]	ICCV'21	80.2	91.4	-	-	29.3	56.1	-	-
IDM [29]	ICCV'21	82.8	93.2	97.5	98.1	35.4	63.6	75.5	80.2
CCL++ [21]	ICCV'21	83.4	94.2	-	-	36.3	66.6	-	-
Fully Unsupervised									
BUC [12]	AAAI'19	29.6	61.9	73.5	78.2	-	-	-	-
SSL [13]	CVPR'20	37.8	71.7	83.8	87.4	-	-	-	-
JVTC [60]	ECCV'20	41.8	72.9	84.2	88.7	15.1	39.0	50.9	56.8
MMCL [59]	CVPR'20	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8
HCT [14]	CVPR'20	56.4	80.0	91.6	95.2	-	-	-	-
CycAs [65]	ECCV'20	64.8	84.8	-	-	26.7	50.1	-	-
GCL [66]	CVPR'21	66.8	87.3	93.5	95.5	21.3	45.7	58.6	64.5
SPCL [16]	NeurIPS'20	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61.2
HCD [64]	ICCV'21	78.1	91.1	96.4	97.7	26.9	53.7	65.3	70.2
ICE [54]	ICCV'21	79.5	92.0	97.0	98.1	29.8	59.0	71.7	77.0
CCL [17]	ACCV'22	82.6	93.0	97.0	98.1	33.3	63.3	73.7	77.8
MCL [32]	MM'22	82.9	92.7	97.6	98.7	38.2	66.5	75.2	79.7
HDCPD [31]	TIP'22	84.5	93.5	97.6	98.6	24.6	50.2	61.4	65.7
PPLR [38]	CVPR'22	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
ISE [33]	CVPR'22	85.3	94.3	98.0	98.8	37.0	67.6	77.5	81.0
Ours	-	85.8	94.5	97.8	98.7	39.5	67.9	78.0	81.6

TABLE II

ABLATION STUDY ON MARKET-1501 AND MSMT17 DATASETS

Method	Market-1501		MSMT17	
	mAP	R1	mAP	R1
Baseline	82.8	92.7	34.2	64.2
Baseline +FP	84.4	93.5	35.2	64.5
Baseline +LP	84.7	93.6	37.6	67.6
Baseline +FP+LP	**85.8**	**94.5**	**39.5**	**67.9**

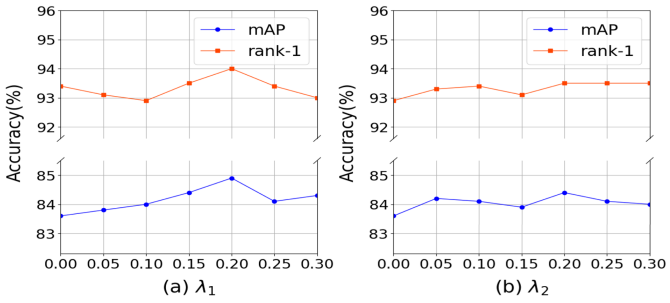


Fig. 5. Impact of hyper-parameter λ_1 and λ_2 of the teacher model on Market-1501 dataset. In (a) λ_2 is fixed to 0.2 while in (b) λ_1 is fixed to 0.15.

achieved on these datasets. The result shows that our proposed two purification modules can work in a mutual benefit way and the baseline with these two modules can achieve the best performance. Furthermore, compare the third line with the first line we can also find that the LP module is more effective on MSMT17 than Market-1501 dataset. The reason is probably that compared with Market-1501 dataset, the MSMT17 dataset is more challenging which contains more occluded images. Thus the LP module can mitigate the severe side effect of label noise introduced in the clustering process.

