Unsupervised Clustering Active Learning for Person Re-identification

Wenjing Gao
wenjinggao@njust.edu.cn
Minxian Li*
minxianli@njust.edu.cn

School of Computer Science and Engineering
Nanjing University of Science and Technology

Abstract

Supervised person re-identification (re-id) approaches require a large amount of pairwise manual labeled data, which is not applicable in most real-world scenarios for re-id deployment. On the other hand, unsupervised re-id methods rely on unlabeled data to train models but performs poorly compared with supervised re-id methods. In this work, we aim to combine unsupervised re-id learning with a small number of human annotations to achieve a competitive performance. Towards this goal, we present a Unsupervised Clustering Active Learning (UCAL) re-id deep learning approach. It is capable of incrementally discovering the representative centroid-pairs and requiring human annotate them. These few labeled representative pairwise data can improve the unsupervised representation learning model with other large amounts of unlabeled data. More importantly, because the representative centroid-pairs are selected for annotation, UCAL can work with very low-cost human effort. Extensive experiments demonstrate the superiority of the proposed model over state-of-the-art active learning methods on three re-id benchmark datasets.

1 Introduction

In recent years, person re-identification (re-id) attracted lots of research attentions because of its practical applications on public security and smart city [11, 19, 29, 45]. Person re-identification aims to recognize the same identity of person across non-overlapped cameras, which is a challenging task in computer vision. Because of the non-overlap of ID labels between training and test set, re-id methods aims to learn a discriminative feature representation model for each person image. Despite promising results reported in previous works, re-id relies heavily the acquisition of ID labels for each person image. Unlike the labelling process for general categories which only requires each image to be labeled, *the acquisition of ID labels need to annotate all pairs of person images, which costs huge human effort.*

Supervised re-id methods can achieve encouraging performances but require a large number of manually labeled identity matching image pairs. However, the manual labeling for all person image pairs is a tedious and expensive process in re-id task. The cost of human labeling effort increases tremendously with the size of camera network and the number of persons in each camera. One the other hand, unsupervised re-id methods can be trained by abundant

[*] Corresponding author, Email: minxianli@njust.edu.cn

This work is supported by National Natural Science Foundation of China (Project No. 62076132), and Natural Science Foundation of Jiangsu (Project No. BK20211194).

unlabeled person images but significantly inferior in re-id accuracy. Without cross-view pairwise ID labels, the re-id model is not able to learn the discriminative feature representation for the significant appearance change across cameras.

To save human labeling effort, active learning aims to select a subset of samples that are the most representative for labeling, and then use them to train the model in order to achieve a competitive performance compared with fully supervised models. Specially, unsupervised active learning (UAL) [16] aims to select representative samples from totally no labeled samples. It is also called early active learning (EAL) [27]. UAL/EAL re-id methods [22, 23, 25, 31, 34] assume that no labeled image pairs are available in training set in the beginning. Pairwise constraint minimization [22], triangle-free subgraph maximization [31], and reinforcement learning [25] techniques are adopted to select a subset of pairwise samples for querying labeling. After selecting a small number of representative data for labeling, existing unsupervised active learning use these pair annotation to train re-id model with reliable supervised models. However, *all the above active learning re-id methods train the re-id model only from the selected image pairs, losing sight of the rest unlabeled samples*. Actually, the unlabeled data also benefit to re-id model to learn discriminative feature representation by exploring the association from the similar samples.

To overcome the above limitations, we consider a unsupervised clustering based active learning re-id framework. Unsupervised clustering based deep model can offer a reliable data structure by generating pseudo labels without any human annotation. Based on unsupervised clustering model, unsupervised active learning aims to select the representative sample pairs to reorganize the cluster structure. The advantages are two-fold: (1) Based on the clustering algorithm, unsupervised active learning method can easily and efficiently find the most representative sample pairs in the global space. (2) Depending on the representative pairs selection, the reorganization of the cluster structure can help the feature representation learning in a more effective way.

In this work, the main contributions are as follows: (1) We formulated an **U**nsupervised **C**lustering **A**ctive **L**earning (**UCAL**) model for person re-identification. This model combines jointly both unsupervised learning and active learning principles in an integrated learning framework. To the best of our knowledge, *this is the first attempt at active learning with unsupervised learning person re-id model.* (2) We propose an effective active learning strategy by the means of selecting the representative centroid-pairs from unsupervised clustering structure. (3) We present a clustering reorganization method (i.e. splitting/merging) to maximize the effect of active learning, so it takes the low-cost human labeling labor.

