# Highly Efficient Active Learning With Tracklet-Aware Co-Cooperative Annotators for Person Re-Identification

Xiao Teng, Long Lan, *Member, IEEE*, Jing Zhao, Xueqiong Li, and Yuhua Tang

***Abstract*— Supervised person re-identification (ReID) has attracted widespread attentions in the computer vision community due to its great potential in real-world applications. However, the demand of human annotation heavily limits the application as it is costly to annotate identical pedestrians appearing from different cameras. Thus, how to reduce the annotation cost while preserving the performance remains challenging and has been studied extensively. In this article, we propose a tracklet-aware co-cooperative annotators' framework to reduce the demand of human annotation. Specifically, we partition the training samples into different clusters and associate adjacent images in each cluster to produce the robust tracklet which decreases the annotation requirements significantly. Besides, to further reduce the cost, we introduce a powerful teacher model in our framework to implement the active learning strategy and select the most informative tracklets for human annotator, the teacher model itself, in our setting, also acts as an annotator to label the relatively certain tracklets. Thus, our final model could be well-trained with both confident pseudo-labels and human-given annotations. Extensive experiments on three popular person ReID datasets demonstrate that our approach could achieve competitive performance compared with state-of-the-art methods in both active learning and unsupervised learning (USL) settings.**

***Index Terms*— Active learning, co-cooperative annotators, person re-identification (ReID).**

## NOMENCLATURE

| Notations | Meaning |
|---|---|
| $X$ | Unlabeled images. |
| $\tilde{Y}$ | Generated pseudo labels. |
| $L$ | Ground truth. |
| $\mathcal{E}_p$ | Human expert annotator. |
| $I$ | Frame numbers. |
| $C$ | Cluster set. |
| $X_t$ | Unlabeled images of the $t$th cluster. |
| $X'_t$ | Tracklet centers of the $t$th cluster. |
| $T_t$ | Tracklet indexes of the $t$th cluster. |
| $\Omega^+$ | Samples whose annotations are consistent with the previous cluster. |
| $\Omega^-$ | Samples whose annotations are conflict with the previous cluster. |
| $f_\theta$ | Model to be trained. |
| $f_{\theta'}$ | Trained teacher model. |

Manuscript received 31 October 2022; revised 17 April 2023; accepted 21 June 2023. This work was supported in part by the National Grand Research and Development Plan under Grant 2020AAA0103501 and in part by the National Natural Science Foundation of China under Grant 61906210. *(Corresponding author: Long Lan.)*

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3289178.

Digital Object Identifier 10.1109/TNNLS.2023.3289178

## I. INTRODUCTION

PERSON re-identification (ReID) aims to retrieve the same person under different camera views, which is a hot topic in the computer vision community due to its potential in real-world applications [1], [2], [3], [4], [5], [6], [7], [8], [9]. In recent years, thanks to the publication of large-scale labeled datasets, numerous methods have been proposed and achieved great performance on the supervised person ReID problem. However, these methods always rely on extensive human annotations of the whole training data, which are time-consuming and expensive to obtain. As a result, the demand of human annotation heavily limits the scalability of these methods as it is unpractical to provide the expensive annotations for all the large-scale datasets in the realistic applications [10], [11], [12]. Thus, how to reduce the annotation cost while preserving the performance remains challenging and has been studied extensively by researchers in recent years.

To tackle the above problem, several methods have been proposed and achieved great progress. These methods can be concluded as three categories. 1) Unsupervised person ReID, which aims to learn the feature representation of the dataset without any human annotations. However, it usually occurs a serious degeneration of the performance as it cannot learn more discriminative feature representations without any pairwise identity labels. 2) Semi-supervised person ReID, which makes the assumption that a subset of identities are fully labeled under different camera views. In this way, they can explore the remaining unlabeled dataset by taking advantage of existing annotations to further close the gap with supervised person ReID. Although great performance can be achieved, they are not practical in real applications as the cost of human annotations is still extremely high, which is comparable with the cost of supervised person ReID. 3) Active learning person ReID, which actively selects the most informative image pairs for annotation. Compared with the first two categories, active learning person ReID is more practical as the annotation cost is limited and controllable [13].

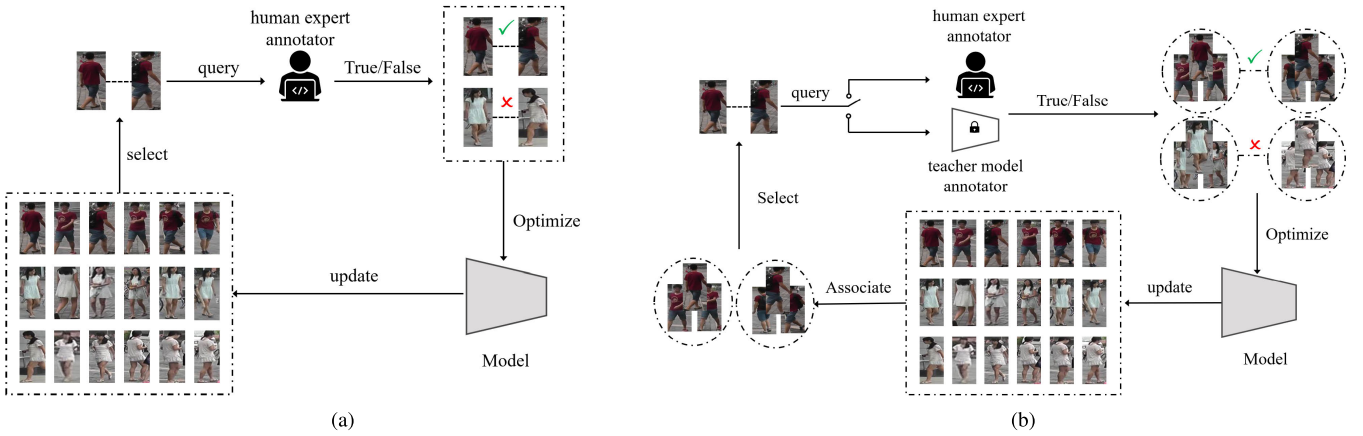



Fig. 1. Comparison of existing active learning person ReID framework and our proposed framework. (a) Existing active learning person ReID framework which solely relies on human experts for annotation. (b) Besides the human experts, a trained teacher model is introduced in our method which serves as a free annotator to deduce the extensive annotation cost.

Active learning person ReID frameworks usually select a set of image pairs actively as the candidate samples for annotation by human experts. Then these labeled samples will serve as the supervision to train the model in the next round [14], [15], [16], [17]. These steps are repeated iteratively until the expected annotation budget exhausted. Although existing active learning person ReID methods can achieve good performance with their carefully designed sampling strategy, the expected annotation cost of their methods is still non-negligible. They mainly have two problems. 1) Most of these methods merely rely on the labeled samples to supervise the training of the model, which is only a small subset of the whole dataset [18]. Compared with these labeled image pairs, a large number of unlabeled samples may contain more valuable information under exploration. 2) These methods only rely on human annotations, which are expensive to acquire. They send selected image pairs to human experts for annotation regard of their uncertainties, which will cause unnecessary extensive annotation cost.

To solve the above problems, we propose the tracklet-aware co-cooperative annotators for person ReID. The main motivation of our method is to reduce the cost of human annotations by taking the offline teacher model to serve as another free "annotator" to assist human experts with less uncertain image pairs as shown in Fig. 1. Specifically, by inheriting the merits of unsupervised person ReID, we first partition the dataset into different clusters and associate adjacent images in each cluster to form tracklets according to their frame numbers, which are easy to obtain in the dataset collection stage. As the same person tends to be captured in consecutive frames, if images in the same cluster are also captured in adjacent frames, then they are more likely to belong to the same identity. Then, center candidates of different tracklets in the same cluster are formed as the candidate samples for annotation by offline teacher model annotator and human expert annotator working in a co-cooperative scheme. Image pairs with less uncertainty are sent to the teacher model annotator for free annotation cost, while harder image pairs are sent to the human expert annotator for more accurate annotations. In this way, the total annotation cost can be reduced to the most extent while performance can be preserved. Although video-based person ReID also aims at matching the same person across video clips, it relies on the human labor or multiple objects tracking algorithm to generate image tracklet in the data collection stage. However, the quality of the generated tracklets can be a limiting factor for model learning if it is not satisfactory. In contrast, our proposed method utilizes only the accessible frame numbers of the cropped images, making it more scalable. Moreover, our approach generates tracklets directly from the model, and the model training stage and tracklet generation stage complement each other during the training process. A better model generates more accurate tracklets, which in turn can be used to train a better model.

The main contributions of our work can be summarized as follows.

1) To our knowledge, we are the first to utilize the information of frame number in active learning person ReID to reduce the annotation cost. To obtain more valuable image pairs, we propose the tracklet-aware sample selection strategy.
2) Based on the phenomenon that the trained model can also serve as an alternative annotator well, we propose the co-cooperative annotators for person ReID, which is first work to utilize the offline teacher model as the annotator to reduce the annotation cost from the human experts.
3) To relieve the influence of label noise in the training process, we propose the selective knowledge distillation (SKD) module to guide the model to learn from the unlabeled data in a more stable way.
4) Extensive experiments are conducted on three popular person ReID benchmarks. By combining the above three modules, we can achieve the state-of-the-art performance in both unsupervised person ReID task and active learning person ReID tasks.