2) Balancing Factors: As the aim of our proposed extra two branches is to encourage the model to explore more discriminative local cues, the weights of these branches play an important role in our experiment. Fig. 5 and Fig. 6 report the result versus different values of balancing factors λ_1 and λ_2 in Eq. (4) and Eq. (8) on Market-1501 and MSMT17 datasets,

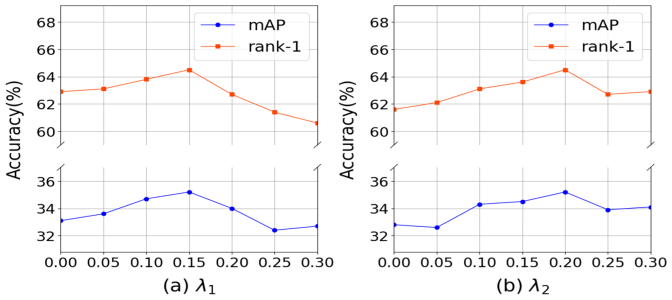


Fig. 6. Impact of hyper-parameter λ_1 and λ_2 of the teacher model on MSMT17 dataset. In (a) λ_2 is fixed to 0.2 while in (b) λ_1 is fixed to 0.15.

respectively. Results in Fig. 5 (a) and Fig. 6 (a) show that the teacher model can achieve the bast results on Market-1501 and MSMT17 with λ_1 set to 0.2 and 0.15, respectively. To balance between different datasets, we set λ_1 to 0.15 by default for all datasets. From Fig. 5 (b) and Fig. 6 (b), we can see that when λ_2 is set to 0.2, the teacher model can achieve the best results on both datasets. As λ_1 and λ_2 are two critical factors in our method, we carefully tune these hyper-parameters to obtain the optimal values for the task as other works [31], [33], [38]. As shown in Fig. 5 and Fig. 6, although the experiment results rely on the hyper-parameters λ_1 and λ_2, generally they can achieve better performance on Market1501 when λ_1 and λ_2 are both setting to 0.2. Specifically, when λ_1 is set in the interval of (0.15, 0.25), the proposed model delivers good results, e.g., in terms of mAP, which is an important evaluation metric for person ReID. While on the more challenging MSMT17 dataset, the improvements are more obvious when λ_1 and λ_2 are setting to 0.15 and 0.2, respectively. For simplicity, in the experiment we set λ_1 and λ_2 to 0.15 and 0.2 for all datasets, respectively.

Hyper-parameter μ in Eq. (9) is another important balancing factor, which determines the weight of the guidance of teacher in the overall training process. If μ is too small, then the student model will learn without enough guidance from the teacher model. On the other hand, if μ is too large, the student model will be forced to mimic the teacher model, which limits the generalization of the learned feature representations. Table III shows the results under different values of μ on Market-1501 and MSMT17 datasets. As can be bseen, the model can achieve the best performance on both Market-1501 and MSMT17 datasets with μ set to 0.5 and 1.0, respectively. To balance between different datasets, we set μ to 1.0 in our experiments for all datasets. It is noteworthy that we assume that the trained teacher model performs better than the initialized student model. Thus, our proposed label noise purification module aims to utilize the trained teacher model to guide the student model to relieve label noise. Although the teacher model may hinder the student model learning in the latter period, we add the hyper-parameter mu in Eq. (9) to determine the weight of the guidance of the teacher model in the training process. While learning from the teacher model, the student model is also encouraged to explore by itself. These two phases are balanced by the hyper-parameter mu to relieve the negative influence of the teacher model to some extent. Therefore, it is reasonable that the student model performs better than the fixed teacher model. As how to determine the turning point from which the teacher model disturbs the student model is challenging due to lack of annotated labels, we leave it in the future work, e.g., by designing some measures in the unsupervised setting [70].

