Cluster and Scatter: A Multi-grained Active Semi-supervised Learning Framework for Scalable Person Re-identification

Bingyu Hu, Zheng-Jun Zha[†], Jiawei Liu, Xierong Zhu, Hongtao Xie
University of Science and Technology of China
{hby0728,zxr8192}@mail.ustc.edu.cn,{zhazj,jwliu6,htxie}@ustc.edu.cn

ABSTRACT

Active learning has recently attracted increasing attention in the task of person re-identification, due to its unique scalability that not only maximally reduces the annotation cost but also retains the satisfying performance. Although some preliminary active learning methods have been explored in scalable person re-identification task, they have the following two problems: 1) the inefficiency in the selection process of image pairs due to the huge search space, and 2) the ineffectiveness caused by ignoring the impact of unlabeled data in model training. Considering that, we propose a Multi-grained Active Semi-Supervised learning framework, named MASS, to address the scalable person re-identification problem existing in the practical scenarios. Specifically, we firstly design a cluster-scatter procedure to alleviate the inefficiency problem, which consists of two components: *cluster step* and *scatter step*. The *cluster step* shrinks the search space into individual small clusters by a coarse-grained clustering method, and the subsequent *scatter step* further mines the hard distinguished image pairs from unlabelled set to purify the learned clusters by a novel centrality-based adaptive purification strategy. Afterward, we introduce a customized purification loss for the purified clustering, which utilizes the complementary information in both labeled and unlabeled data to optimize the model for solving the ineffectiveness problem. The cluster-scatter procedure and the model optimization are performed in an iterative fashion to achieve the promising performance while greatly reducing the annotation cost. Extensive experimental results have demonstrated that MASS can even achieve a competitive performance with fully supervised methods in the case of extremely less annotation requirements.

CCS CONCEPTS

• **Information systems** → **Top-k retrieval in databases**.

KEYWORDS

Person re-identification; active learning; semi-supervised learning

[†] Corresponding author.

MM ’21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475435

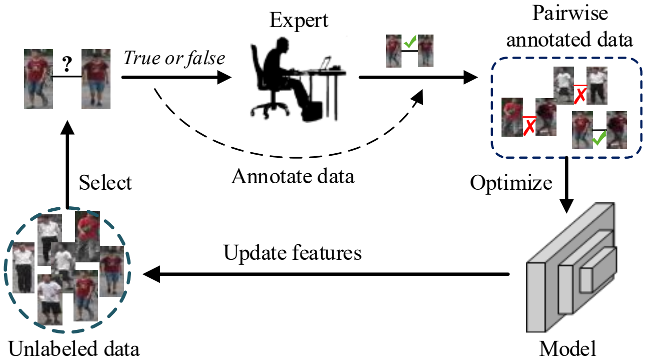


Figure 1: The active learning framework for person re-identification. Some image pairs from an unlabelled set are selected and annotated by an expert, and then the labeled image pairs are employed for model training. The iterations repeat until the annotation budget runs out.

ACM Reference Format:
Bingyu Hu, Zheng-Jun Zha[†], Jiawei Liu, Xierong Zhu, Hongtao Xie. 2021. Cluster and Scatter: A Multi-grained Active Semi-supervised Learning Framework for Scalable Person Re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475435

1 INTRODUCTION

Person re-identification is a common and hot research topic in the computer vision community, which aims to retrieve target person images across non-overlapping camera views given query images. Recently, plenty of research works have made significant progress on the supervised person re-identification problem [9, 14, 15, 24, 25, 37, 53]. Nevertheless, the prerequisite for prominent performance improvement in these methods is a large amount of pre-labeled training data as the supervision to adequately optimize the models. Therefore, they cannot be flexibly applied to large-scale practical deployments without sufficient annotations [26, 39]. Consequently, it has attracted growing attention from researchers to reduce the annotation cost while retaining the satisfying performance in large-scale person re-identification scenes.

Inspired by this, there are mainly two types of methods proposed to handle large-scale person re-identification. The first category is unsupervised learning methods [11, 17, 22, 23, 35, 36, 47, 49], which do not demand any human annotations. Their performance dramatically degrades on large-scale datasets because of the incremental visual discrepancy of the same pedestrian such as body pose, viewpoint and illumination. The second category is semi-supervised learning methods [4, 20, 21, 28, 42, 44–46]. Most of these works require making strong assumptions about datasets that parts of the identities (*e.g.*, one-third of the training set) are fully labeled under every camera view, which remains a serious challenge for large-scale surveillance networks. To further reduce the requirements of annotations for person re-identification, some researchers attempt to introduce the active learning technique to actively select the most informative image pairs to annotate, without requiring exhaustively labeling for every pair of cameras [26, 39, 48].

The pipeline of active learning framework for person re-identific ation is shown in Fig. 1. Firstly, some candidate samples are selected from the unlabeled dataset for annotations by experts. These labeled samples are then utilized to supervise model training in the next round [29, 48]. The whole process is iteratively repeated until the annotation budget is exhausted. Consequently, the effective sample selection strategy is the key component of the active learning framework. Note that for the person re-identification task, the annotation is performed in a pairwise fashion (a pair of images is truly matched or not), since it's difficult to directly distinguish the identity of a pedestrian image without knowing the overall number of identities beforehand. Wang *et al.* [39] selected the top-50 negative retrieval images from the whole gallery set with a query image to annotate, which provides the most informative feedback on model mistakes. Liu *et al.* [26] utilized a reinforcement learning agent to select informative samples (measuring the uncertainty of samples) from the whole gallery set for a given input query.

Although these two active learning methods make preliminary attempts in reducing the labeling cost for large-scale person re-identification, they still have two serious problems: 1) the number of total image pairs grows quadratic with the number of images, resulting in that the search space of pairwise annotation expands rapidly. The overwhelming global space increases the difficulty and decreases the efficiency in searching valuable image pairs for annotation; 2) existing active learning methods only employ the labeled data to optimize the model, but completely ignore the abundant complementary information of the unlabeled data [26]. Even though the most valuable data are actively selected and annotated, those less valuable ones (unannotated) are still worth considering due to their vast quantity. Moreover, the unsupervised methods have also shown that the unlabeled data contain abundant complementary visual cues [47, 49]. To sum up, there remain two major challenges in reducing the labeling cost for large-scale person re-identification: (1) how to efficiently perform image pair selection for annotation in a huge search space; (2) how to effectively incorporate both labeled and unlabeled data into model training.