Section II provides a review of related works, Section III introduces our proposed tracklet-aware co-cooperative annotators' framework, Section IV presents a theoretical analysis of our approach, Section V reports the experimental results, Section VI discusses the limitations and future directions

of our work, and Section VII concludes the article with a summary of our contributions and potential avenues for future research.

## II. Related Work

### A. Unsupervised Person ReID

Unsupervised person ReID aims to learn the feature representation directly from the unlabeled target dataset. These methods can be summarized as two categories: unsupervised domain adaptation (UDA) person ReID and purely unsupervised learning (USL) person ReID [27], [28], [29]. The former aims to transfer the knowledge from the annotated source domain to the unlabeled target domain, which is based on the assumption that the discrepancy between source domain and target domain is not significant [4], [30], [31]. The latter directly learns from the unlabeled target domain, thus is more scalable compared with the former. Here, we mainly concentrate on the latter as it is more related to our work.

In the USL person ReID, to fully exploit the unlabeled dataset, pseudo-labels are generated with existing clustering algorithm as the supervision to train the model. Thus, how to generate more accurate pseudo-labels is the key problem for USL person ReID. In [32], a bottom-up clustering framework is proposed to optimize the model and the relationship of samples in a joint way. PLM [33] proposes a novel progress learning method with a multiscale fusion network, which can directly exploit inference from the available abundant data without any annotations. To depress the influence of label noise in the isolated clustering process, GLT [34] proposes the group-aware label transfer algorithm to enhance the connection between pseudo-label generation process and feature representation learning process. To avoid the label noise accumulation, MMT [31] learns the feature representation through offline refined hard pseudo-labels and online refined soft pseudo-labels. ISE [35] 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. Existing state-of-the-art USL person ReID methods are mainly established on memory-based contrastive learning frameworks. These methods first utilize Kmeans [36] or DBSCAN [37] to generate pseudo-labels for the unlabeled training data. Then the model can be trained with contrastive learning and features of images stored in the memory bank. Specifically, SPCL [4] proposes a self-paced method which gradually creates more reliable clusters to refine the hybrid memory and learning targets. To solve the inconsistency problem in the memory update process, CCL [5] proposes the cluster contrast learning framework by taking advantage of cluster memory banks. As shown in Table I, our method stands out from the existing unsupervised person ReID methods by introducing a novel approach that leverages a reliable teacher model and readily accessible frame numbers to refine the pseudo-labels. By contrast, previous methods rely on the model itself to refine the pseudo-labels or features, leading to severe label noise in the early stages of training due to the substantial discrepancy between the pretrained parameters and the target dataset. Although HDCPD [25] employs an exponential moving average (EMA) updated teacher model, it still confronts this challenge, as their teacher model is also initialized with ImageNet pretrained parameters. In contrast, our approach utilizes a well-trained teacher model and easily obtained frame numbers to guide the student model in the fully unsupervised person ReID setting.

TABLE I
Comparison Between Existing Unsupervised Person ReID Methods and Our Proposed Method

| Method | Label Refinery | Frame Number | Teacher Model |
|---|---|---|---|
| SPCL [4] | ✓ | ✗ | ✗ |
| GCL [19] | ✗ | ✗ | ✗ |
| HCD [20] | ✓ | ✗ | ✗ |
| ICE [21] | ✗ | ✗ | ✗ |
| CCL [5] | ✗ | ✗ | ✗ |
| SECRET [22] | ✓ | ✗ | ✗ |
| MCRN [23] | ✓ | ✗ | ✗ |
| PPLR [24] | ✓ | ✗ | ✗ |
| HDCPD [25] | ✗ | ✗ | ✓ |
| Ours [26] | ✓ | ✓ | ✓ |

### B. Active Learning Person ReID

Active learning person ReID aims to select the most informative image pairs for annotation to reduce the annotation cost while preserving the performance of the model. Thus, how to design the sample selection strategy is the key problem for active learning person ReID. In [40], a novel early active learning algorithm is proposed to enforce the closeness of similar representations of instances with pairwise constraint. To reduce the annotation cost, [41] regards the annotation cost problem as the subset selection task and tries to optimize the problem to get the optimal subset of images pairs for annotation. ARR [42] proposes the uncertainty criterion to select informative samples with the prediction of the model for these samples. To further exploit the information of unlabeled samples, UCAL [43] proposes the unsupervised clustering active learning method which aims to combine unsupervised person ReID method with active learning to achieve a more satisfying performance. MASS [18] proposes the clustering purification-based active learning framework to select more valuable image pairs. SPAL [13] and [39] combine memory-based contrastive learning USL ReID frameworks with active learning and achieve great improvements.

Although other active learning methods also design selection criteria to select samples for annotation, in their methods all selected samples are sent to human experts for annotation, resulting in unnecessary annotation costs for hard samples and relatively hard samples with less uncertainty. Our proposed method aims to address this issue by carefully distinguishing between hard and relatively hard samples and utilizing a trained teacher model for less uncertain samples. This allows for relatively accurate, free annotations for the latter, thereby optimizing the use of limited annotation budgets. As shown in Table II, similar to SPAL [13] and AE [39], our method

TABLE II
COMPARISON BETWEEN EXISTING ACTIVE LEANING PERSON REID METHODS AND OUR PROPOSED METHOD

| Method | Clustering | Frame Number | Auxiliary Annotator | Annotation Efficient |
|---|---|---|---|---|
| HVIL [38] | ✗ | ✗ | ✗ | ✗ |
| DRAL [26] | ✗ | ✗ | ✗ | ✗ |
| MASS [18] | ✗ | ✗ | ✗ | ✗ |
| SPAL [13] | ✓ | ✗ | ✗ | ✗ |
| AE [39] | ✓ | ✗ | ✗ | ✓ |
| Ours [26] | ✓ | ✓ | ✓ | ✓ |

also applies the clustering algorithm to generate pseudo-labels. However, unlike them, we further reduce the manual annotation cost by utilizing easily obtained frame numbers and regarding the trained teacher model as an auxiliary free annotator. Furthermore, those methods whose annotation cost is lower than the dataset size are considered as annotation efficient. Compared to AE, our proposed method further reduces the annotation budget with the advantages mentioned above.

### C. Knowledge Distillation

Knowledge distillation aims to transfer the knowledge from the teacher model to the student model [44], [45], [46], [47], [48]. The original idea of knowledge distillation is to compress the large-scale teacher model to the smaller student model for computation efficiency. In recent years, increasingly more researchers focus on self-knowledge distillation, which keeps the teacher and student models with the same structure [49], [50], [51]. Specifically, these methods usually directly align the probability distribution or the feature representations of the teacher model with the student model. In [44], a simple baseline is proposed by utilizing the soft label of the teacher model. CS-KD [49] proposes the regularization which distills the prediction distributions of the similar samples of the same class in the training. Similar to our work, a probability distillation module is proposed in [25] to match the probability distribution between the network and the teacher network updated by EMA method. However, the teacher network updated in the online scheme is still limited in the feature representation and suffers from severe label noise [52]. Unlike these methods, we propose the SKD module, which selects some confident samples and uses them to learn from the teacher model in an offline scheme.

## III. METHOD

### A. Overview

Given N unlabeled images $X = \{x_1, x_2, \ldots, x_N\}$, their identity labels are denoted as $L = \{l_i\}_{i=1}^N$, which are unknown in advance. A human expert $\mathcal{E}_p$ can provide the accurate pairwise annotation in the training process. Given a pair of images $(I_i, I_j)$, the human expert can provide a binary label whether these two images belong to the same identity, i.e., $\mathcal{E}_p(i, j) = \mathbb{1}[l_i = l_j]$. Besides, a trained teacher model $f_{\theta'}$ is also available which serves as another annotator for free annotations. Unlike existing works, a set containing the frame numbers of these images $\mathcal{I} = \{I_i\}_{i=1}^N$ is also used in our setting, which represents which frames these cropped images belong to. The aim of active learning for person ReID is to train a discriminative model $f_\theta$ while reducing the annotation cost of human expert $\mathcal{E}_p$.

In order to achieve this objective, we propose the tracklet-aware co-cooperative annotators' framework. As illustrated in Fig. 2, our method aims to utilize the frame number information and the trained teacher model to relieve the excessive annotation cost of human expert. Given the unlabeled images $X$, the model $f_\theta$ is used to encode these images into feature vectors, then the existing clustering algorithm is applied to partition them into different clusters $C = \{c_1, c_2, \ldots, c_k\}$, where $k$ is the number of clusters. The tracklet-aware sample selection module will merge adjacent images in the same cluster to form tracklets according to their frame numbers. Then the center samples of different tracklets in the same cluster are formed as the candidates for annotation by human experts or free teacher model in the co-cooperative annotators' module. Finally, to fully utilize these annotated image pairs and unlabeled dataset, a hybrid loss is proposed, which includes contrastive learning, triplet loss, and the proposed SKD loss. The overall framework of our proposed method is illustrated in Algorithm 1, some detailed descriptions about these modules will be discussed in the following sections, and the notations are summarized in Sections III-B–III-E.