TABLE III
IMPACT OF HYPER-PARAMETER μ ON MARKET-1501 AND MSMT17 DATASETS

μ	Market-1501		MSMT17	
	mAP	R1	mAP	R1
0.5	**86.1**	94.3	39.3	67.5
1.0	85.8	**94.5**	**39.5**	**67.9**
1.5	85.8	93.9	38.8	67.4
2.0	85.8	94.0	38.7	67.6
2.5	85.9	94.4	37.9	66.8
3.0	85.8	94.3	37.5	66.7

TABLE IV
PERFORMANCE OF THE MODEL WITH DIFFERENT TEACHER MODELS ON MARKET-1501 AND MSMT17 DATASETS

Model	Market-1501		MSMT17	
	mAP	R1	mAP	R1
Teacher Model-1	75.1	88.8	29.6	57.6
Student Model-1	84.8	94.0	36.8	65.6
Teacher Model-2	82.7	92.9	32.0	60.6
Student Model-2	**86.3**	**94.6**	38.0	66.6
Teacher Model-3	84.4	93.5	34.8	64.4
Student Model-3	85.8	94.5	**39.5**	**67.9**

As the teacher model in our method is trained in advance, we further analyse the performance of the model with different teacher models. To obtain diverse teacher models, we change the hyper-parameter epoch to obtain a series of teacher models. The result is shown in Tab. IV. As can be seen from the table, generally the student model performs better with a stronger teacher model except for the Market-1501 dataset. As how to obtain the appropriate trained teacher model is challenging due to lack of annotations, we leave it in the future research, e.g., by designing some measures in the unsupervised setting [70].

3) Compared GEM Pooling With Other Types of Pooling: In our method, we use GEM pooling [53] to obtain feature vectors. GEM pooling is a strong trick, which is widely adopted in many existing popular unsupervised person ReID methods, including CCL [17], HDCPD [31], ISE [33] and MCL [32], etc. We also compare Gem pooling with GAP/GMP in our experiment and results are shown in Tab. V. From the result we can find that the combination of our model and Gem pooling can achieve the best result. The reason could be that gem pooling can obtain more suitable feature vectors adaptively as GAP and GMP are just special cases of Gem pooling. For more details about Gem pooling please refer to [53].