To this end, we propose a Multi-grained Active Semi-Supervised learning framework (MASS) for scalable person re-identification. Specifically, we draw on the active learning paradigm, but with some significant differences. To tackle the inefficiency problem in image pair selection, we design a novel procedure, termed **cluster-scatter**, consisting of (1) *cluster step*: cluster the samples with the extracted features from a base CNN model, and (2) *scatter step*: scatter out the impurities of each cluster with experts' annotations. At the *cluster step*, we significantly shrink the search space from the whole dataset into individual small clusters, and concentrate on the wrongly clustered impurities, which significantly influence re-identification performance. Moreover, to support the later cluster purification, we construct a weighted graph for each cluster and search the candidate pairs in the graph separately. At the *scatter step*, we propose a centrality-based adaptive purification strategy (CAPS), which efficiently deduces the impurities and purifies the coarse-grained clusters with low annotation cost. For addressing the ineffectiveness challenge in model training, we propose a purification loss for the output purified clustering, which is used to optimize the base model. The purification loss naturally explores the complementary information in both labeled and unlabeled data. The pipeline of MASS iterates until achieving both a high-quality clustering and a discriminative base model. The joint efficient image pair selection and effective model training lead to high re-identification performance and low annotation cost. Extensive experimental results have demonstrated the efficiency and effectiveness of MASS.

The main contributions of our work can be summarized as follows: (1) We introduce a multi-grained active semi-supervised learning framework, termed MASS, for scalable person re-identification, which achieves satisfying re-identification performance and extremely reduces human annotation cost on three large-scale datasets. (2) We propose a novel cluster-scatter procedure, where the *cluster step* shrinks the search space with coarse-grained clustering and the *scatter step* efficiently mines wrongly clustered samples by the proposed centrality-based adaptive purification strategy. (3) We integrate the purified clustering into the model optimization by introducing a customized purification loss, which effectively utilizes the complementary information of both labeled and unlabeled data.

2 RELATED WORK

Active Learning. Active learning aims to select the most informative samples from the unlabeled dataset and hand it over to the expert for labeling, so as to reduce the cost of labeling as much as possible while still maintaining performance. The key component in this process is the sample selection strategy [2, 29]. For example, Das *et al.* [5] conducted the entropy-based criterion and Liu *et al.* [26] utilized reinforcement learning to design a powerful sample selection strategy given human feedback annotations. However, all existing active learning methods for person re-identification only utilize a spot of labeled data to train the model, ignoring the vast potential of unlabeled data for boosting feature representation [16, 33]. The idea of combining active learning with semi-supervised learning, although quite natural, has received relatively little attention. Rhee *et al.* [30] proposed an active semi-supervised system which demonstrates superior performance in the object detection task. Sinha *et al.* [34] trained a variational auto-encoder and an adversarial network by using both labeled and unlabeled samples to infer the representativeness of unlabeled samples during the sampling process. However, these methods focus on closed-set tasks, which cannot be directly applied to the open-set retrieval task of person re-identification due to requiring pairwise annotation.

Semi-supervised Person Re-identification. Most state-of-the-art semi-supervised works in person re-identification[4, 20, 21, 28, 42, 44–46] adopt the pattern of training a model initially on the labeled portion and assigning pseudo-labels to the unlabeled examples by exploiting the inherent relationships between them. For example, Xin *et al.* [44] proposed a multi-view clustering method, which uses several different CNN to extract features and combines them to cluster unlabeled samples and generate pseudo labels. Wu *et al.* [42] proposed a learning method for the unlabeled data which contains the exclusive loss and a pseudo-labels estimation technique. However, the annotation cost in these methods is still large, such as 1/3 identities require to be labeled for every camera view. Besides, there exists redundancy in labeling all images if they can be distinguished by the inherent similarities, wasting the labeling efforts. Consequently, we focus on active learning to annotate the most informative samples so as to reduce the annotating cost.

Unsupervised Person Re-identification. The unsupervised person re-identification methods are mainly divided into two categories. The first category is the cross-domain person re-identification [17, 36, 49–51], which requires auxiliary source domain datasets for pre-training or joint training. However, the source domain datasets may not be available because of privacy problems in real scenes. Another category is the purely unsupervised methods requiring no labels [22, 35, 47]. The visual discrepancy in the images of the same pedestrian under different camera views raises the difficulty in feature representation, thus degrading their performance. Nevertheless, these methods inspire us that the abundant inherent information of unlabeled data is beneficial for optimizing the model.

3 METHOD

3.1 Problem Definition and Overview

We denote the samples as $S = \{1, 2, ..., N\}$ in the dataset, and their identity labels as $L = \{l_i\}_{i=1}^N$, where N is the total number of samples. In our setting, L is totally unknown in advance and an expert $\mathcal{E}$ provides pairwise annotation during the training stage. Given a pair of images (i, j), the expert provides a binary response indicating whether they belong to the same identity, $\mathcal{E}(i, j) = \mathbb{1}[l_i = l_j]$. Moreover, a base CNN model $\mathcal{M}$ is employed to extract discriminative features for the samples $X = \{x_i\}_{i=1}^N$. The objective of active learning for person re-identification is to learn a high-quality discriminative model $\mathcal{M}$ while keeping low annotation cost, given the initially unlabeled dataset S and the expert $\mathcal{E}$.

To achieve this goal, we propose a novel **M**ulti-grained **A**ctive **S**emi-supervised framework for **S**calable person re-identification (MASS). As illustrated in Fig 2, MASS incorporates the selective pairwise annotation and the model optimization in a cooperative and iterative fashion. At the iteration t, we firstly employ the current model $\mathcal{M}_t$ to extract the features of the whole dataset, $X_t = \left\{x_i^t\right\}_{i=1}^N$. Then we conduct a cluster-scatter procedure, consisting of a coarse-grained cluster step which clusters the samples, and a fine-grained scatter step which purifies the clusters obtained before. Specifically, at the cluster step, we perform the FINCH clustering algorithm on the samples based on their extracted features X_t, and construct a so-called NN-relation graph for each cluster. In the scatter step, we design a novel centrality-based adaptive purification strategy (CAPS) to purify the coarse-grained clusters with pairwise annotations according to the structural information contained in the NN-relation graphs. Summarily, the cluster-scatter procedure organizes the samples as a group of purified clusters. Finally, we propose the purification loss to train $\mathcal{M}_t$ with the purified clusters. The updated model $\mathcal{M}_{t+1}$ is able to extract more discriminative features in the next iteration and produce more high-quality clustering. The repeating iterations construct a positive feedback loop between model discriminativeness and clustering quality. The overall flowchart is illustrated in Algo. 1 and the details will be discussed in the rest of the section.