### B. Tracklet-Aware Sample Selection Strategy

Given the unlabeled images $X = \{x_1, x_2, \ldots, x_N\}$, their feature vectors can be extracted by the model $f_\theta$, then DBSCAN [37] clustering algorithm is utilized to partition these samples into different disjoint clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, where $k$ is the number of clusters and $c_i$ is the $i$th cluster which contains indexes of images belong to this cluster. DBSCAN is a density-based clustering algorithm that is capable of discovering clusters of arbitrary shape. Due to its efficiency and effectiveness, it has been extensively utilized in recent active learning and unsupervised person ReID studies, including AE [39], CCL [5], and ICE [21]. In our work, we also adopt DBSCAN for generating pseudo-labels. Compared with other popular clustering algorithms such as Kmeans, DBSCAN is more appropriate for unsupervised person ReID tasks because it does not require the number of clusters in advance and is capable of detecting clusters of varying shapes and sizes. Therefore, we consider DBSCAN to be more suitable for unsupervised person ReID tasks and hence utilize it in our approach. Then, for images in the same cluster, adjacent images are associated with form tracklets by their corresponding frame numbers $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, which can be obtained easily in the data collection stage. Our motivation is that as the same person tends to be captured in consecutive frames, if images in the same clustering are also cropped from adjacent frames, then those images are likely to have the same identity.

Specifically, for a cluster $C_t$, images in this cluster and their corresponding frame numbers can be obtained by the

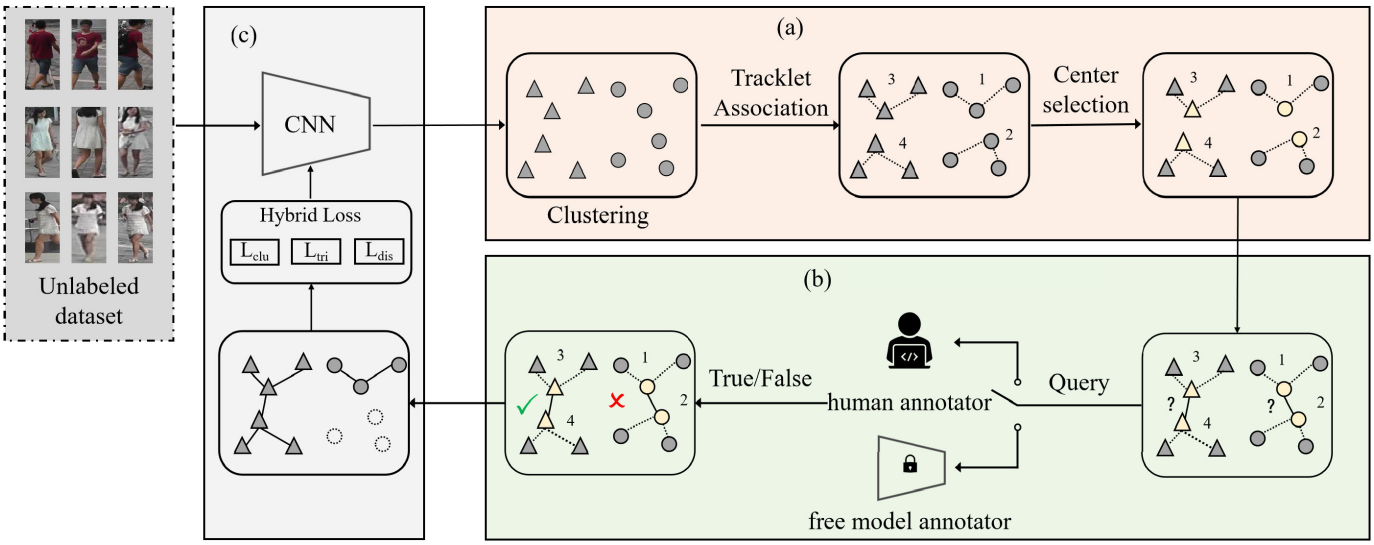



Fig. 2. Overview of the proposed tracklet-aware co-cooperative annotators' framework. (a) Tracklet-aware sample selection module aims to associate adjacent images into a tracklet and select the center sample for each tracklet. (b) Co-cooperative annotators' module aims to take advantage of the free teacher model and the human expert to judge whether different tracklets belong to the same identity by comparing center points. (c) Model is optimized with the proposed hybrid loss, which includes cluster contrast loss, triplet loss, and SKD loss.

ascending order as $X_t = \{x_i\}_{i \in C_t}$ and $\mathcal{I}_t = \{I_i\}_{i \in C_t}$, respectively, which satisfies $\mathcal{I}_t^i < \mathcal{I}_t^{i+1}$, where $\mathcal{I}_t^i$ is the $i$th element of $\mathcal{I}_t$. Then, images in the same cluster can be partitioned into different tracklets by associating adjacent images by their frame numbers. In this way, the tracklet indexes of $X_t$ can be denoted as $T_t = \{t_i\}_{i \in C_t}$, where $t_i$ represents the index of tracklet the $i$th sample belongs to, and their tracklet indexes satisfies the following constraint:

$$\mathbb{1}(T_t^i = T_t^{i+1}) = \begin{cases} 1, & \text{if } \left(\mathcal{I}_t^{i+1} - \mathcal{I}_t^i\right) < \lambda \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $T_t^i$ is the $i$th element of $T_t$, and $\lambda$ is a hyperparameter constrains whether a pair of images is adjacent, which is set to 300 in our experiments. In this way, if the margin of adjacent images is smaller than $\lambda$, then these images will be merged to the same tracklet. Finally, the center candidates of different tracklets in the same cluster are formed as queries for annotation, and the center candidate is selected as the sample whose feature representation is nearest to the mean feature representation of the tracklet. For simplicity, we omit the detailed description of this selection process.

### C. Co-Cooperative Annotators

For a cluster $C_t$, the center candidates $X'_t = \{x_{c_1}, x_{c_2}, \ldots, x_{c_m}\}$ could be obtained as described in Section III-B, where $c_i$ is the index of the center candidate and $m$ is the number of tracklets in this cluster. Then, a pair of center candidates $(x_{c_i}, x_{c_j})$ can be selected as a query for annotation. Unlike existing methods which solely rely on human experts for annotation, we propose to introduce the trained teacher model serving as another free annotator to reduce the extensive annotation cost. Thus, there are mainly two key problems to be resolved. 1) Given a pair of images, how to select the annotator for annotation. 2) Given the labeled image pairs annotated by human expert annotator and the free teacher model annotator, how to use these samples in the training process.

To solve the above problems, we propose a simple framework, which aims to take the human expert annotator and the free teacher model annotator working in a co-cooperative way. Intuitively, the model suffers from severe label noise at the beginning as it is initialized with parameters pretrained on ImageNet dataset, which has the significant discrepancy with person ReID datasets. As the model converges, it gains the ability of discriminating from different persons. Thus, some image pairs which are hard for the model to judge may be easily distinguished by a trained model, and it is unnecessary to annotate them by the expensive human expert. And, we propose a simple threshold-based mechanism to find these pairs by using the distance between the image pair as the criterion as follows:

$$\mathcal{E}(x_{c_i}, x_{c_j}) = \begin{cases} 1, & \text{if } d\left(f_{\theta'}(x_{c_i}), f_{\theta'}(x_{c_j})\right) < \omega - \delta \\ 0, & \text{if } d\left(f_{\theta'}(x_{c_i}), f_{\theta'}(x_{c_j})\right) > \omega + \delta \\ \mathcal{E}_p(x_{c_i}, x_{c_j}), & \text{otherwise} \end{cases} \tag{2}$$

where $\omega$ and $\delta$ are two hyperparameters which divide image pairs into different groups. $d(f_{\theta'}(x_{c_i}), f_{\theta'}(x_{c_j})) = |f_{\theta'}(x_{c_i}) - f_{\theta'}(x_{c_j})|_2^2$ and $f_{\theta'}(\cdot)$ is the output feature vector of the trained teacher model. Intuitively, since the trained teacher model gains the ability of discriminating from different person and achieves better performance compared with the initialized student model. We can resort to the trained teacher model and the human expert for help to relieve the severe label noise of the model. As shown in (2), for a pair of center candidates $(x_{c_i}, x_{c_j})$, feature vectors $(f_{\theta'}(x_{c_i}), f_{\theta'}(x_{c_j}))$ can be extracted from them by the teacher model $f_{\theta'}$. Specifically, given a pair of images, if the distance between them calculated by the teacher model is smaller/larger than $\omega - \delta/\omega + \delta$, then they will be judged as positive/negative pair by the teacher model with less uncertainly, otherwise we will resort to the human expert for accurate but expensive annotations. As the budget of human annotations is limited, when the budget exhausted,

**Algorithm 1** Training Process of the Student Model

**Require:** Unlabeled training data $X$
**Require:** Frame numbers $I$
**Require:** Initialize the encoder $f_\theta$ with ImageNet-pretrained ResNet-50
**Require:** The trained teacher encoder $f_{\theta'}$
**Require:** Balancing factors $\lambda_1$ and $\lambda_2$ for (7)

```
for n in [1,num_epochs] do
    Initial Ω+ and Ω−;
    Extract feature vector sets F from X by f_θ;
    Clustering F into clusters C with DBSCAN;
    Initialize memory dictionaries φ individually with
     the mean feature vector of each cluster;
    for each cluster C_t in C do
        Associate images from C_t into tracklets with
         (1) and frame numbers I, corresponding
         tracklet indexes are T_t = {t_i}_{i∈C_t};
        Obtain centers of tracklets
         X'_t = {x_{c_1}, x_{c_2}, ·, x_{c_m}} by selecting the sample
         nearest to the mean feature of its tracklet;
        Find the longest tracklet, and select its center
         x_{c_*} as the pivot;
        for i in [1, num_tracklets] do
            Obtain the annotation whether the ith
             center in X'_t and pivot x_{c_*} belong to the
             same identity according to (2);
            if they have the same identity then
                Add the ith tracklet to Ω+;
            else
                Add the ith tracklet to Ω−;
            end
        end
    end
    for iter in [1,num_iterations] do
        Sample a batch of hybrid samples from Ω+
         and Ω−;
        Compute objective function with (7) ;
        Update cluster feature vectors with (4);
    end
end
```

we will fully rely on the trained teacher model for annotations. Given a pair of images, when the distance between them calculated by the teacher model is smaller/larger than $\omega$, then they will be directly judged as positive/negative pair. In this way, all the pairs of center candidates can be annotated by our proposed selective annotation mechanism, and the annotations of other images are consistent with their corresponding center candidates.