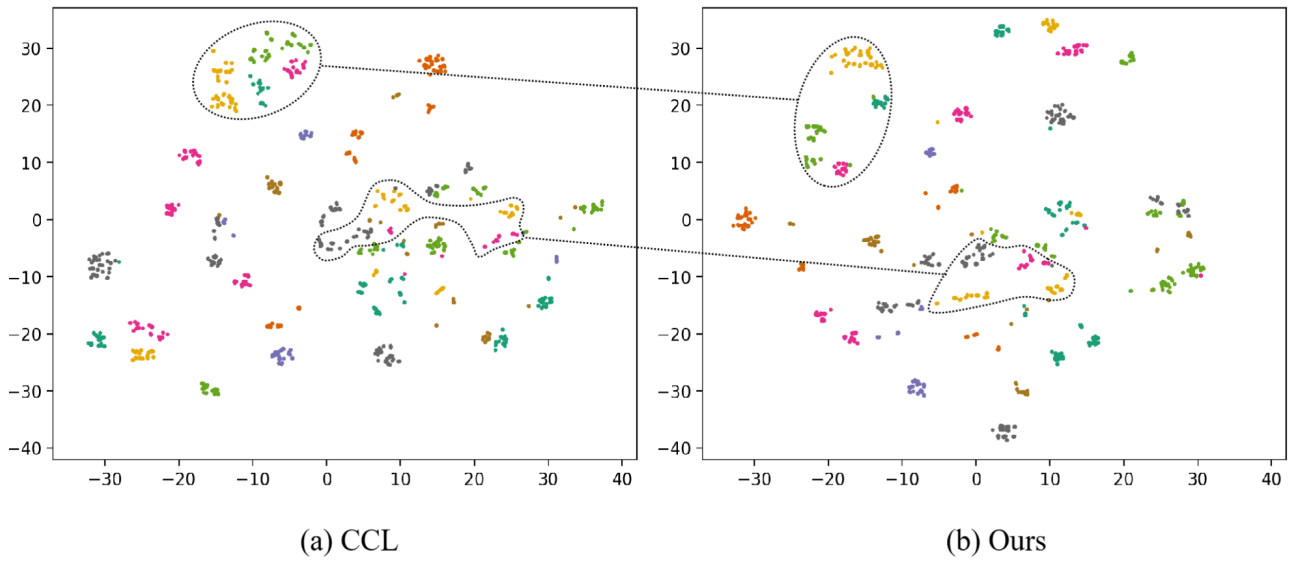


Fig. 7. T-SNE visualization of the learned features on a subset of MSMT17 training set. Points of the same color in the same dashed line represent features of the same identity. Compared with CCL, our method is more discriminative for the hard negative samples while have more compact features for the hard positive samples.

TABLE V
IMPACT OF DIFFERENT POOLING OPERATIONS ON MARKET-1501 AND MSMT17 DATASETS

Method	Market-1501		MSMT17	
	mAP	R1	mAP	R1
Ours (GAP)	84.9	94.0	37.4	66.4
Ours (GMP)	82.7	92.8	34.6	64.3
Ours (GEM)	**85.8**	**94.5**	**39.5**	**67.9**

TABLE VI
RESULTS OF OUR METHOD WITH DIFFERENT PSEUDO LABEL GENERATION MECHANISMS ON MARKET-1501 AND MSMT17 DATASETS. THE INDIVIDUAL LABELS MECHANISM MEANS USING INDIVIDUAL PSEUDO LABELS FOR EACH GLOBAL AND LOCAL BRANCHES. THE AGGREGATED LABELS MECHANISM MEANS COMBINING ALL BRANCHES TO FORM THE SAME PSEUDO LABELS FOR THESE BRANCHES, WHICH IS THE MECHANISM USED IN OUR METHOD

dataset	pseudo labels	mAP	R1	R5	R10
Market-1501	Individual Labels	85.4	94.0	97.7	98.6
	Aggregated Labels	**85.8**	**94.5**	**97.8**	**98.7**
MSMT17	Individual Labels	38.4	67.3	77.2	81.0
	Aggregated Labels	**39.5**	**67.9**	**78.0**	**81.6**

4) Individual Pseudo Labels for Each Branch: We use the same pseudo labels for these different branches mainly for two reasons. On the one hand, using the same pseudo labels for these branches is more efficient, which only needs to apply clustering algorithm on the final distance matrix to generate pseudo labels once for different branches before each epoch. On the other hand, we aim to aggregate all these global and local branches to refine pseudo labels for them and obtain more discriminative and consistent feature representations. To analyse the effect of such aggregation mechanism, we also conduct experiments by adopting individual pseudo labels for each of global and local branches. As shown in Tab. VI, the performance of our method degenerates slightly when individual pseudo labels are adopted for different branches, and the reason may be that individual pseudo labels generated for local branches suffer from serious label noise as similar local parts can be easily merged to the same cluster.

5) Combine Local Features in the Inference Stage: In our method, as local features are only involved in the training process, we also add the experiment to analyse the performance of our method by combining local features in the inference stage as $D = (1-2\alpha)D^{gb} + \alpha D^{up} + \alpha D^{dw}$, where D^{gb}, D^{up} and D^{dw} are Jaccd Distance matrices calculated from global and local feature vectors, α is the hyper-parameter to fuse these matrices. D is the final distance matrix to obtain the retrieval results. Tab. VII shows the results under different values of α in the inference stage on Market-1501 and MSMT17 datasets. As can be seen, the performance of the model shows no obvious fluctuations with different values of α. The reason could be that in the training process, global and local branches of the model share the same pseudo labels, thus these branches tend to capture more consistent semantic representations.