Extensive comparative experiments demonstrate the advantages of UCAL over the state-of-the-art active learning re-id approaches on three popular benchmarks: Market-1501[45], DukeMTMC-ReID[29, 46], and MSMT17[36]. Especially, the proposed UCAL model achieves the best performance with the lowest annotation cost.

2 Related Work

Unsupervised Learning in Re-ID. According to whether auxiliary data is used in training stage, unsupervised learning methods can be devided into two groups: (1) Pure unsupervsed learning. Most existing unsupervised learning re-id methods [2, 20, 21, 33, 42] adopt clustering-based approaches to produce pseudo labels, and then update the feature representation model by these pseudo labels. The challenge is how to obtain the precise cluster structure and how to alleviate the negative impact by noisy pseudo labels. (2) Unsupervised domain adaptation (UDA). UDA approaches aim to transfer the learned knowledge from a labeled source domain to an unlabeled target domain. The methods can be classified into three groups: Source domain pre-trained methods [7, 8, 40, 43, 44], Image-synthesis based methods [3, 5, 14, 32, 36], and Joint-learning based methods [9, 10, 28, 47, 48]. Generally, the performance of unsupervised learning approaches is limited, because of the lack of the pair relation supervision.

Semi-Supervised Learning in Re-ID. Semi-supervised learning methods [1, 8, 17, 18, 24, 37, 41] train the re-id model both on pre-labeled data and unlabeled data. Liu et al.[24] proposed a coupled dictionary learning by randomly pre-labeling one-third training data. Most of these methods [1, 17, 18, 37, 41] work on one-shot learning setting. *One-shot learning requires selecting all or most person identities, and then labeling one image or tracklet for each identity*. Bak et al. [1] learned a texture metric and a color metric on each camera pair by a one-shot metric learning approach. Ye et al. [41] proposed a label estimation approach to learn feature representations by using the pre-labeled data to formulate an anchor graph. Wu et al. [37] initialized a CNN model using pre-labeled data per ID, and then adopted a step-wise learning approach to update the CNN model. Li et al. [17, 18] used pre-labeled within-camera tracklet per ID to initialize a deep model, and then incrementally discover cross-camera tracklet association to improve the representation capability of deep model.

Although semi-supervised learning methods improve re-id performance by one-shot labeling strategy, this labeling setting is actually not practical for re-id task. In practice, the total number of individual ID is hard to know, not mention to label one instance for each ID.

Active Learning and Human-in-the-loop in Re-ID. For reducing the annotation cost in a more practical way, *pair-wise labeling* re-id approaches are proposed. According to the different applied stages, these methods can be divided into two categories, including active learning methods [22, 23, 25, 26, 31, 34] and human-in-the-loop methods [4, 27, 35]: (1) Active learning re-id methods aim at selecting a small number of image pairs to query human labeling in the training stage. Liu et al. [22] proposed an early active learning algorithm (EALPC) with a pairwise constraint to select the most representative samples for labeling. Roy et al. [31] presented a pairwise training subset selection framework to minimize human annotation effort. Liu et al. [25] designed a deep reinforcement active learning (DRAL) model to minimize human effort in the training stage. (2) Human-in-the-loop re-id methods focus on optimising the ranking list of every probe by human feedback directly in the test stage. Liu et al. [27] learned a post-rank function for re-ordering the rank list during a re-identification process. Wang et al. [35] proposed a distance metric learning method to incrementally optimise each new probe by human annotation in the deployment stage.

Compared with unsupervised learning and semi-supervised re-id methods,active learning and human-in-the-loop re-id methods reasonably add a handful of pair-wise labels to improve the performance. This paper follows active learning scheme, and our objective is to minimize human labeling effort to improve the performance by jointly unsupervised learning and active learning.