In this way, the original pseudo-labeled dataset $(X, \tilde{Y})$ can be divided into positive dataset $\Omega^+ = (X^+, \tilde{Y}^+)$ and negative dataset $\Omega^- = (X^-, \tilde{Y}^-)$, where the former set contains samples whose annotations are consistent with the previous cluster, while the latter set contains samples whose annotations conflict with the previous cluster, $\tilde{Y}^+$ and $\tilde{Y}^-$ contain the previous cluster indexes. To learn the main pattern from the data, cluster contrast loss [5] is utilized to learn from the positive dataset. Specifically, the corresponding cluster contrast loss can be described as follows:

$$L_{\text{clu}} = \frac{1}{N_1} \sum_{(x_q, y_q) \in \Omega^+} -\log \frac{\exp\left(f_\theta(x_q) \cdot \phi_+ / \tau\right)}{\sum_{k=0}^{C} \exp\left(f_\theta(x_q) \cdot \phi_k / \tau\right)} \tag{3}$$

where $f_\theta(x_q)$ is the feature vector extracted by the student model, and $\phi_k$ is the centroid feature vector representing the $k$th cluster stored in the memory. $\phi_+$ is the centroid feature vector representing the cluster which $x_q$ belongs to. $\tau$ is the temperature hyperparameter, $C$ is the number of the clusters and $N_1$ is the number of samples evolved in the cluster contrast loss in a batch. Similar with [5], the centroid feature vector stored in the memory dictionary sets can be updated in the following way:

$$\phi_k = m\phi_k + (1 - m) f_\theta(x_q) \tag{4}$$

where $k$ is the index of the cluster query sample belongs to and $m$ is the momentum updating factor, which is set to 0.1 as the same as [5]. To further explore the information of the annotations, a triplet loss is used to mine the relationship among these samples. Specifically, we use samples belong to the same cluster to form the triplet as follows:

$$L_{tri} = \frac{1}{N_2} \sum_{\substack{(x_i, y_i), (x_j, y_j) \in \Omega^+ \\ (x_k, y_k) \in \Omega^- \\ y_i = y_j = y_k}} \max\left\{d\left(x_i, x_j\right) + m - d(x_i, x_k), 0\right\} \tag{5}$$

where $x_i$ and $x_i$ are sampled from the positive dataset while $x_k$ is sampled from the negative dataset, and these three samples belong to the same cluster. $m$ is the margin of the triplet loss and $N_2$ is the number of triplets in a batch. The triplet consists of samples from the same cluster, thus the model can fully explore the correlations of these hard samples and learn more discriminative feature representations.

### D. Selective Knowledge Distillation

As the model is trained with the generated pseudo-labels, the label noise will be accumulated in the training process as the model is initialized with parameters pretrained on ImageNet dataset at the beginning, which has the significant discrepancy with person ReID datasets. To relieve the issue, we propose the SKD module. Our motivation is that although the structure of the teacher model has no superiority over the student model, the trained model can achieve more accurate retrieval results than the initialized model, thus the feature representation of the teacher model can be utilized to guide the student model toward a more robust representation by converging fast. Some existing works also apply knowledge distillation in person ReID and achieved great progress [25], [53]. However, these methods directly align the probability distribution or the feature representations of the teacher model with the student model, which may hinder the learning process of the student model by excessively mimicking the behavior of the teacher model. If the teacher model is trained with the false pseudo-labels, then it will learn the biased feature representation, which will degenerate the performance of the

student model through knowledge distillation. To conquer the limitation while guiding the training process of the student model, we propose to select some confident samples to learn from the teacher model by combining knowledge distillation with our active learning framework as follows:

$$L_{\text{dis}} = \frac{1}{N_1} \sum_{(x_q, y_q) \in \Omega^+} \left| \frac{f_\theta(x_q)}{\left\| f_\theta(x_q) \right\|} - \frac{f_{\theta'}(x_q)}{\left\| f_{\theta'}(x_q) \right\|} \right|_2^2 \tag{6}$$

where $f_\theta(x_q)$ and $f_{\theta'}(x_q)$ are the feature vectors of $x_q$ extracted by the student model and the teacher model, respectively, $N_1$ is the number of samples evolved in the knowledge distillation in a batch. As those samples in $\Omega^-$ which are conflict with the previous clustering may also exist in the learning process of the teacher model and lead to biased feature representation. Thus, we only take samples in $\Omega^+$ for more confident knowledge distillation to expedite the convergence process of the student model.

### E. Final Objective Function

Given the unlabeled dataset $X$, pseudo-labels $\tilde{Y}$ can be obtained using DBSCAN [37] clustering algorithm. Then, we can use our proposed method to divide the whole dataset into positive dataset $\Omega^+ = (X^+, \tilde{Y}^+)$ and negative dataset $\Omega^- = (X^-, \tilde{Y}^-)$, where the former set contains samples whose annotations are consistent with the previous cluster, while the latter set contains samples whose annotations conflict with the previous cluster. To make the model fully explore the inherent structure of the unlabeled samples and utilize annotations, we propose the hybrid loss, which consists of the cluster contrast loss, triplet loss, and the knowledge distillation loss as follows:

$$\begin{aligned} L_{\text{clu}} = & \frac{1}{N_1} \sum_{(x_q, y_q) \in \Omega^+} -\log \frac{\exp\left(f_\theta(x_q) \cdot \phi_+ / \tau\right)}{\sum_{k=0}^{C} \exp\left(f_\theta(x_q) \cdot \phi_k / \tau\right)} \\ & + \frac{\lambda_1}{N_1} \sum_{(x_q, y_q) \in \Omega^+} \left| \frac{f_\theta(x_q)}{\left\| f_\theta(x_q) \right\|} - \frac{f_{\theta'}(x_q)}{\left\| f_{\theta'}(x_q) \right\|} \right|_2^2 \\ & + \frac{\lambda_2}{N_2} \sum_{\substack{(x_i, y_i), (x_j, y_j) \in \Omega^+ \\ (x_k, y_k) \in \Omega^- \\ y_i = y_j = y_k}} \max\left\{ d(x_i, x_j) + m - d(x_i, x_k), 0 \right\} \end{aligned} \tag{7}$$

where $\lambda_1$ and $\lambda_2$ are balancing factors, $N_1$ and $N_2$ are numbers of individual samples and triplets of the dataset. To make it practical with the training batch manner, for each batch we sample 160 individual samples and 32 triplets, thus the batch size is 256, which is the same with [5], i.e., $N_1 = 160$ and $N_2 = 32$ for each batch. More training details can refer to Section V-B. The teacher model is trained in advance following existing CCL framework, due to limited page we omit the description of this part and interested readers can refer to [5] for more details. In the experiment, we also regard CCL as the baseline for fair comparison. It is noteworthy that we assume that the trained teacher model is more accurate than the initialized student model. Thus, our proposed SKD module aims to utilize the trained teacher model to guide the student model to relieve label noise. Although the teacher model performs worse than the student model and may even hinder the student model learning in the latter period of the training process, we add the hyperparameter $\lambda_1$ in (7) to determine the weight of the guidance of the teacher model. While learning from the teacher model, the student model is also encouraged to explore by itself. These two phases are balanced by the hyperparameter $\lambda_1$ to relieve the negative influence of the teacher model to some extent. Therefore, it is reasonable that the student model will perform better than the teacher model when it converges. On the other hand, as those samples whose annotations conflict with the previous clustering may also exist in the learning process of the teacher model and lead to biased feature representation, we only take confident samples for more reliable knowledge distillation to guide the student model toward faster convergence in the training process. As it is hard to determine the turning point from which the teacher model disturbs the student model due to lack of annotated labels, we leave it in the future work, e.g., by designing some measures in the unsupervised setting [54].

Besides, our proposed framework utilizes a clustering algorithm to generate pseudo-labels, which serves as the primary supervision signal in the optimization process. To improve the quality of the pseudo-labels, we employ the cooperative annotators' framework to provide annotations for center candidate pairs, and any samples whose annotations conflict with the previous clusters are treated as negative samples. To leverage these negative samples, we incorporate them into the triplet loss of the final objective function. Given the limited annotation budget, we choose not to reassign these samples to the new cluster as it would result in additional annotation costs. Although further exploiting the labels of these negative samples would be beneficial, it is challenging due to the budget constraints, and we leave it in the future work.