6) Update the Teacher Model in a Momentum Manner: As our label noise purification module is proposed based on the phenomenon that the trained model is more accurate than the initialized model, we utilize the trained teacher model to help the student model relieve the influence of label noise. Although such mechanism can perform well in the early stage, it may hinder the student model learning in the later training stage as the parameters of the teacher model are fixed during

TABLE VII
IMPACT OF HYPER-PARAMETER α ON MARKET-1501 AND MSMT17 DATASETS

α	Market-1501		MSMT17	
	mAP	R1	mAP	R1
0.0	**85.8**	94.5	39.5	67.9
0.05	**85.8**	**94.7**	39.6	**68.0**
0.1	85.7	94.6	**39.7**	67.9
0.15	85.6	94.5	**39.7**	67.9
0.2	85.5	94.6	**39.7**	**68.0**

TABLE VIII
IMPACT OF MOMENTUM HYPER-PARAMETER γ ON MARKET-1501 AND MSMT17 DATASETS. THE TRAINED MODEL IS USED AS THE INITIAL TEACHER MODEL

γ	Market-1501		MSMT17	
	mAP	R1	mAP	R1
0.99	83.6	93.1	39.3	67.1
0.999	85.3	94.1	**40.8**	**69.0**
1.0	**85.8**	**94.5**	39.5	67.9

the training process. We also conduct experiments to study whether it can further improve the performance of the student model if the parameters of the trained teacher model are updated in a momentum manner, i.e., the teacher model is trained on the ReID dataset in advance, then it continues to update its parameters in a momentum scheme with the parameters of the student model as follows,

$$\theta_t = \gamma\theta_t + (1-\gamma)\theta_s, \tag{12}$$

where θ_t and θ_s are the parameters of the teacher model and student model, respectively. γ is the momentum hyper-parameter. Note that in our method we use the fixed trained teacher, and it can be regarded as a special case where $\gamma = 1$. The experiment results are shown in Tab. VIII. As can be seen from the table, when the momentum hyper-parameter γ is set to 0.999, our method can achieve better performance on MSMT dataset while lower results on Market-1501 dataset with the teacher model further updated in a momentum manner. The reason could be that compared with Market-1501 dataset, MSMT17 dataset is more challenging and the trained teacher model cannot obtain relatively discriminative feature representations as other datasets. Thus, using the fixed trained teacher model may hinder the student model learning in the later training stage.

In the online knowledge distillation method [31], the teacher model and the student model are both initialized with ImageNet-pretrained parameters and then updated. The discrepancy between ImageNet and person ReID datasets has a side effect on the prediction of the teacher model, leading to noisy pseudo labels, which will produce erroneous supervisory signals and mislead the training process. Different from the online knowledge distillation scheme, we aim to resort to the well-trained teacher model for help in the unsupervised person ReID setting. As the trained teacher model is more accurate than the initialized student model, the teacher model can guide the student model to relieve the label noise in the early period of the training phase. We also add the experiment to replace our teacher model with ImageNet-pretrained model and update it in an online manner. As shown in Tab. IX, the performance of the model drops greatly compared with our method, which supports the claim and validates the effectiveness of the proposed method.

TABLE IX
IMPACT OF MOMENTUM HYPER-PARAMETER γ ON MARKET-1501 AND MSMT17 DATASETS. THE IMAGENET-PRETRAINED MODEL IS USED AS THE INITIAL TEACHER MODEL

γ	Market-1501		MSMT17	
	mAP	R1	mAP	R1
0.99	**83.1**	**93.2**	**37.0**	**66.3**
0.999	81.8	92.6	33.5	62.4

TABLE X
PERFORMANCE OF THE STUDENT MODEL TRAINED DIRECTLY WITH PSEUDO LABELS GENERATED BY THE TEACHER MODEL ON MARKET-1501 AND MSMT17 DATASETS

dataset	mAP	R1	R5	R10
Market-1501	84.8	94.0	97.5	98.5
MSMT17	37.4	66.7	77.3	81.2

7) Generate Pseudo Labels With the Teacher Model: In our proposed label noise purification module, knowledge distillation is utilized to make the student model learn from the trained teacher model with MSE loss. To further analyse the influence of the trained teacher model, we use trained teacher model to generate pseudo labels for student model without MSE loss. Specifically, in the training stage of the student model, pseudo labels are generated by the trained teacher model before each epoch. The result is shown in Tab. X. As can be seen from the table, only using the teacher model to generate pseudo labels for student model is not as efficient as our original label noise purification module. The reason could be that the student model may easily overfit the label noise caused by the fixed teacher model, and in our label noise purification module, the student model is also encouraged to explore the pseudo labels by itself while learning from the teacher model.