Algorithm 1 MASS Framework

```
Input: An unlabeled dataset $S = \{1, 2, ..., N\}$, a base CNN model $\mathcal{M}$ pre-trained with ImageNet, a human expert $\mathcal{E}$, and the number of iterations $T$.
Output: An optimized model $\mathcal{M}_T$.
for $t = 0, ..., T - 1$ do
    Extract features $X_t$ with $\mathcal{M}_t$;
    Coarse-grained clustering: $\Gamma = \left\{\hat{C}_1, ..., \hat{C}_{|\Gamma|}\right\} = \text{FINCH}(X_t)$;
    Construct an NN-relation graph $G$ for each cluster $\hat{C} \in \Gamma$;
    Cluster purification: $C = \text{CAPS}(\hat{C}, \mathcal{E}, G)$ for $\hat{C} \in \Gamma$ (Algo. 2);
    Training with purification loss (Eq. (7)): $\mathcal{M}_{t+1} \longleftarrow \mathcal{M}_t$;
return $\mathcal{M}_T$
```

3.2 Cluster Step

Given the extracted features of the samples, the cluster step aims to perform a coarse-grained clustering. In our MASS framework, we leverage the recently proposed FINCH algorithm [32], which is a parameter-free hierarchical clustering method based on the nearest neighbor (NN) relations between the samples. Moreover, instead of solely acquiring the resulting clusters, we propose to build an NN-relation graph for each cluster, providing additional information to support the cluster purification in the scatter step.

Clustering algorithm. Generally, FINCH works by linking the samples with nearest neighbor relations and grouping each connected component as a cluster. The process is executed recursively to obtain a hierarchical clustering by treating the centroids of the clusters as a new sample set. In detail, an adjacency matrix is defined between all sample pairs:

$$A(i, j) = \begin{cases} 1 & \text{if } j = \kappa_i^1 \text{ or } \kappa_j^1 = i \text{ or } \kappa_i^1 = \kappa_j^1 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where κ_i^1 represents the first nearest neighbor (1NN) of sample i. Summarily, the adjacency matrix joins each sample i to their 1NN via $j = \kappa_i^1$, enforces symmetry through $\kappa_j^1 = i$, and links the samples that have a shared nearest neighbor (SNN) $\kappa_i^1 = \kappa_j^1$. To produce a clustering tree that provides meaningful ways of interpreting data at different levels of granularity, FINCH recursively merges the clusters in an agglomerative fashion by averaging the data samples in each cluster and using the mean vectors for computation in the next iteration, until there leaves only one cluster, *i.e.*, the whole dataset. We adopt FINCH for three reasons. Firstly, FINCH requires neither hyper-parameters nor prior knowledge such as the number of clusters, and thus can be easily used in practice. Secondly, FINCH keeps a low computational overhead which leads to good scalability. Thirdly, apart from data partitions, FINCH provides semantic structure via 1NN and SNN relations, benefiting cluster purification potentially, which will be discussed in detail next.

NN-relation graph. Considering that the clusters require to be purified later, there are supposed to be more details about the internal information of each cluster, *e.g.*, *which pair of samples probably have the same label*, *which sample is more likely to be an*

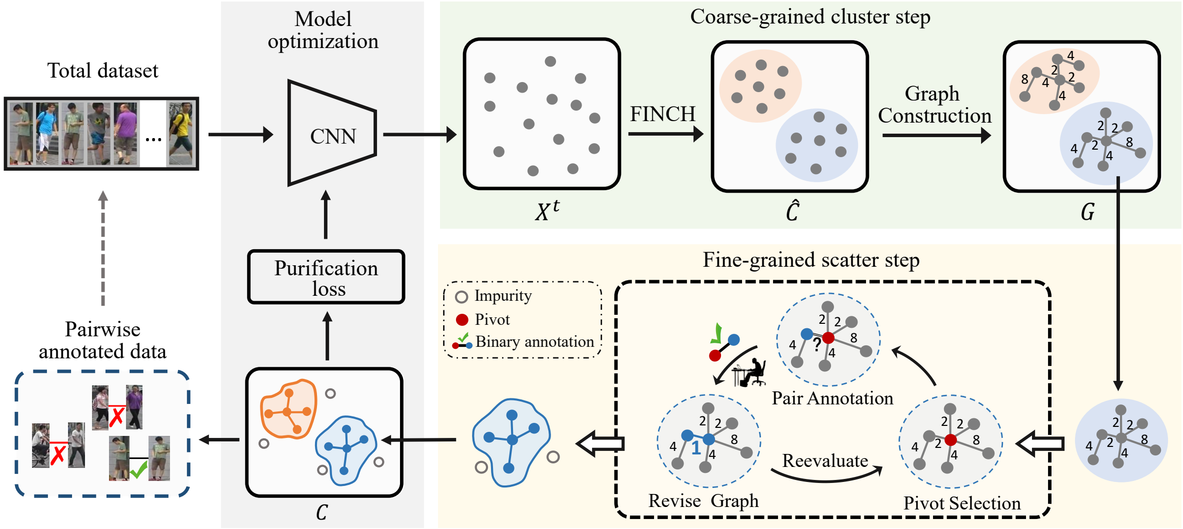


Figure 2: The Multi-grained Active Semi-supervised Learning Framework for Scalable Person Re-identification. The cluster-scatter procedure efficiently performs image pair selection for annotation and retains the purified clustering. A base CNN model is effectively optimized with the purified clustering by the purification loss and then updates the feature for clustering. This iterative process stops when it reaches the annotation budget.

impurity, *etc.* In FINCH, such information is inherently contained in the nearest neighbor (NN) relations among the samples. Intuitively, we model each cluster as a graph structure as below. Basically, the samples are connected according to Eq. (1). Moreover, note that FINCH works recursively, when two clusters are merged at the k-th level, we interconnect the samples of the two clusters in pairs. To distinguish the connections at different partition levels, we introduce weights for the edges so that high-level connections are assigned with longer distances. Formally, we define an NN-relation graph $G = (V, E, w)$. The vertex set $V = \{1, ..., N\}$ represents the samples. The edge set $E \subset V \times V$ links vertices with NN relations: $\forall u, v \in V, (u, v) \in E$ iff u and v (or the centroids of the clusters they belong to) are linked by an NN-relation. The weight $w : E \longrightarrow \mathbb{R}$ assigns a distance to each edge: $\forall (u, v) \in E, w(u, v) = F(k)$, where $k \in \mathbb{Z}^+$ is the clustering level where u and v is linked, and $F(\cdot)$ is a positive monotonically increasing function, which is empirically set as $F(k) = 2^k$.