## IV. Theoretical Analysis

The aim of person ReID aims to learn a feature classifier $f$ with the dataset to minimize the following risk:

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[\ell(f(x), y)\right] \tag{8}$$

where $\ell$ is the loss function and $\mathbb{D}$ is the distribution of the dataset, $\nabla\ell$ is bounded. In our proposed approach, although the number of clusters varies before each epoch, we can obtain cluster centers by computing the average of feature vectors in every cluster. This process enables us to use the model as a classifier by combining the encoder and cluster centers. More specifically, when presented with a query sample, the class distribution can be obtained by applying a softmax activation function to the dot-product between the feature vector of the query sample and the cluster centers. Thus, even though the model $f_\theta$ is primarily a feature extractor that maps input images to feature vectors, it can be considered a classifier $f$ due to the use of cluster centers. In the unsupervised person ReID, the ground-truth of the dataset is unavailable, thus the pseudo-labels are generated to construct the noisy dataset and

the actual learning objective becomes

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathbb{D}}}\big[\ell\big(f(x),\tilde{y}\big)\big]. \tag{9}$$

Due to the existence of the pseudo-labels, the distribution of the dataset has been corrupted from $\mathbb{D}$ to $\tilde{\mathbb{D}}$. Let $\delta$ and $\tilde{\delta}: X \to \Delta_K$ denote the class probability functions of clean dataset distribution and noisy dataset distribution, respectively, where $\Delta_K$ denotes the $K-$simplex. Then, $f(x) = \arg\max_{1\le i\le K}\delta_i$ and $\tilde{f}(x) = \arg\max_{1\le i\le K}\tilde{\delta}_i$ can be easily obtained. Assume a transition matrix $T$ exists which works as $T_{ij}(x) = \mathbb{P}(\tilde{y} = j \mid y = i, x), \forall x \in X$, we have

$$\begin{aligned}\tilde{\delta}_i(x) &= \mathbb{P}\big(\tilde{Y} = i \mid X = x\big)\\ &= \sum_j \mathbb{P}\big(\tilde{Y} = i \mid Y = j, X = x\big)\mathbb{P}(Y = j \mid X = x)\\ &= \sum_j T_{ji}(x)\delta_j(x)\\ &= T_i^{\top}(x)\delta(x) \quad \forall x \sim X.\end{aligned} \tag{10}$$

Thus, we have $\tilde{\delta}(x) = T^{\top}(x)\delta^s(x)$ and it can be directly derived that

$$\delta(x) = \big(T^{\top}(x)\big)^{-1}\tilde{\delta}(x) \tag{11}$$

if $T^{\top}(x)$ is non-invertible, the Moore–Penrose pseudo-left inverse $(T(x)T^{\top}(x))^{-1}T(x)$ can be used here. Then, we will provide a risk bound on clean dataset distribution for corrected model $f(x) = \arg\max_{1\le i\le K}((T^{\top}(x))^{-1}\tilde{\delta}(x))_i$, which is learned on noisy dataset distribution. For simplicity, we define $\mathcal{I}(\delta(x), y) := \ell(f(x), y) = \ell(\arg\max_{1\le i\le K}\delta_i(x), y)$. In our active learning setting, for each learning iteration, we relabel the examples depend on the human annotations and teacher model's outputs, and the assumed transition matrix changed accordingly. We thus interest in how our active learning strategy influencing the student model learning process. We first give a regret risk of the learned model about the transition matrix.

*Theorem 1:* Let $f^* = \arg\min_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mathbb{D}}[\ell(f(x), y)]$ with $\|\nabla\mathcal{I}\|_2 \le L$. For any $f = \arg\max_{1\le i\le K}\delta_i \in \mathcal{F}$ learned on noisy dataset distribution, we have

$$\mathcal{R}(f) \le \mathcal{R}\big(f^*\big) + L\cdot\mathbb{E}_{\mathbb{D}_X}\big\|T^{\top}(x)^{-1}\big\|_2\big\|\tilde{\delta}(x) - \tilde{\delta}^*(x)\big\|_2 \tag{12}$$

where $\tilde{\delta}(x) = T^{\top}(x)\delta(x)$ and $\tilde{\delta}^*(x) = T^{\top}(x)\delta^*(x)$.

*Proof:*

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}\big(f^*\big) &= \mathbb{E}_{\mathbb{D}}\big[\ell(f(x), y) - \ell\big(f^*(x), y\big)\big]\\ &= \mathbb{E}_{\mathbb{D}}\big[\mathcal{I}(\delta(x), y) - \mathcal{I}\big(\delta^*(x), y\big)\big]\\ &= \mathbb{E}_{\mathbb{D}}\Big[\mathcal{I}\Big(\big(T^{\top}(x)\big)^{-1}\tilde{\delta}(x), y\Big)\\ &\qquad - \mathcal{I}\Big(\big(T^{\top}(x)\big)^{-1}\tilde{\delta}^*(x), y\Big)\Big]\\ &= \mathbb{E}_{\mathbb{D}}\Big[\nabla\mathcal{I}(\xi, y)\Big(\big(T^{\top}(x)\big)^{-1}\tilde{\delta}(x)\\ &\qquad - \big(T^{\top}(x)\big)^{-1}\tilde{\delta}^*(x)\Big)\Big]\\ &\le \mathbb{E}_{\mathbb{D}}\|\nabla\mathcal{I}(\xi, y)\|_2\big\|T^{\top}(x)^{-1}\big\|_2\big\|\tilde{\delta}(x) - \tilde{\delta}^*(x)\big\|_2\\ &\le L\cdot\mathbb{E}_{\mathbb{D}_X}\big\|T^{\top}(x)^{-1}\big\|_2\big\|\tilde{\delta}(x) - \tilde{\delta}^*(x)\big\|_2\end{aligned} \tag{13}$$

where $\xi \in [0, 1]^K$ is the intermediate value. Hence, we prove this theorem. □

*Remark 1:* Theorem 1 indicates that the regret risk of the learned corrected model on clean dataset distribution $\mathbb{D}$ is bounded by $L\cdot\mathbb{E}_{\mathbb{D}_X}\|T^{\top}(x)^{-1}\|_2\|\tilde{\delta}(x) - \tilde{\delta}^*(x)\|_2$. The model is pretrained on ImageNet dataset, which has the significant discrepancy with ImageNet dataset. Thus, it suffers from severe label noise in the start training stages. Our carefully designed active learning module selects hard image pairs for accurate annotations that those most confusing samples are sent to human experts while relatively easier samples are resorted to the trained teacher model. Thus, reliable annotations could be obtained in this way. By taking advantage of them in the training process, the student model could obtain more discriminative feature representations for these confusing samples. Therefore, the probabilistic value of the transition matrix can be more concentrated on the certain possible classes, $\|T^{\top}(x)\|_2$ will increase and $\|T^{\top}(x)^{-1}\|_2$ will decrease as those irrelevant classes will bother less in the transition matrix $T$. As a result, the regret risk of the learned model on clean dataset distribution will also decrease accordingly.

TABLE III
STATISTICS OF THREE PERSON REID DATASETS USED IN OUR EXPERIMENTS

| Datasets | Cameras | Training | | Testing | |
|---|---|---|---|---|---|
| | | IDs | Images | Query | Gallery |
| Market-1501 [63] | 6 | 751 | 12,936 | 3,368 | 19,732 |
| DukeMTMC-reID [64] | 8 | 702 | 16,522 | 2,228 | 17,661 |
| MSMT17 [65] | 15 | 1,041 | 32,621 | 11,659 | 51,027 |

## V. EXPERIMENT

### A. Datasets and Evaluation Protocol

We conduct our experiments on three public benchmarks, including Market-1501 [63], DukeMTMC-reID [64], and MSMT17 [65]. Market-1501 dataset includes 32 668 images of 1501 IDs captured by six different cameras. DukeMTMC-reID dataset is another large-scale person ReID dataset, which includes 36 441 images of 702 IDs captured by eight different cameras. While MSMT17 dataset includes 126 441 images of 1041 IDs captured by 15 different cameras. These three datasets are widely used the person ReID task and the details of these datasets are described in Table III.

Following existing person ReID works [4], [5], [63], we adopt the mean average precision (mAP) and cumulated matching characteristics (CMC) as the evaluation metrics. In the CMC evaluation metrics, we report Top-1, Top-5, and Top-10 in the result. For fair comparison, no post-processing technique is adopted in our experiment. Similar with other person ReID works, in the unsupervised setting, only unlabeled target dataset is used to train our model. While in the active learning setting, the amount of manual pairwise annotations is calculated as the budget.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID. TA AND PA REPRESENT OUR PROPOSED TRACKLET-AWARE FREE TEACHER MODEL ANNOTATOR AND HUMAN ANNOTATOR. $N$ IS THE SIZE OF THE DATASET, WHICH IS 12 936 AND 16 522 IN MARKET-1501 AND DUKEMTMC-REID, RESPECTIVELY