8) Use More Local Parts in the Training Stage: As persons in the dataset are not strictly aligned, we only select two parts to roughly align most of them. We also add the experiment by using more parts to mine local cues, and the result is shown in Tab. XI. We can find that the accuracy of the model becomes lower when increasing the number of parts. The reason could be that it is hard to align these fine local parts as most persons in the dataset are not strictly aligned.

9) Compared With IBN-ResNet-50 and ResNet-50 Backbones: As Instance Normalization (IN) can learn features that are invariant to appearance changes, while Batch Normalization (BN) is essential for preserving content related

TABLE XI
IMPACT OF THE NUMBER OF LOCAL PARTS N_p ON MARKET-1501 AND MSMT17 DATASETS

N_p	Market-1501		MSMT17	
	mAP	R1	mAP	R1
2	**85.8**	**94.5**	**39.5**	**67.9**
3	85.4	94.2	38.1	67.1
4	84.9	93.8	36.6	65.4

TABLE XII
RESULTS OF DIFFERENT METHODS WITH IBN-RESNET-50 BACKBONE ON MARKET-1501 AND MSMT17 DATASETS

dataset	method	mAP	R1	R5	R10
Market-1501	SPCL [16]	73.8	88.4	95.3	97.3
	CCL [17]	84.1	93.2	97.6	98.2
	Ours	**86.9**	**94.4**	**97.7**	**98.5**
MSMT17	SPCL [16]	24.0	48.9	61.8	67.1
	CCL [17]	41.1	69.1	79.3	83.1
	Ours	**47.9**	**74.5**	**83.8**	**86.7**

information, IBN-Net [71] can achieve better performance by integrating Instance Normalization and Batch Normalization. Thus, IBN-ResNet-50 can be regarded as a stronger baseline by replacing the BN in ResNet-50 with IBN. We also conduct experiments to compare these two encoders, while keeping other experiment settings the same. Comparing the results in Table XII and Table I, we can find that our proposed method can achieve better performance with the IBN-ResNet-50 backbone than the ResNet-50 backbone.

10) Pre-Train the Model With Different Source Datasets: As pre-training stage is important for unsupervised person ReID methods, a better pre-trained model can also boost the final performance of the trained model. Recently, some works [72], [73] use larger person ReID dataset, named LUPerson, to pre-train the model, and their results show that model pretrained on LUPerson can achieve better performance on unsupervised person ReID than the model pre-trained on ImageNet. We believe that using more advanced pre-trained model can further improve our proposed method. We also add the experiemnt to replace the initial model with the ResNet-50 pre-trained on LUPerson-NL, which is released in [72]. As shown in Tab. XIII, our model can also benefit from the initialization of model weights pretrained on the large-scale LUPerson-NL dataset and outperforms the ImageNet-initialized counterpart significantly.

11) Qualitative Analysis of Visualization: To further understand the discrimination ability of our method, we utilize t-SNE [74] to visualize the features learned by the baseline and our method. We provide the visualization on the more challenging MSMT17 dataset and resize them to the same scale. As shown in Fig. 7, features of the same identity are usually clustered together in both CCL and our proposed method, which verifies the effectiveness of CCL and our method. More specifically, points of the same color in the same closed dotted line represent images of the same identity. As can be seen, compared with CCL, feature representations extracted from our method are more discriminative for different persons while more compact for the same person.