3.3 Scatter Step

At the cluster step, we obtain a coarse-grained graph-structured clustering of the samples. Typically, each cluster contains a majority of samples having the same label called *dominant samples*, and some minor misclustered samples called *impurities*. Such impurities are exactly the hard samples degrading the performance of the model, which is difficult to directly find in a large-scale dataset. As the search space has been shrunk into individual small clusters, at the scatter step, we mine such hard samples by purifying each cluster, *i.e.*, distinguish impurities from dominant samples with pairwise annotations. The difficulty lies in achieving the goal while keeping low annotation cost. To this end, we propose a centrality-based adaptive purification strategy (CAPS) which leverages the NN-relation graphs mentioned before.

Intuition. We firstly discuss our insight behind the idea. Supposing a cluster consisting of n samples, all the possible relations counts to $\frac{n(n-1)}{2}$. It is obviously unnecessary for exhaustive annotations. The key observation is that, if we can locate one of the dominant samples, which we call a *pivot*, the rest becomes naive. We just require to ask the expert to compare this pivot with other samples one by one, and scatter out those who have a negative relationship with the pivot. Under this circumstance, the annotation cost reduces to only $n - 1$ pairs. However, as the labels of the samples are unknown in prior, the pivot can only be chosen heuristically and the choice may be wrong (*i.e.*, the chosen pivot is indeed an impurity). Thereby, instead of keeping a static pivot, we should make an adaptive and dynamic choice based on the annotated pairs from the expert. Intuitively, the more positive relations the pivot has with other samples, the more likely it is to be a real dominant sample, vice versa. Moreover, we consider a situation in which the label distribution of the samples is both diverse and even so that there is no dominant label. In this case, every sample in the cluster can be treated as an impurity so that the purification is meaningless and wasteful. Instead, we require a clever stopping condition to realize such chaos in advance and dismiss the cluster immediately. Illustrated in Algo. 2, our CAPS strategy can be stated as follows:

(1) Select a sample as the pivot.
(2) Select a pair associated with the pivot to annotate.
(3) Re-evaluate the heuristic about the pivot.
(4) Exit if the stopping condition is met, otherwise return to (1).

Now we elaborate on the details of the four parts above.

Pivot selection. The key to implementing the pivot-centered strategy mentioned above is introducing the concept of *centrality* in the NN-relation graph G, which is widely used in graph theory to characterize the importance of a vertex [3, 7, 38, 40]. In CAPS, the Harmonic centrality [27] is adopted, which is formulated as below:

$$H_G(v) = \frac{1}{|V| - 1} \sum_{u \neq v} \frac{1}{d(u, v)}, \tag{2}$$

where $d(u, v)$ is the shortest distance between u and v, and $\frac{1}{d(u,v)} :=$ 0 if there is no path from u to v. Principally, $H_G(v)$ measures how close on average v is to the other vertices; the higher $H_G(v)$, the more vertices tend to have the same label as v. Thereby, we choose the sample with the highest centrality as the pivot, which is heuristically likely to be a dominant sample.

Pair annotation. Given a pivot, we choose a target sample and query the relationship between the pivot and the target sample. We give priority to the sample which has the least distance to the pivot because they are prone to have a positive relationship. Compared to negative ones, positive annotations have the advantage of being transitive. Specifically, the samples that have been connected with positive relationships make up a transitive closure. When a pair of samples are annotated, the relationship will propagate through the transitive closures they belong to. As a result, preferentially obtaining positive annotations could help infer more relationships and potentially save the annotation cost.

Heuristic re-evaluation. According to the obtained annotations, the heuristic of pivot selection should be re-evaluated. We employ the annotations as constraints to revise the graph structure, so that the centrality of the pivot will increase if its associated annotation is positive and vice versa. Formally, we denote $\Omega = \Omega^+ \cup \Omega^- \subset E$ as the set of pairs whose relationship we have known by either annotation or inference; Ω^+ consists of positive pairs and Ω^- consists of negative pairs. Once we obtain a new annotation, we update Ω and revise the graph $G = (V, E, w)$ to $G' = (V', E', w')$ as follows:

$$\begin{aligned} V' &= V, \\ E' &= (E \setminus \Omega^-) \cup \Omega^+, \\ w'(u, v) &= \begin{cases} 1 & \text{if } (u, v) \in \Omega^+ \\ w(u, v) & \text{otherwise} \end{cases}. \end{aligned} \tag{3}$$

Summarily, if the link (u, v) is positive, it will be shortcut with a lower bound distance; otherwise, it will be disconnected. After obtaining the revised graph G', we recompute the centrality of vertices $H_{G'}(v), v \in V'$ for the next pair annotation.

Stopping condition. With the heuristic reevaluation above, the pivot may change over time. But ultimately, the transitive closures consisting of dominant samples will become a stable pivot, and the strategy will repeat querying the relationship between it and the rest samples. Thus when there is no sample with uncertain relationship to the currently selected pivot, the CAPS strategy stops normally. In addition, considering the circumstance with a chaotic cluster as mentioned before, CAPS has an early-stopping condition. When no dominant sample exists, the annotated pairs are mostly negative, resulting in the decrease of the centrality of all vertices. Therefore, we propose to early-stop the strategy when the maximum centrality is lower than a threshold $\beta \in (0, 1)$ (note that $H_G(v) \in [0, 1]$ because $d(u, v) \in [1, +\infty)$).