| Method | Reference | Market-1501 | | | | | DukeMTMC-reID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R10 | Budget | mAP | R1 | R5 | R10 | Budget |
| Unsupervised Person ReID | | | | | | | | | | | |
| SSL [55] | CVPR'20 | 37.8 | 71.7 | 83.8 | 87.4 | 0 | 28.6 | 52.5 | 63.5 | 68.9 | 0 |
| JVTC [56] | ECCV'20 | 41.8 | 72.9 | 84.2 | 88.7 | 0 | 42.2 | 67.6 | 78.0 | 81.6 | 0 |
| MMCL [57] | CVPR'20 | 45.5 | 80.3 | 89.4 | 92.3 | 0 | 40.2 | 65.2 | 75.9 | 80.0 | 0 |
| HCT [58] | CVPR'20 | 56.4 | 80.0 | 91.6 | 95.2 | 0 | 50.7 | 69.6 | 83.4 | 87.4 | 0 |
| CycAs [59] | ECCV'20 | 64.8 | 84.8 | - | - | 0 | 60.1 | 77.9 | - | - | 0 |
| GCL [19] | CVPR'21 | 66.8 | 87.3 | 93.5 | 95.5 | 0 | 62.8 | 82.9 | 87.1 | 88.5 | 0 |
| SPCL [4] | NeurIPS'20 | 73.1 | 88.1 | 95.1 | 97.0 | 0 | 65.3 | 81.2 | 90.3 | 92.2 | 0 |
| HCD [20] | ICCV'21 | 78.1 | 91.1 | 96.4 | 97.7 | 0 | 65.6 | 79.8 | 88.6 | 91.6 | 0 |
| ICE [21] | ICCV'21 | 79.5 | 92.0 | 97.0 | 98.1 | 0 | 67.2 | 81.3 | 90.1 | 93.0 | 0 |
| CCL [5] | ACCV'22 | 82.6 | 93.0 | 97.0 | 98.1 | 0 | 72.8 | 85.7 | 92.0 | 93.5 | 0 |
| MCRN [23] | AAAI'22 | 80.8 | 92.5 | - | - | 0 | 69.9 | 83.5 | - | - | 0 |
| SECRET [22] | AAAI'22 | 81.0 | 92.6 | - | - | 0 | 63.9 | 77.9 | - | - | 0 |
| PPLR [24] | CVPR'22 | 81.5 | 92.8 | 97.1 | 98.1 | 0 | - | - | - | - | - |
| HDCPD [25] | TIP'22 | 84.5 | 93.5 | 97.6 | **98.6** | 0 | 73.5 | 85.4 | 92.2 | 94.5 | 0 |
| Ours (TA) | - | **86.2** | **94.6** | **98.0** | **98.6** | 0 | **75.1** | **86.0** | **92.6** | **94.7** | 0 |
| Active Learning Person ReID | | | | | | | | | | | |
| QIC [60] | SIGIR'94 | 45.0 | 67.8 | 85.7 | 91.1 | - | 36.8 | 56.8 | 74.2 | 79.3 | - |
| QBC [61] | ICML'98 | 46.3 | 68.4 | 86.1 | 91.2 | - | 40.8 | 61.1 | 77.4 | 82.4 | - |
| GD [62] | CVPR'12 | 49.3 | 71.4 | 87.1 | 91.4 | - | 33.6 | 53.5 | 70.0 | 75.8 | - |
| HVIL [38] | ECCV'16 | - | 78.0 | - | - | - | - | - | - | - | - |
| DRAL [26] | ICCV'19 | 66.3 | 84.2 | 94.3 | 96.6 | 10N | 56.0 | 74.3 | 84.8 | 88.4 | 10N |
| MASS [18] | MM'21 | 81.7 | 93.5 | 97.9 | 98.6 | 5N | 72.9 | 86.1 | **93.9** | **95.9** | 5N |
| AE [39] | Arxiv'22 | 84.1 | 93.3 | 97.4 | - | **1000** | - | - | - | - | - |
| AE [39] | Arxiv'22 | 85.6 | 93.6 | 97.7 | 98.4 | 3000 | 75.3 | **86.9** | 92.8 | 94.7 | 3000 |
| Ours (TA+PA) | - | **86.8** | **94.7** | **98.0** | **98.8** | **1000** | **75.8** | 86.4 | 93.1 | 95.0 | **1000** |

### B. Implementations Details

We use the Resnet-50 [66] initialized with the parameters pretrained on the ImageNet [67] as the backbone encoder. Following existing ReID framework [5], all sub-module layers after layer-4 are removed and a GEM pooling is added followed by batch normalization layer [68] and L2-normalization layer. During training, we use the DBSCAN [37] as clustering algorithm to generate pseudo-labels before each epoch.

For training, each mini-batch contains 256 images, which includes 160 individual samples and 32 triplets, which are resized as $256 \times 128$. 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 factors $\lambda_1$ and $\lambda_2$ in (7) are set to 0.5 and 0.2, respectively. To select the appropriate annotator, $\omega$ in (2) is set to 0.4 for Market-1501 and DukeMTMC-reID while 0.8 for MSMT17, $\delta$ is set to 0.15 for all these datasets. For DBSCAN clustering algorithm, the minimal number of neighbors is set to 4 and the maximum distance $d$ is set to 0.6 for Market-1501 and DukeMTMC-reID while 0.7 for MSMT17. To associate adjacent images into the same tracklet, we set $\lambda$ in (1) to 300 in the experiment. As the frame number information is not provided in the MSMT17 dataset, we regard each image as a single tracklet. To make full use of the annotations of the human expert, we begin to provide manual annotations after ten epochs, and the triplets in (7) are sampled from images annotated by the human expert. As the student model is initialized with parameters pretrained on ImageNet, it may produce low-quality center pair candidates for annotation due to the large discrepancy between ImageNet and person ReID datasets. If we apply manual annotations on these candidates at the beginning of the training procedure, less informative samples can be mined, and some of the valuable manual annotation budget may be wasted. Therefore, to make better use of the valuable annotations of the human expert, we utilize them after ten epochs to mine more informative image pairs. By this way, the student model can leverage the knowledge learned from the unsupervised training stage to generate higher quality center pair candidates, which can be further refined with the manual annotations.

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSMT17. TA AND PA REPRESENT OUR PROPOSED TRACKLET-AWARE FREE TEACHER MODEL ANNOTATOR AND HUMAN ANNOTATOR. $N$ IS THE SIZE OF THE DATASET, WHICH IS 32 621 IN MSMT17

| Method | Reference | MSMT17 | | | | |
|---|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R10 | Budget |
| Unsupervised Person ReID | | | | | | |
| JVTC [56] | ECCV'20 | 15.1 | 39.0 | 50.9 | 56.8 | 0 |
| MMCL [57] | CVPR'20 | 11.2 | 35.4 | 44.8 | 49.8 | 0 |
| CycAs [59] | ECCV'20 | 26.7 | 50.1 | - | - | 0 |
| GCL [19] | CVPR'21 | 21.3 | 45.7 | 58.6 | 64.5 | 0 |
| SPCL [4] | NeurIPS'20 | 19.1 | 42.3 | 55.6 | 61.2 | 0 |
| HCD [20] | ICCV'21 | 26.9 | 53.7 | 65.3 | 70.2 | 0 |
| ICE [21] | ICCV'21 | 29.8 | 59.0 | 71.7 | 77.0 | 0 |
| MCRN [23] | AAAI'22 | 31.2 | 63.6 | - | - | 0 |
| SECRET [22] | AAAI'22 | 31.3 | 60.4 | - | - | 0 |
| PPLR [24] | CVPR'22 | 31.4 | 61.1 | 73.4 | 77.8 | 0 |
| HDCPD [25] | TIP'22 | 24.6 | 50.2 | 61.4 | 65.7 | 0 |
| CCL [5] | Arxiv'21 | 33.3 | 63.3 | 73.7 | 77.8 | 0 |
| Ours (TA) | - | **35.2** | **64.2** | **74.6** | **78.3** | 0 |
| Active Learning Person ReID | | | | | | |
| QIC [60] | SIGIR'94 | 21.5 | 31.3 | 52.6 | 59.8 | - |
| QBC [61] | ICML'98 | 24.9 | 39.4 | 57.0 | 61.3 | - |
| GD [62] | CVPR'12 | 25.4 | 42.5 | 58.2 | 63.6 | - |
| MASS [18] | MM'21 | 30.0 | 54.1 | 65.4 | 70.4 | 5N |
| AE [39] | Arxiv'22 | 35.5 | 63.4 | 74.5 | 79.0 | 3000 |
| Ours (TA+PA) | - | **36.5** | **65.5** | **75.9** | **79.8** | **1000** |

TABLE VI
ABLATION STUDY ON MARKET-1501

| Method | Market-1501 | | | |
|---|---|---|---|---|
| | mAP | R1 | R5 | R10 |
| Baseline | 83.0 | 92.7 | 97.1 | 98.2 |
| Baseline+TA | 85.7 | 94.2 | 97.9 | 98.6 |
| Baseline+TA+PA | 86.3 | 94.3 | **98.0** | **98.8** |
| Baseline+TA+PA+SKD | **86.8** | **94.7** | **98.0** | **98.8** |

In our proposed co-cooperative annotators' framework, the issue of redundant annotations may arise due to the appearance of hard samples annotated by the human expert in the subsequent rounds. Our approach to address this issue involves two strategies. First, we incorporate the samples annotated by human experts into the training process and use them to construct the triplet loss in the final objective function. By doing so, the model can learn the main patterns from these samples, guided by the accurate annotations, which may result in these samples being regarded as more confident samples. Second, we use annotations from human experts only every five epochs, to generate different results and reduce the possibility of redundant annotations.

### C. Comparison With State-of-the-Arts

We compare our proposed method with the state-of-the-art fully unsupervised person ReID methods and active learning person ReID methods on three public person ReID datasets in Tables IV and V. Fully unsupervised methods include SSL [55], JVTC [56], MMCL [57], HCT [58], CycAs [59], GCL [19], SPCL [4], HCD [20], ICE [21], CCL [5], MCRN [23], SECRET [22], PPLR [24], and HDCPD [25]. By taking advantage of our proposed tracklet-aware free teacher model (TA), our method can achieve the best results on these three datasets in the fully unsupervised setting. Specifically, compared with CCL, we can achieve the improvements of 3.6%, 2.3%, and 1.9% in terms of mAP on Maret-1501, DukeMTMC-reID, and MSMT17, respectively.