3 Methodology

In this section, we introduce the overall framework of our Unsupervised Clustering Active Learning (UCAL) method for person re-identification. An overview of the proposed UCAL model is depicted in Fig. 1. Given N unlabeled training person images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_N\}$ extracted from multiple camera views. The training objective is to select M pairs of unlabeled

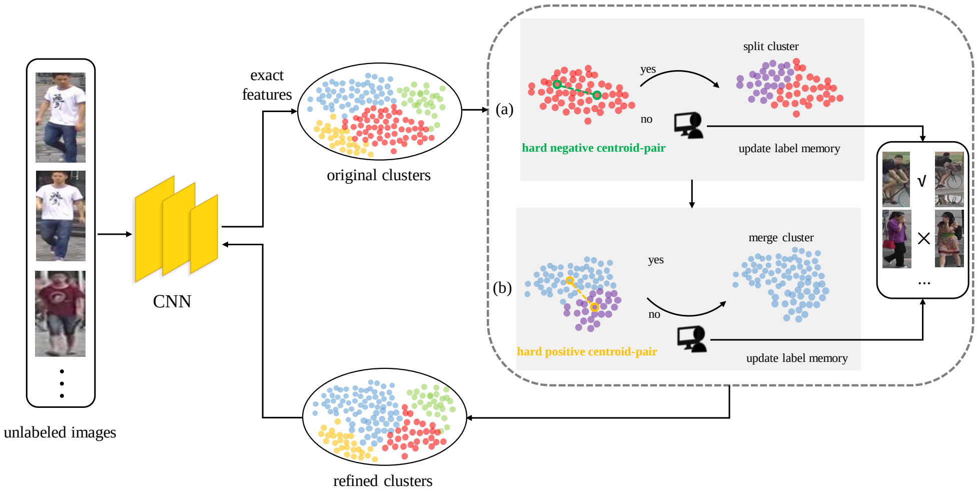


Figure 1: An overview of the proposed *Unsupervised Clustering Active Learning* (UCAL) method for re-id model learning. The UCAL takes as unlabeled person images from all the camera views. The objective is to derive a person discriminative feature representative feature representation module by unsupervised active learning. To this end, we formulate the UCAL model (see Sec. 3.2) with (a) Split by Negative centroid-Pair Selection (SNPS) module and (b) Merge by Positive centroid-Pair Selection (MPPS) module in an unsupervised clustering framework. Different colours correspond to different clusters. Best viewed in colour.

data to ask for human's labelling and learn a discriminative person re-id model. The proposed UCAL framework is driven by both of unlablled data and labeled data. In particular, we first adopt an unsupervised clustering method to discover the data distribution structure. And then, we select the key *centroid-pairs* from the clustering structure by *Negative centroid-Pair Selection (SNPS)* module and *Merge by Positive centroid-Pair Selection (MPPS)* module, and annotate the positive/negative relationships for these pairs. At last, the model is updated by the re-organized clustering structure in an iterative way to learn a discriminative feature representation.

3.1 Base Unsupervised Clustering Model

We adopt a DBSCAN based unsupervised clustering method [6, 10] as the base clustering framework. In this framework, a CNN-based [12] encoder f_θ is used to represent person image feature. Given the unlabeled training data $\mathcal{I}$, we adopt a self-paced clustering strategy [10] to group the data into clusters. According to the results of clustering, we update the parameters θ of the feature encoder f_θ. We enforce $||\boldsymbol{x}|| = 1$ via a L2-normalization layer. The loss function is an unsupervised contrastive learning which is similar with [10, 38], but computed only by the unlabeled data.

$$Loss(\boldsymbol{x}) = -log \frac{exp(\boldsymbol{x}^T \boldsymbol{c}_k/\tau)}{\sum_{i=1}^{n} exp(\boldsymbol{x}^T \boldsymbol{c}_i/\tau)} \quad (1)$$

where $\boldsymbol{x}^T\boldsymbol{c}_k$ expresses the similarity between the sample $\boldsymbol{x}$ and the k-th cluster centroid $\boldsymbol{c}_k$, n is the number of clusters, and τ is a temperature parameter that controls the concentration of the distribution [13].

3.2 Unsupervised Active Learning Framework

The base unsupervised clustering algorithm inevitably generates the incorrect pseudo labels, which will be harmful to the CNN model. For refining the clustering results, we aim to explore two kinds of representative sample pairs from the clustering structure: ***hard negative centroid-pair*** and ***hard positive centroid-pair***. The *hard negative centroid-pair* is the pair between two group centroids which have two IDs but wrongly grouped into the same cluster. The *hard positive centroid-pair* is the pair between two cluster centroids which have the same ID but wrongly grouped into two different clusters. After obtaining these two kinds of sample pairs candidates, positive/negative relationships are required to label by human experts. After that, two categories of manipulations: **split** and **merge** are adopted to refine the current clustering structure. The CNN model is then updated by the refined clustering result.