Algorithm 2 Centrality-based adaptive purification strategy

```
Input: A coarse-grained cluster $\hat{C}$ represented as an NN-relation
    graph $G = (V, E, w)$, a human expert $\mathcal{E}$, and a threshold hyper-
    parameter $\beta \in (0, 1)$.
Output: A purified cluster $C \subset \hat{C}$ consisting of dominant samples
Initialization $\Omega = \Omega^+ \cup \Omega^- \subset E$ represents the set of known
    pairs, where $\Omega^+/\Omega^-$ denotes positive/negative pairs;
while True do
    Compute centrality $H_G(v)$ for $v \in V$ (Eq. (2));
    if $\max_{v \in V} H_G(v) \leq \beta$ then
        $C = \emptyset$; // Early stopping with no dominant samples
        break;
    Select a pivot $s = argmax_{v \in V} H_G(v)$;
    Get target candidates $D = \{v \in V | (v, s) \notin \Omega\}$;
    if $D = \emptyset$ then
        $C = \{v \in V | (v, s) \in \Omega^+\}$;
        break;
    Select a target sample $t = argmin_{v \in V} d(v, s)$;
    Query for expert $\mathcal{E}$ to annotate pair $(s, t)$;
    Update $\Omega$ with $\mathcal{E}(s, t)$ taking transitivity into consideration;
    Revise the graph $G$ according to Eq. (3);
    return $C$;
```

3.4 Model Training

At the last of an iteration, we turn to find an effective objective function for the output purified clustering, which is used to optimize the base model $\mathcal{M}$ and obtain more robust and discriminative pedestrian representations. The purified clustering is formed by purified clusters and isolated impurities. Their quantity varies continuously during the training process. The isolated instances are proved to be valuable for training the model by unsupervised learning methods [11]. Such an impurity contains extra information beneficial to model training: as the impurity is misclustered into a specific cluster, it is worthy to explicitly impose the model to distinguish it from that cluster. Consequently, an ideal learning objective not only exploits the class-wise supervisory knowledge and instance-wise supervisory knowledge in the purified clusters, but also leverages the extra information in the isolated impurities.

To this end, we design the purification loss $\mathcal{L}$ from two aspects, *i.e.*, cluster loss $\mathcal{L}_C$ and scatter loss $\mathcal{L}_\mathcal{S}$. We utilize the cluster loss $\mathcal{L}_\mathcal{S}$ to properly model intra-/inter-class affinities in the purified clustering. Due to the dynamic number of clusters and isolated instances, we build our formulation upon a non-parametric softmax operation [43], which defines the probability that an arbitrary feature $\boldsymbol{x}$ is recognized as the i-th image to be:

$$P(i|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{x}/\tau)}{\sum_{j=1}^{N} \exp(\boldsymbol{x}_j^T \boldsymbol{x}/\tau)}, \tag{4}$$

where $\tau \in [0, 1]$ is a fixed temperature hyper-parameter, and all $\boldsymbol{v}$ are normalized by $||\boldsymbol{x}||_2 = 1$.

The insight behind is that each image instance is treated as a distinct class and aims to learn the representations that capture the apparent similarity among instances. Now that a pure cluster is made up of the samples which must belong to the same pedestrian, we regard the output pure clusters and isolated instances as the distinct classes uniformly and employ the centroids as the class prototypes. The probability of $\boldsymbol{x}$ belongs to i-th class is defined as:

$$P(\boldsymbol{c_i}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{c_i}^T\boldsymbol{x}/\tau)}{\sum_{j=1}^{L} \exp(\boldsymbol{c_j}^T\boldsymbol{x}/\tau)}, \tag{5}$$

where $\boldsymbol{c}_i$ is the centroid of the i-th class and L is the total number of clusters and isolated instances.

$$\boldsymbol{c}_i = \frac{1}{|I_i|} \sum_{\boldsymbol{x}_k \in I_i} \boldsymbol{x}_k, \tag{6}$$

where I_i denoted the i-th class set. when the class is indeed an isolated instance, $\boldsymbol{c}_i = \boldsymbol{x}_i$. Finally, we formulate the cluster loss as:

$$\mathcal{L}_C = -\log \frac{\exp(\boldsymbol{c_i}^T\boldsymbol{x}/\tau)}{\sum_{j=1}^{L} \exp(\boldsymbol{c_j}^T\boldsymbol{x}/\tau)}. \tag{7}$$

As the number of clusters and isolated instances may change with the alternate clustering process, the class prototypes for the cluster loss are built in a dynamic manner. Following [12, 55], we maintain a memory bank to store an average approximation to all features $\overline{\boldsymbol{F}} = \{\overline{\boldsymbol{x_1}}, \overline{\boldsymbol{x_2}}, ..., \overline{\boldsymbol{x_N}}\}$ and continuously update the class prototype, *i.e.*, centroids. After extracting a new feature $\boldsymbol{x}$, we update its corresponding instance entry $\overline{\boldsymbol{x}}$ as follows:

$$\overline{\boldsymbol{x}}_i \longleftarrow m\overline{\boldsymbol{x}}_i + (1-m)\boldsymbol{x}_i, \tag{8}$$

where $m \in [0, 1)$ is the momentum coefficient for updating features. After that, the class prototype (Eq. (6)) can be rewritten by replacing the feature $\boldsymbol{x}$ with $\overline{\boldsymbol{x}}$.

On the other hand, we design the scatter loss to leverage the information containing in the impurities, denoted as $\{x^{im}\}_{k=1}^{M}$. Suppose the impurity x_k^{im} is weeded out by the cluster c_k, x_k^{im} is hard to be distinguished by clustering based on their apparent similarity. This extra information can be leveraged to improve the model discriminativeness. The impurity should be pulled away from its corresponding cluster in the next iteration. Consequently, the scatter loss is formulated as below:

$$\mathcal{L}_S = \sum_{k=1}^{M} \max(0, \overline{\boldsymbol{d}_k} - \boldsymbol{d}(x_k^{im}, c_k)), \tag{9}$$

where $\boldsymbol{d}$ is the Euclidean distance, $\overline{\boldsymbol{d}_k}$ is the average of the Euclidean distances between each sample and centroid in the cluster.

In conclusion, the overall purification loss is formulated as :

$$\mathcal{L} = \mathcal{L}_C + \lambda\mathcal{L}_S, \tag{10}$$

where λ is the hyper-parameter for balancing the two loss terms.