TABLE XIII
PERFORMANCE OF THE MODEL INITIALIZED WITH DIFFERENT SOURCE DATASETS ON MARKET-1501. CCL [17] IS USED AS THE BASELINE

Source Dataset	Method	Market-1501	
		mAP	R1
ImageNet	Baseline	82.8	92.7
	Ours	**85.8**	**94.5**
LUPerson-NL	Baseline	84.3	93.3
	Ours	**88.0**	**95.0**

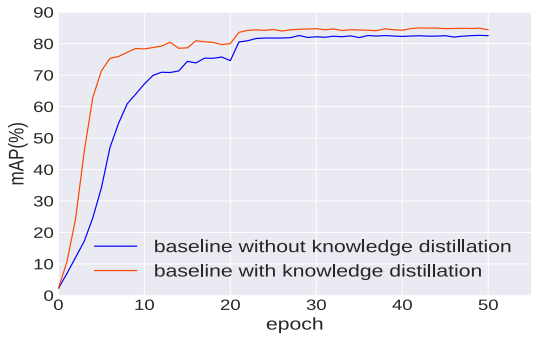


Fig. 8. Accuracy of our model with/without knowledge distillation during training on Market-1501.

V. CONCLUSION

In the paper we propose the purification method for unsupervised person ReID. Two novel purification modules are devised. Specifically, the feature purification module takes into account the features from two local views to enrich the feature representation to purify the inherent feature bias of the global feature involved. The label noise purification module helps purify the label noise by taking advantage of the knowledge of teacher model in an offline scheme. Extensive experiments on two challenging person ReID datasets demonstrate the superiority of our method over state-of-the-art methods.

APPENDIX
INFLUENCE OF KNOWLEDGE DISTILLATION

To investigate the influence of knowledge distillation, we show the test accuracy of the baseline with/without knowledge distillation in each epoch. Due to the lack of ground truth labels, the model has to be trained with the pseudo labels generated by clustering algorithm. In this way, the noise will be inevitably introduced in the convergence process as the model initialized with ImageNet pre-trained ResNet-50 performs poorly on these person ReID datasets. As shown in Fig. 8 and Fig. 9, the student model with the offline knowledge distillation converges faster than its counterpart without

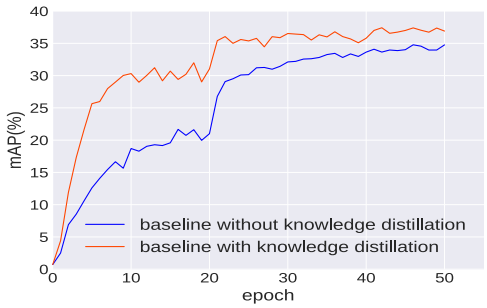


Fig. 9. Accuracy of our model with/without knowledge distillation during training on MSMT17.

knowledge distillation, since it mitigates the interference of noisy labels.

REFERENCES

[1] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7789–7817, May 2021.

[2] Z. Hu, C. Zhu, and G. He, "Hard-sample guided hybrid contrast learning for unsupervised person re-identification," 2021, *arXiv:2109.12333*.

[3] D. Kumar, P. Siva, P. Marchwica, and A. Wong, "Unsupervised domain adaptation in person re-ID via k-reciprocal clustering and large-scale heterogeneous environment synthesis," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2634–2643.

[4] Y. Ge, F. Zhu, R. Zhao, and H. Li, "Structured domain adaptation with online relation regularization for unsupervised person re-ID," 2020, *arXiv:2003.06650*.

[5] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[6] G. Wei, C. Lan, W. Zeng, and Z. Chen, "MetaAlign: Coordinating domain alignment and classification for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16638–16648.

[7] N. Xiao and L. Zhang, "Dynamic weighted learning for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15237–15246.

[8] J. Na, H. Jung, H. J. Chang, and W. Hwang, "FixBi: Bridging domain spaces for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1094–1103.

[9] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[10] L. Gao, J. Zhang, L. Zhang, and D. Tao, "DSP: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2825–2833.

[11] W. Wang et al., "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1730–1738.

[12] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8738–8745.

[13] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3387–3396.

[14] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13654–13662.