3.2.1 Split by Hard Negative Centroid-Pair Selection

A single cluster generated the base clustering algorithm may contain some small groups belonged to different ground truth IDs. This kind of relationship between these groups is called *hard negative*. This result gives incorrectly the same pseudo label to these individual samples grouped into one cluster, which be harmful to update the feature representation CNN model. To alleviate this problem, we propose a Split by hard Negative centroid-Pair Selection (SNPS) method to mine the hard negative centroid-pairs within the same cluster.

Given n_t clusters $\mathcal{C} = \{C_1, C_2, \cdots, C_{n_t}\}$ at epoch t, we use k-medoids [15] algorithm to generate k group candidates for each cluster. For j-th group G_j, $\boldsymbol{g}_j$ represents the centroid of G_j (Fig.1(a)). However, the real number of groups which have negative relations depends on the different clusters. That is, k is not fixed for each cluster. Therefore, we design a reliability criterion to decide k for each cluster. As discussed in [10], a reliable cluster should have two properties: high independence and high compactness. Thus, we aim to find k groups for each cluster with high independence and high compactness as splitting candidates. According to this criterion, an algorithm is proposed to decide k as follow:

$$comp_j = \min Sim_{intra}(G_j) \in [0,1] \tag{2}$$

$$indep_j = 1 - \max Sim_{inter}(G_j, G) \in [0,1] \tag{3}$$

$$k^* = \mathop{\arg\max}_{k \in [2, k_{max}]} \sum_{j=1}^{k} comp_j \times indep_j \tag{4}$$

where Sim_{intra} represents the similarities between samples within group G_j, Sim_{inter} represents the similarities between samples from group G_j and samples from other groups, k_{max} is set to $\frac{\sqrt{|C_i|/2}}{2}$.

Discussion. Although the similar concept of independence and compactness is proposed in [10], the objective is quite different in this paper. The objective of independence and compactness in [10] is to make each cluster more reliable. In SNPS, we aim to obtain a *reliable splitting structure* by measuring the independence and compactness of several groups controlled by k. Moreover, the criterion of good independence and compactness in [10] is decided by the density threshold, which is difficult to choose appropriately. In this paper, we adopt the maximum product of independence and compactness to select the most reliable splitting structure.

3.2.2 Merge by Hard Positive Centroid-Pair Selection

The individual samples which have the same person ID are very likely separated into different clusters by the clustering algorithm. This improper clustering structure generates the different pseudo labels to those clusters which should be grouped into the same cluster. This kind of relationship between these clusters is called *hard positive*. The existence of hard positive pairs harms the representation ability of CNN model. To mine the hard positive relationship between clusters, we propose a **M**erge by hard **P**ositive centroid-**P**iar **S**election (**MPPS**) method.

Given n_t clusters $\mathcal{C} = \{C_1, C_2, \cdots, C_{n_t}\}$ at epoch t, the cluster centroid of cluster C_i is defined as $\boldsymbol{c}_i$, and $s(\boldsymbol{c}_i, \boldsymbol{c}_j) \in [0,1]$ represents the similarity between $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$. For each centroid $\boldsymbol{c}_i$, a rank list $[s_1, s_2, \cdots, s_l, \cdots, s_{l_{max}}]$ is computed by the similarity between $\boldsymbol{c}_i$ and other centroids *from high to low*. l_{max} is the maximum number of candidate clusters. If all centroid-pairs are provided to human experts, the cost will be very high. To lower the labeling cost, we design a measure function to decide how many the similar centroid-pairs are selected to labeling. We take the similarity difference between the adjacent centroids in the rank list as the measure value.

$$\boldsymbol{d} = \{d_l = s_l - s_{l+1} | l \in [1, l_{max} - 1]\} \tag{5}$$

$$d_l^* = \frac{d_l - min(\boldsymbol{d})}{max(\boldsymbol{d}) - min(\boldsymbol{d})} \in [0,1] \tag{6}$$

where d_l^* is the min-max normalization value.

Inspired by [30], the high value of d_l^* indicates a large margin of similarity difference between s_l and s_{l+1}. It means the positive probabilities of pair c_i-c_l and pair c_i-c_{l+1} are uncertainty. So, pair c_i-c_l and pair c_i-c_{l+1} need to be labeled by users. We select the first l centroids with $\boldsymbol{c}_j$ as the labeling pairs when $d_l^* > \delta$. For the centroid-pairs labelled as positive, the corresponding clusters will be merged to one cluster, which is given the same pseudo label during the current training epoch.