4 EXPERIMENTS

4.1 Datasets and Settings

Datasets. To verify the effectiveness of our proposed method, we conduct several experiments on three popular benchmarks. Market-1501 [52] dataset contains 32,668 images of 1501 identities captured from 6 cameras. DukeMTMC-ReID [31] consists of 36,411 images belonging to 1,404 identities from 8 cameras. MSMT17 [41] is the largest image dataset in person re-identification so far, which includes 4,101 identities and 126,441 images captured by 15 cameras.

Evaluation protocols. Two evaluation metrics, *i.e.*, Cumulated Matching Characteristics (CMC) at Rank-k and mean average precision (mAP), are adopted to evaluate the performance.

Implementation details. The implementation of the proposed method is based on the PyTorch framework with four NVIDIA GTX1080Ti GPUs. We employ the ResNet-50 [6, 13] as the base model $\mathcal{M}$. Adam [18] optimizer with learning rate of 0.00035 and betas of [0.9,0.999] is adopted. We use clusters from the second-level partition of the FINCH algorithm. The threshold hyper-parameter β of early-stopping is set to 0.2. For the purification loss, we set the temperature $\tau = 0.05$ and the momentum coefficient $m = 0.2$. The budget of the reported results is $5n$, where n is the number of training images. The total number of iterations T is set to 50 and the balance parameter λ in Eq. (10) is set to 0.05.

4.2 Comparison to State-of-the-Arts

Comparison to active learning methods. As little attention has been drawn to active learning for person re-identification so far, we not only compare MASS with the state-of-the-art methods, *i.e.*, HVIL [39] and DRAL [26], but also three traditional active learning methods, *i.e.*, Query Instance Uncertainty (QIU) [19], Query By Committee (QBC) [1] and Graph Density (GD) [8]. In detail, HVIL labels the top-50 candidates of each query instance while DRAL using the triplet loss as the reward function of a reinforced agent to mine hard samples. As for the traditional methods, QIU queries the samples with the highest uncertainty. QBC selects the instances which cause maximum disagreement among the committee. GD constructs a graph to search highly connected nodes and select the most representative samples. Table 1, 2, 3 report the Rank-1, Rank-5, Rank-10 accuracy and mAP score on three benchmarks respectively. The proposed MASS outperforms all these methods by a large margin with much lower annotation cost. Specifically, MASS achieves 93.5%, 86.1%, 54.1% Rank-1 accuracy and 81.7%, 72.9%, 30.0% mAP score on the Market-1501, DukeMTMC-ReID, MSMT17 datasets. Taking Market-1501 as an example, MASS achieve 93.5% Rank-1 accuracy and 81.7% mAP score under the annotation budget of $5n$, which improves the state-of-the-art method DRAL by 9.3% Rank-1 accuracy and 15.4% mAP score under the larger annotation budget of $10n$, indicating the effectiveness and efficiency of MASS.

Comparison to semi-supervised learning methods. Since there are few previous works focusing on MSMT17 dataset, we will describe the results from different semi-supervised learning methods on Market-1501, DukeMTMC-ReID and MSMT17, respectively. No doubt that the larger labeled set, the better performance of the model. Consequently, for Market-1501 and DukeMTMC-ReID datasets, we choose the semi-supervised methods with the largest labeled data, *i.e.*, one third of identities, generally reporting the best results. Table 1, 2 show the performance comparisons of the proposed MASS against three existing methods, which include MVC [45], SPC [46] and TSSML [4]. Besides, for MSMT17 dataset, we compare our method with the methods in Table 3, *i.e.*, TAUDL [20], UTAL [21] and PCSL [28]. These methods adopt the setting of providing the intra-camera labels but no inter-camera labels. The proposed MASS outperforms all the semi-supervised methods on Rank-1 accuracy, and improves the second best works by 9.6%, 13.4%, 5.8% Rank-1 accuracy on Market-1501, DukeMTMC-ReID and MSMT17 dataset, respectively. This result verifies that

Table 1: Performance (%) comparison to active learning, semi-supervised learning and unsupervised learning methods on Market-1501 dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
Cross-Domain Unsupervised Methods				
MLC [35]	84.4	92.8	95.0	60.4
SNR [17]	85.5	-	-	65.9
AD-Cluster [49]	86.7	94.4	96.5	68.3
SPCL [11]	90.3	96.2	97.7	76.7
Pure Unsupervised Methods				
SSL [22]	71.7	83.8	87.4	37.8
HCT [47]	80.0	91.6	95.2	56.4
MLC [35]	80.3	89.4	92.3	45.5
SPCL [11]	88.1	95.1	97.0	73.1
Semi-Supervised Methods				
SPC [46]	71.5	86.2	90.4	53.2
MVC [45]	75.2	-	-	52.6
TSSML [4]	83.9	93.1	-	65.6
Active Learning Methods				
Random	58.0	79.1	85.7	35.2
QIU [19]	67.8	85.7	91.1	45.0
QBC [1]	68.4	86.1	91.2	46.3
GD [8]	71.4	87.1	91.4	49.3
HVIL [39]	78.0	-	-	-
DRAL [26]	84.2	94.3	96.6	66.3
MASS(ours)	**93.5**	**97.9**	**98.6**	**81.7**

designing an effective active selection strategy to obtain more valuable labeled data does benefit person re-identification performance.

Comparison to unsupervised learning methods. Finally, we compare MASS to state-of-the-art unsupervised learning methods. The compared approaches belong to two categories, *i.e.*, the pure unsupervised methods including SSL [22], HCT [47], MLC [35] and SPCL [11], which do not require any annotation and the cross-domain unsupervised methods including MLC [35], SNR [17], AD-Cluster [49], SPCL [11], ECN [54] and SSG [10], which require source dataset to pretrain the model. The proposed MASS outperforms the 2nd best methods by 3.2%, 3.2%, 0.4% Rank-1 accuracy and 5.0%, 4.1%, 3.2% mAP score on Market-1501, DukeMTMC-ReID and MSMT17 datasets respectively. Moreover, the proposed MASS surpasses all the cross-domain unsupervised methods, where each image only requires comparing to other five images on average for annotation. These results clearly verify the effectiveness of MASS, indicating that a discriminative model can be learned effectively without annotating large quantities of training data.