[15] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Energy clustering for unsupervised person re-identification," *Image Vis. Comput.*, vol. 98, Jun. 2020, Art. no. 103913.

[16] Y. Ge, F. Zhu, D. Chen, and R. Zhao, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11309–11321.

[17] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1142–1160.

[18] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[19] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[20] X. Lin, P. Ren, C.-H. Yeh, L. Yao, A. Song, and X. Chang, "Unsupervised person re-identification: A systematic survey of challenges and solutions," 2021, *arXiv:2109.06057*.

[21] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, and S. Wang, "Towards discriminative representation learning for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8506–8516.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Oakland, CA, USA,1967, pp. 281–297.

[23] M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, no. 34, 1996, pp. 226–231.

[24] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z. Zha, "Group-aware label transfer for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5306–5315.

[25] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[26] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[27] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[28] Y. Ge, F. Zhu, D. Chen, R. Zhao, X. Wang, and H. Li, "Structured domain adaptation with online relation regularization for unsupervised person re-ID," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 18, 2022, doi: 10.1109/TNNLS.2022.3173489.

[29] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L. Duan, "IDM: An intermediate domain module for domain adaptive person re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11844–11854.

[30] Y. Dai, Y. Sun, J. Liu, Z. Tong, Y. Yang, and L.-Y. Duan, "Bridging the source-to-target gap for cross-domain person re-identification with intermediate domains," 2022, *arXiv:2203.01682*.

[31] D. Cheng, J. Zhou, N. Wang, and X. Gao, "Hybrid dynamic contrast and probability distillation for unsupervised person re-ID," *IEEE Trans. Image Process.*, vol. 31, pp. 3334–3346, 2022.

[32] X. Jin et al., "Meta clustering learning for large-scale unsupervised person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2163–2172.

[33] X. Zhang et al., "Implicit sample extension for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7359–7368.

[34] D. Li et al., "Self-guided hard negative generation for unsupervised person re-identification," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–7.

[35] T. Si, F. He, Z. Zhang, and Y. Duan, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Trans. Multimedia*, early access, May 11, 2022, doi: 10.1109/TMM.2022.3174414.

[36] X. Han et al., "Rethinking sampling strategies for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 29–42, 2023.

[37] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L. Duan, "Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 7815–7829, 2021.

[38] Y. Cho, W. J. Kim, S. Hong, and S. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7298–7308.

[39] M. Li, C. Li, and J. Guo, "Cluster-guided asymmetric contrastive learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 3606–3617, 2022.

[40] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3435–3444.

[41] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[42] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. H. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, vol. 31, pp. 379–391, 2022.

[43] J. Zhang, Z. Chen, and D. Tao, "Towards high performance human keypoint detection," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2639–2662, Sep. 2021.

[44] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–16.

[45] Y. Fu et al., "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8295–8302.

[46] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.

[47] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6111–6120.

[48] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13873–13882.

[49] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 610–619.

[50] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 266–282.

[51] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, 2020.

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[53] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.

[54] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," 2021, *arXiv:2103.16364*.

[55] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3652–3661.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2015, pp. 1116–1124.

[58] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.

[59] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10978–10987.

[60] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 483–499.

[61] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 87–104.

[62] G. Chen, Y. Lu, J. Lu, and J. Zhou, "Deep credible metric learning for unsupervised domain adaptation person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 643–659.

[63] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 594–611.

[64] Y. Zheng et al., "Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8351–8361.

[65] Z. Wang et al., "CycAs: Self-supervised cycle association for learning re-identifiable descriptions," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 72–88.

[66] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2004–2013.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[69] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.

[70] K. Saito, D. Kim, P. Teterwak, S. Sclaroff, T. Darrell, and K. Saenko, "Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9164–9173.

[71] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 464–479.

[72] D. Fu et al., "Large-scale pre-training for person re-identification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2476–2486.

[73] D. Fu et al., "Unsupervised pre-training for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14745–14754.

[74] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.