Compared with the state-of-the-art active learning person ReID methods, our method also achieves better performance. These methods include QIC [60], QBC [61], GD [62], HVIL [38], DRAL [26], MASS [18], and AE [39]. As shown in Tables IV and V, our method can outperform all of them on these three datasets in the active learning setting with fewer annotations. Compared with the fully unsupervised setting, we can further improve the performance of our method with very limited annotations. Compared with the state-of-the-art active learning ReID framework AE [39], we can achieve competitive performance with fewer labels annotated by human experts. The reason is probably that existing active learning frameworks utilize their designed criteria to select uncertain samples and send all of them to the human expert for annotations. Thus, these methods have to afford extensive annotation cost as those relatively hard samples but with less uncertainty will also consume the annotation budget equally. Unlike these active learning frameworks, our proposed method further distinguishes these samples by our carefully designed selection criteria. For those hard samples, we will send them to the human expert for more accurate but expensive annotations. For those relatively hard but with less uncertainty, we will resort to the trained teacher model for relatively accurate but free annotations. In this way, we can improve the performance of the model with very limited annotation budget. Furthermore, we acknowledge that our proposed method achieved surprising improvements on Market-1501 and DukeMTMC-reID datasets, but the performance on MSMT17 dataset is relatively limited compared with state-of-the-art methods. The reason for this could be attributed to the fact that the MSMT17 dataset does not provide frame numbers, which limits our ability to associate adjacent images and produce robust tracklets. Instead, we have to treat each image as a single tracklet, which could lead to reduced performance. Nonetheless, we found that our proposed method can still achieve significant improvement on the MSMT17 dataset without the frame number information.

### D. Ablation Studies

In this section, we study effectiveness of different components and hyperparameters in our proposed method. Our method is implemented based on the CCL baseline [5], and hyperparameters introduced in our work include hyperparameters $\lambda_1$, $\lambda_2$, $m$ in (7), and $\omega$, $\delta$ in (2).

*1) Different Combinations of the Components:* Our method can be regarded as a combination of three modules, including

TABLE VII
ABLATION STUDY ON MSMT17

| Method | MSMT17 | | | |
|---|---|---|---|---|
| | mAP | R1 | R5 | R10 |
| Baseline | 33.0 | 63.0 | 73.7 | 77.7 |
| Baseline+TA | 35.1 | 64.3 | 74.7 | 78.5 |
| Baseline+TA+PA | 35.7 | 64.7 | 75.3 | 79.2 |
| Baseline+TA+PA+SKD | **36.5** | **65.5** | **75.9** | **79.8** |

TABLE VIII
IMPACT OF HYPERPARAMETER $\lambda_1$ ON MARKET-1501

| $\lambda_1$ | Market-1501 | | | |
|---|---|---|---|---|
| | mAP | R1 | R5 | R10 |
| 0.0 | 86.3 | 94.3 | **98.0** | **98.8** |
| 0.5 | **86.8** | **94.7** | **98.0** | **98.8** |
| 1.0 | 86.6 | 94.5 | 97.9 | **98.8** |
| 1.5 | 86.3 | 94.4 | 97.9 | **98.8** |
| 2.0 | 85.9 | 94.1 | 97.7 | 98.6 |
| 2.5 | 85.7 | 94.1 | 97.7 | 98.6 |

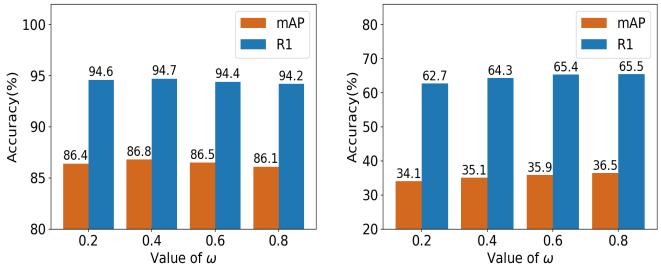
TABLE IX
IMPACT OF HYPERPARAMETER $\lambda_2$ ON MARKET-1501

| $\lambda_2$ | Market-1501 | | | |
|---|---|---|---|---|
| | mAP | R1 | R5 | R10 |
| 0.0 | 86.4 | 94.0 | 97.9 | 98.8 |
| 0.1 | 86.7 | 94.2 | 98.0 | 98.6 |
| 0.2 | **86.8** | **94.7** | 98.0 | 98.8 |
| 0.3 | **86.8** | 94.3 | **98.2** | **98.9** |
| 0.4 | 86.8 | 94.3 | 98.0 | 98.6 |
| 0.5 | 86.6 | 94.5 | 98.1 | 98.8 |

tracklet-aware teacher model annotator (TA), tracklet-aware human expert annotator (PA), and SKD. As our work is implemented on the CCL [5], we take CCL as a baseline in our experiment and the result is shown in Tables VI and VII. As shown in tables, the first line means the performance of CCL on Market-1501 and MSMT17 datasets, CCL can achieve good performance by taking advantage of contrastive learning and cluster memory, but it is still limited by the label noise introduced in the clustering stage. The second line is the result of the combination of CCL and our proposed TA module, compared with the first line we can find that our proposed TA module can improve the baseline by 2.7%, 2.1% in terms of mAP on Market-1501 and MSMT17 datasets, which indicates that our introduced trained teacher model can serve as a free annotator and help the student model relieve the severe label noise. The third line is the result of the combination of CCL, TA, and our proposed PA modules, compared with the first line, improvements of 3.3% and 2.7% in terms of mAP on Market-1501 and MSMT17 datasets can be further achieved by taking advantage of free teacher model and very limited human annotations. The last line denotes the result of the combination of CCL, TA, PA, and SKD modules, compared with the first line, improvements of 3.8% and 3.5% can be achieved by our proposed three modules. The result shows that our proposed three modules can work in a mutual benefit way and the baseline with these three modules can achieve the best performance.

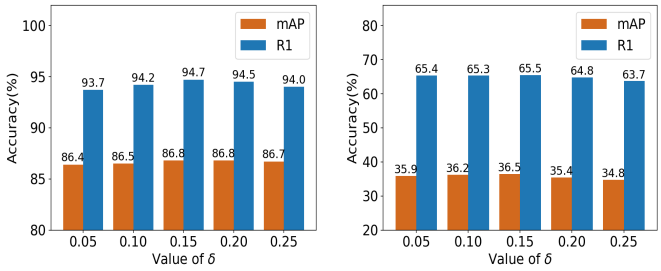
TABLE X
IMPACT OF HYPERPARAMETER $m$ ON MARKET-1501

| $m$ | Market-1501 | | | |
|---|---|---|---|---|
| | mAP | R1 | R5 | R10 |
| 0.0 | 86.6 | 94.5 | 98.0 | 98.7 |
| 0.1 | 86.7 | 94.2 | 98.0 | 98.7 |
| 0.2 | 86.8 | 94.6 | **98.1** | **98.8** |
| 0.3 | 86.8 | **94.7** | 98.0 | **98.8** |
| 0.4 | **86.9** | 94.2 | **98.1** | **98.8** |
| 0.5 | 86.6 | 94.2 | 97.9 | 98.7 |




(a) (b)

Fig. 3. Impact of hyperparameter $\omega$ on Market-1501 and MSMT17. (a) Hyperparameter $\omega$ on Market-1501. (b) Hyperparameter $\omega$ on MSMT17.




(a) (b)

Fig. 4. Impact of hyperparameter $\delta$ on Market-1501 and MSMT17. (a) Hyperparameter $\delta$ on Market-1501. (b) Hyperparameter $\delta$ on MSMT17.

TABLE XI
PERFORMANCE OF OUR METHOD ON MARKET-1501 AND DUKEMTMC-REID WITH DIFFERENT SOURCE DATASETS PRETRAINED MODELS

| Source Dataset | Budget | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|---|
| | | mAP | R1 | mAP | R1 |
| ImageNet | 0 | 86.2 | 94.6 | 75.1 | 86.0 |
| | 1000 | **86.8** | **94.7** | **75.8** | **86.4** |
| LUPerson-NL | 0 | 87.9 | 94.7 | 78.1 | 87.1 |
| | 1000 | **88.3** | **95.1** | **78.3** | **87.9** |

*2) Impact of Hyperparameters in the Objective Function:* We analyze the effect of hyperparameters $\lambda_1$, $\lambda_2$, and $m$ in (7). Tables VIII–X show the performance of our method over different values of $\lambda_1$, $\lambda_2$, and $m$ on Market-1501, respectively. Specifically, when $\lambda_1$ increases, the performance


Fig. 5. Top 6 retrieval results of some hard queries on Market-1501 dataset. Note that the green/red boxes denote true/false retrieval results, respectively.

of our method degenerates slightly as it relies more on the output of the trained teacher model. When hyperparameters $\lambda_2$ and $m$ change, our method shows similar performance, which shows our method is robust against these two hyperparameters. In the experiment, we set $\lambda_1$, $\lambda_2$, and $m$ to 0.5, 0.2, and 0.3 for all three datasets and both unsupervised and active learning settings.