Table 2: Performance (%) comparison to active learning, semi-supervised learning and unsupervised learning methods on DukeMTMC-ReID dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
Cross-Domain Unsupervised Methods				
MLC [35]	72.4	82.9	85.0	51.4
AD-Cluster [49]	72.6	82.5	85.5	54.1
SNR [17]	78.2	-	-	61.6
SPCL [11]	82.9	90.1	92.5	68.8
Pure Unsupervised Methods				
SSL [22]	52.5	63.5	68.9	28.6
HCT [47]	65.2	75.9	80.0	40.2
MLC [35]	69.6	83.4	87.4	50.7
SPCL [11]	81.2	90.3	92.2	65.3
Semi-Supervised Methods				
MVC [45]	57.6	-	-	37.8
SPC [46]	58.5	73.7	79.3	37.4
TSSML [4]	72.7	85.2	-	53.2
Active Learning Methods				
Random	44.7	63.6	70.7	25.7
GD [8]	53.5	70.0	75.8	33.6
QIU [19]	56.8	74.2	79.3	36.8
QBC [1]	61.1	77.4	82.4	40.8
DRAL [26]	74.3	84.8	88.4	56.0
MASS(ours)	**86.1**	**93.9**	**95.9**	**72.9**

Table 3: Performance (%) comparison to active learning, semi-supervised learning and unsupervised learning methods on MSMT17 dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
Cross-Domain Unsupervised Methods				
ECN [54]	30.2	41.5	46.8	10.2
SSG [10]	32.2	-	51.2	13.3
MLC [35]	43.6	54.3	58.9	16.2
SPCL [11]	53.7	65.0	69.8	26.8
Pure Unsupervised Methods				
MLC [35]	35.4	44.8	49.8	11.2
SPCL [11]	42.3	55.6	61.2	19.1
Semi-Supervised Methods				
TAUDL [20]	28.4	-	-	12.5
UTAL [21]	31.4	-	-	13.1
PCSL [28]	48.3	62.8	68.6	20.7
Active Learning Methods				
Random	23.8	42.5	50.2	12.4
QIU [19]	31.3	52.6	59.8	21.5
QBC [1]	39.4	57.0	61.3	24.9
GD [8]	42.5	58.2	63.6	25.4
MASS(ours)	**54.1**	**65.4**	**70.4**	**30.0**

4.3 Comparison with different budget

In this work, annotation and model optimization work in a step-by-step pattern, where some image pairs are selected from clustering for annotation, and then the purified clustering is utilized for model training. The iterations repeat until the annotation budget is exhausted. With the train set n, the budget is $5n$ for the reported MASS result. We further compare the performance of the proposed MASS approach in a varying amount of labeled data, *i.e.*, different budgets on Market-1501. As shown in Table 4, we present the performance with the budget setting of n, $3n$ and $5n$. In all settings, the total number of iteration is set as 50 to ensure the convergence of the model. During the training iterations, when the budget runs out, there is no more purification operation from now on and the model

Table 4: Performance (%) comparison with different budget on Market-1501 dataset, where n indicates the total training images number.

Method	Market-1501				Budget
	Rank-1	Rank-5	Rank-10	mAP	
SOTA(DRAL)	84.2	94.3	96.6	66.3	10n
MASS	82.0	91.8	94.8	62.9	n
	90.7	96.5	97.8	76.7	3n
	93.5	97.9	98.6	81.7	5n

Table 5: Ablation study on Market-1501 dataset.

Model	Rank-1	Rank-5	Rank-10	mAP
MASS *w/o* al	82.9	90.7	95.2	64.3
MASS *w/o* ssl	86.2	94.1	96.6	70.0
MASS	93.5	97.9	98.6	81.7

is directly trained with the coarse-grained clustering result. We can observe that with more annotation budget, the model becomes stronger. Besides, the performance of MASS outperforms the state-of-the-art method by a large margin while the annotation cost is much lower than it, which demonstrates the effectiveness of MASS.

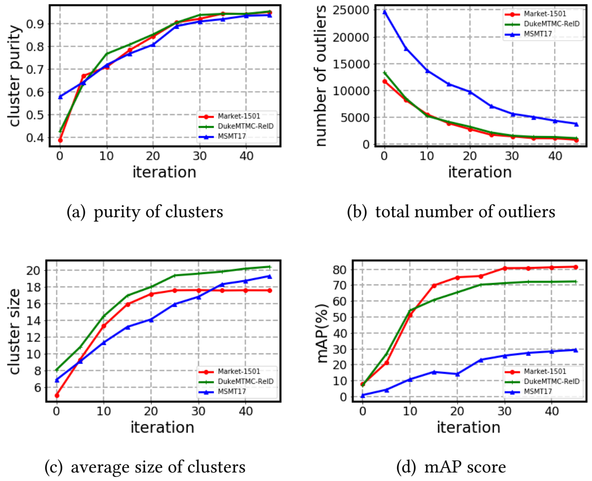
4.4 Ablation Study

In order to verify the necessity of the efficient pairwise selection and the effective model training, we conduct ablation experiments, and the results on Market-1501 dataset are shown in Table 5. MASS *w/o* al refers to the proposed MASS without the active learning technique, which means that we directly utilize the coarse-grained clustering for model training rather than the purified clustering. By comparing the MASS w/o al to MASS, we can observe the Rank-1 accuracy drops 10.6%, indicating the effectiveness of introducing the cluster-scatter procedure for active learning. MASS *w/o* ssl refers to the proposed MASS without using the unlabeled data for the active learning technique. We only use labeled pairs to train the model with the triplet loss. The Rank-1 accuracy and mAP score of MASS *w/o* ssl are both inferior to MASS, which demonstrates that learning from the unlabeled set does benefit the model performance.

Improvement during iterative training. We further inspect how clustering quality and model performance form a positive circle across iterations. To measure the clustering quality, we adopt the *purity* metrics, which is given by,

$$purity(C, L) = \frac{1}{N} \sum_{k} \max_{j} |c_k \cap l_j|, \tag{11}$$

where N is the total number of instances, $C = \{c_1, c_2, ..., c_K\}$ is the clustering results and $L = \left\{l_1, l_2, ..., l_J\right\}$ is the ground-truth classes, respectively. The value of purity is between 0 and 1, where 1 represents the perfect partition of data points. Fig. 3 (a) shows how the purity of coarse-grained clusters improves with respect to the iterations. Besides, we count the total number of outliers and the average size of clusters in each iteration as shown in Fig. 3 (b), (c). Taking Market-1501 dataset as an example, the initial purity of clusters is 0.39. Along with the training, data points belonging to the same class are gradually grouped together, ending up with 0.96 on purity. This suggests that the discriminativeness of the learned representations promotes gradually during the training process.