*3) Impact of Hyperparameters in the Sample Selection:* $\omega$ and $\delta$ in (2) are two hyperparameters determining which samples are selected as positive/negative or hard/simple samples for annotation. As shown in Fig. 3, $\omega$ is a key factor which determines the performance of our method. When $\omega$ increases/decreases, less/more samples will be regarded as positive samples for training, and our method shows different patterns with different values of $\omega$. In the experiment, we set $\omega$ to 0.4 for Market-1501 and DukeMTMC-reID while 0.8 for MSMT17. As shown in Fig. 4, our method achieves similar results on Market-1501 and MSMT17 when $\delta$ changes in the appropriate range, which shows our method is robust against $\delta$. In the experiment, we set $\delta$ to 0.15 for all datasets.

*4) Pretrain the Model With Different Source Datasets:* Due to lack of ground truth, most active learning/unsupervised person ReID methods utilized clustering algorithm to generate pseudo-labels for the dataset. Then the generated pseudo-labels are used to train the model. Pretraining on ImageNet is the key to the success of unsupervised person ReID methods, which can guarantee that the model can discover general patterns in the pseudo-labels generation process. If the encoder is initialized randomly, then it is hard to train the model as the generated pseudo-labels are very noisy.

As pretraining is significant for active learning/unsupervised person ReID, a better pretrained model can also boost the performance of the trained model. Recently, some works [70], [71], [72] use larger person ReID dataset, such as LUPerson and LUPerson-NL, to pretrain the model and achieve better performance on downstream ReID tasks than the model pretrained on ImageNet. We believe that using more advanced pretrained model can further improve our proposed method. We also add the experiment to replace the initial model with the ResNet50 pretrained on LUPerson-NL, which is released in [70]. As shown in Table XI, 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.

*5) Qualitative Analysis of Visualization:* We present some retrieval examples with top 6 retrieved images in Fig. 5. Our proposed method can achieve great improvements of the baseline CCL. In the first two rows, CCL gets some false results for the query due to the high similarity between different persons, in terms of clothes, gender, and bicycle. However, our proposed method can find the true retrieval results with limited annotations from the free teacher model and the human annotator, which indicates that our method can learn more discriminative representations for differing different persons. In the last two rows, CCL could suffer from accumulation of label noise in the training process due to the similarity of different persons. As a result, these similar images could be easily merged to the same cluster and make the model biased. But our method can deliver true retrieval results, which verifies the necessity of our proposed modules to relieve the accumulation of label noise.

## VI. Limitation and Future Work

Although our proposed method can achieve great improvements with very limited annotation budget, our method still has two main limitations. On the one hand, our method requires

the frame number information. Although it is easy to obtain such information in the data collection stage, the frame number information is still unavailable in some public person ReID datasets. On the other hand, our proposed SKD module relies on a fixed trained teacher model. Although it can help the student model relieve the severe label noise in the early period, it may also hinder the student model learning in the latter period of the training process. As these problems are hard to be solved, we leave how to improve them as future directions, for examples, by replacing the frame number information with neighbor information, or designing a better iterative knowledge distillation mechanism.

## VII. Conclusion and Discussion

In the article, we propose a highly efficient active learning framework for person ReID. To possibly reduce the demand of annotations from the human expert, we introduce the trained teacher model to serve as a free annotator and propose the tracklet-aware co-cooperative framework by taking advantage of frame number of hybrid annotators. To further relieve the influence of label noise, we propose the SKD module to guide the model to learn from the unlabeled data in a more stable way. Extensive experiments on three popular datasets demonstrate that our approach can achieve competitive performance compared with state-of-the-art methods in both USL and active learning settings with very limited annotations.

Compared with unsupervised person ReID, existing active learning person ReID methods typically require a large number of human annotations, which can limit scalability in real-world applications. We believe that designing efficient active learning frameworks is essential for advancing the field, and our proposed method could facilitate future research in this direction.

## References


[1] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "IAUnet: Global context-aware feature learning for person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4460–4474, Oct. 2021.

[2] Q. Zhou, B. Zhong, X. Liu, and R. Ji, "Attention-based neural architecture search for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6627–6639, Nov. 2022.

[3] K. Zhu, H. Guo, S. Liu, J. Wang, and M. Tang, "Learning semantics-consistent stripes with self-refinement for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 17, 2022, doi: 10.1109/TNNLS.2022.3151487.

[4] 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.

[5] 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.

[6] L. Wu, Y. Wang, L. Shao, and M. Wang, "3-D PersonVLAD: Learning deep global representations for video-based person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3347–3359, Nov. 2019.

[7] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, "Global-local multiple granularity learning for cross-modality visible-infrared person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 17, 2021, doi: 10.1109/TNNLS.2021.3085978.

[8] Z. Zheng, X. Wang, N. Zheng, and Y. Yang, "Parameter-efficient person re-identification in the 3D space," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 31, 2022, doi: 10.1109/TNNLS.2022.3214834.

[9] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4624–4634, Sep. 2022.

[10] S. Kothawade, N. Beck, K. Killamsetty, and R. Iyer, "SIMILAR: Submodular information measures based active learning in realistic scenarios," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18685–18697.

[11] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

[12] Y. Wang, "Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–25, Jan. 2021.

[13] D. Jin and M. Li, "Towards fewer labels: Support pair active learning for person re-identification," 2022, *arXiv:2204.10008*.

[14] P. Ren et al., "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021.

[15] Z. Zha, M. Wang, Y. Zheng, Y. Yang, R. Hong, and T. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, Feb. 2012.

[16] J. W. Cho, D. Kim, Y. Jung, and I. S. Kweon, "MCDAL: Maximum classifier discrepancy for active learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 22, 2022, doi: 10.1109/TNNLS.2022.3152786.

[17] G. Yu, Y. Xing, J. Wang, C. Domeniconi, and X. Zhang, "Multiview multi-instance multilabel active learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4311–4321, Sep. 2022.

[18] B. Hu, Z.-J. Zha, J. Liu, X. Zhu, and H. Xie, "Cluster and scatter: A multi-grained active semi-supervised learning framework for scalable person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2605–2614.

[19] 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.

[20] 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.

[21] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14940–14949.

[22] T. He, L. Shen, Y. Guo, G. Ding, and Z. Guo, "SECRET: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 879–887.

[23] Y. Wu et al., "Multi-centroid representation network for domain adaptive person re-ID," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2750–2758.

[24] 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.

[25] 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.

[26] Z. Liu, J. Wang, S. Gong, D. Tao, and H. Lu, "Deep reinforcement active learning for human-in-the-loop person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6121–6130.

[27] 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.

[28] Y. Yang, G. Wang, P. Tiwari, H. M. Pandey, and Z. Lei, "Pixel and feature transfer fusion for unsupervised cross-dataset person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2021.

[29] 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.

[30] J. Han, Y.-L. Li, and S. Wang, "Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 790–798.

[31] 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. 1–11.

[32] 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.*, vol. 33, 2019, pp. 8738–8745.

[33] Y. Wang, J. Peng, H. Wang, and M. Wang, "Progressive learning with multi-scale attention network for cross-domain vehicle re-identification," *Sci. China Inf. Sci.*, vol. 65, no. 6, Jun. 2022, Art. no. 160103.

[34] 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.

[35] 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.

[36] 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.

[37] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, vol. 96. no. 34, 1996, pp. 226–231.

[38] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 405–422.

[39] L. Xue, Y. Zou, P. Peng, Y. Tian, and T. Huang, "Annotation efficient person re-identification with diverse cluster-based pair selection," 2022, *arXiv:2203.05395*.

[40] W. Liu, X. Chang, L. Chen, and Y. Yang, "Early active learning with pairwise constraint for person re-identification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2017, pp. 103–118.

[41] S. Roy, S. Paul, N. E. Young, and A. K. Roy-Chowdhury, "Exploiting transitivity for learning person re-identification models on a budget," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7064–7072.

[42] X. Xu, L. Liu, X. Zhang, W. Guan, and R. Hu, "Rethinking data collection for person re-identification: Active redundancy reduction," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107827.

[43] W. Gao and M. Li, "Unsupervised clustering active learning for person re-identification," 2021, *arXiv:2112.13308*.

[44] 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. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 266–282.

[45] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 893–906, Feb. 2022.

[46] Q. Zhao, J. Dong, H. Yu, and S. Chen, "Distilling ordinal relation and dark knowledge for facial age estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3108–3121, Jul. 2021.

[47] T. Zhang, X. Wang, B. Liang, and B. Yuan, "Catastrophic interference in reinforcement learning: A solution based on context division and knowledge distillation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 19, 2022, doi: 10.1109/TNNLS.2022.3162241.

[48] C. Tan and J. Liu, "Improving knowledge distillation with a customized teacher," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 25, 2022, doi: 10.1109/TNNLS.2022.3189680.

[49] 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.

[50] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6547–6556.

[51] 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.

[52] 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.

[53] L. Lan, X. Teng, J. Zhang, X. Zhang, and D. Tao, "Learning to purification for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 3338–3353, 2023.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *ACM Sigir Forum*, vol. 29, no. 2. New York, NY, USA: ACM, 1995, pp. 13–19.

[61] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, Madison, WI, USA, Jul. 1998, pp. 1–9.

[62] S. Ebert, M. Fritz, and B. Schiele, "RALF: A reinforced active learning formulation for object class recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3626–3633.

[63] 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. (ICCV)*, Dec. 2015, pp. 1116–1124.

[64] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 17–35.

[65] 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.

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

[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] 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. 1–11.

[71] 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.

[72] Z. Yang, X. Jin, K. Zheng, and F. Zhao, "Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14278–14287.