(a) purity of clusters

(b) total number of outliers

(c) average size of clusters

(d) mAP score

Figure 3: The purity, average size of cluster, total number of outliers and mAP with respect to the iterations

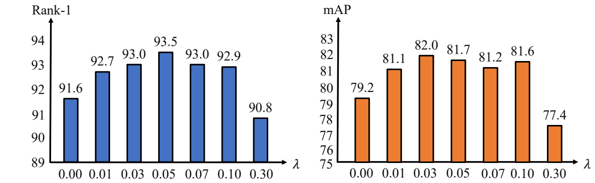


Figure 4: The influence of λ to Rank-1 accuracy and mAP score.

Moreover, Fig. 3 (d) further shows that the mAP score increases at the same pace as the clustering performance, demonstrating that the clustering and training promote each other simultaneously.

Balance factor λ. The balance factor λ in Eq. (10) affect re-identification performance. Fig. 4 shows experimental results with different λ on Market-1501. MASS performs best when $\lambda = 0.05$.

5 CONCLUSION

In this work, we propose a multi-gained active semi-supervised learning framework for scalable person re-identification (MASS) to minimize the annotation cost while retaining the satisfying performance. In order to efficiently select valuable image pairs for annotation, we design a cluster-scatter procedure with a centrality-based adaptive purification strategy to obtain a purified clustering. In order to effectively incorporate both labeled and unlabeled data, we train the model by using the clustering with a customized purification loss. After repeating the workflow iteratively, MASS achieves a high-quality clustering and a discriminative model. Extensive experiments show the effectiveness and efficiency of MASS.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025 and China Postdoctoral Science Foundation Funded Project under Grant 2020M671898.

REFERENCES

[1] Naoki Abe and Hiroshi Mamitsuka. 1998. Query Learning Strategies Using Boosting and Bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1–9.

[2] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 42–51.

[3] Phillip Bonacich. 1987. Power and centrality: A family of measures. *American journal of sociology* 92, 5 (1987), 1170–1182.

[4] Xinyuan Chang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2020. Transductive semi-supervised metric learning for person re-identification. *Pattern Recognition* 108 (2020), 107569.

[5] Abir Das, Rameswar Panda, and Amit Roy-Chowdhury. 2015. Active image pair selection for continuous person re-identification. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4263–4267.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Dongfang Du, Hao Wang, Tong Xu, Yanan Lu, Qi Liu, and Enhong Chen. 2017. Solving link-oriented tasks in signed network via an embedding approach. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 75–80.

[8] Sandra Ebert, Mario Fritz, and Bernt Schiele. 2012. Ralf: A reinforced active learning formulation for object class recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3626–3633.

[9] Chanho Eom and Bumsub Ham. 2019. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems*. 5297–5308.

[10] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 6112–6121.

[11] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2020. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. *arXiv preprint arXiv:2006.02713* (2020).

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Bingyu Hu, Jiawei Liu, and Zheng-jun Zha. 2021. Adversarial Disentanglement and Correlation Network for Rgb-Infrared Person Re-Identification. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[15] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. 2020. Real-world person re-identification via degradation invariance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14084–14094.

[16] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5070–5079.

[17] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3143–3152.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, 3–12.

[20] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2018. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*. 737–753.

[21] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2019. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[22] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. 2020. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3390–3399.

[23] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7202–7211.

[24] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–19.

[25] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*. 192–196.

[26] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. 2019. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 6122–6131.

[27] Massimo Marchiori and Vito Latora. 2000. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications* 285, 3-4 (2000), 539–546.

[28] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. 2020. Progressive Cross-camera Soft-label Learning for Semi-supervised Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[29] PengZhen Ren, Yun Xiao, XiaoJun Chang, Po-Yao Huang, Zhihui Li, XiaoJiang Chen, and Xin Wang. 2020. A survey of deep active learning. In *arXiv:2009.00236.*

[30] Phill Kyu Rhee, Enkhbayar Erdenee, Shin Dong Kyun, Minhaz Uddin Ahmed, and Songguo Jin. 2017. Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research* 45 (2017), 109–123.

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*. Springer, 17–35.

[32] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. 2019. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8934–8943.

[33] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 299–315.

[34] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 5972–5981.

[35] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised Person Re-identification via Multi-label Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10981–10990.

[36] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. 2020. Smoothing Adversarial Domain Attack and P-Memory Reconsolidation for Cross-Domain Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10568–10577.

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 274–282.

[38] Hao Wang, Enhong Chen, Qi Liu, Tong Xu, Dongfang Du, Wen Su, and Xiaopeng Zhang. 2018. A united approach to learning sparse attributed network embedding. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 557–566.

[39] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. 2016. Human-in-the-loop person re-identification. In *European Conference on Computer Vision*. Springer, 405–422.

[40] Hao Wang, Tong Xu, Qi Liu, Defu Lian, Enhong Chen, Dongfang Du, Han Wu, and Wen Su. 2019. MCNE: an end-to-end framework for learning multiple conditional network representations of social network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1064–1072.

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79–88.

[42] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. 2019. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing* 28, 6 (2019), 2872–2881.

[43] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3733–3742.

[44] Xiaomeng Xin, Jinjun Wang, Ruji Xie, Sanping Zhou, Wenli Huang, and Nanning Zheng. 2019. Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition* 88 (2019), 285–297.

[45] Xiaomeng Xin, Jinjun Wang, Ruji Xie, Sanping Zhou, Wenli Huang, and Nanning Zheng. 2019. Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition* 88 (2019), 285–297.

[46] Xiaomeng Xin, Xindi Wu, Yuechen Wang, and Jinjun Wang. 2019. Deep Self-Paced Learning for Semi-Supervised Person Re-Identification Using Multi-View Self-Paced Clustering. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2631–2635.

[47] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. 2020. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13657–13665.

[48] Zheng-Jun Zha, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua. 2011. Interactive video indexing with statistical active learning. *IEEE Transactions on Multimedia* 14, 1 (2011), 17–27.

[49] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9021–9030.

[50] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. 2021. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3538–3546.

[51] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5310–5319.

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[53] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2138–2147.

[54] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 598–607.

[55] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. 2019. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*. 6002–6012.