3.3 Overall Model Training

We adopt a self-paced clustering strategy [10] as the base clustering algorithm, and the loss function (Eq. (1)) for the CNN parameters updating. During the first 15 training epochs, the above pure unsupervised method works as the initialized model. After the first 15 epochs of the training process, we deploy SNPS and MPPS modules to minimise the negative effect of unstable clustering structure. Specifically, the UCAL model first deploy SNPS module (Sec. 3.2.1) to find the hard negative centroid-pairs and split the original cluster into some smaller clusters. And then, we deploy MPPS module (Sec. 3.2.2) to find the hard positive centroid-pairs and merge these clusters to a bigger one. To lower the labeling cost, we also design a *label memory* (Fig. 1) to record the positive/negative relationships between centroids labeled by human expert. If the relationship between two centroids has been labeled in the label memory, the labeling request of this relationship is no need in the subsequent labeling process.

4 Experiments

4.1 Datasets and Evaluation Protocol

Datasets. To evaluate the proposed UCAL model, we reports results on three large person re-identification datasets: Market-1501[45], DukeMTMC-ReID[46], MSMT17[36].

(1) Market-1501[45]: Market-1501 is widely used large-scale re-id dataset. It contains 32,668 images of 1,501 person identities from 6 camera views. There are 12,936 images of 751 identities in training set, and 3,368 queries of 750 identities are used as the query set to search the true match among the remained 19,732 images.

(2) DukeMTMC-ReID[46]: DukeMTMC-ReID is another one of the most popular large scale re-id dataset which consists 36,411 pedestrian images from 1,812 person identities captured from 8 different cameras. Specifically, 16,522 images (702 identities) are adopted for training, 2,228 (702 identities) images are used as query to be retrieved from the remaining 17,661 images (1,110 identities).

(3) MSMT17[36]: MSMT17 is a larger and more challenging dataset, which contains 4,101 identities and 126,441 pedestrian images. There are 32,621 images of 1,041 identities in training set, and 93,820 images of 3,060 identities in test set. In the test set, 11,659 pedestrian images are randomly selected as probe images, and the other 82,161 images are treated as gallery images.

Evaluation Protocol. In the experiments, we used the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics to measure the methods' performance. The cost of human labeling effort is computed as follow:

$$cost = \frac{M}{N*(N-1)/2} * 100\% \tag{7}$$

where N represents the number of unlabeled training samples, and M represents the labeled pairs from human experts.

4.2 Implementation Details

We adopt an ImageNet pre-trained ResNet-50 [12] as the backbone for our UCAL model. After pooling-5 layer,we remove subsequent layers and add a 1D BatchNorm layer and an L2-normalization layer to derive the feature representations. Person bounding box images are resized to 256×128 as input. To ensure each training mini-batch has person images from all cameras, we set the batch size to 64 for all datasets mentioned in this paper. We adopted Adam optimiser with a weight decay of 0.0005, the learning rate is initialized to 3.5×10^{-4}. We train the model for 50 epochs, and the learning rate is divided by 10 after every 20 epochs. From the 15th epoch and every subsequent epoch,we use SNPS and MPPS modules to generate labeled pairs. By default, we set the maximum number of clusters that can be merged in each epoch to 20% of the total number of clusters and $\delta = 0.3$. All the experiments on three datasets follow the same settings as above.

Table 1: Comparison of proposed UCAL approach with state-of-the-art unsupervised, domain adaptation, semi-supervised and active learning approaches on Market-1501, DukeMTMC-ReID, and MSMT17.

	Methods		Market-1501 [45]					DukeMTMC-ReID [46]					MSMT17 [36]				
			R1	R5	R10	mAP	cost(%)	R1	R5	R10	mAP	cost(%)	R1	R5	R10	mAP	cost(%)
unsup	OIM[39]	CVPR'17	38.0	58.0	66.3	14.0	0	24.5	38.8	46.0	11.3	0	-	-	-	-	-
	BUC[20]	AAAI'19	66.2	79.6	84.5	38.3	0	47.4	62.6	68.4	27.5	0	-	-	-	-	-
	SSL[21]	CVPR'20	71.7	83.8	87.4	37.8	0	52.5	63.5	68.9	28.6	0	-	-	-	-	-
	MMCL[33]	CVPR'20	80.3	89.4	92.3	45.5	0	65.2	75.9	80.0	40.2	0	35.4	44.8	49.8	11.2	0
	HCT[42]	CVPR'20	80.0	91.6	95.2	56.4	0	69.6	83.4	87.4	50.7	0	-	-	-	-	-
domain	PTGAN[36]	CVPR'18	38.6	-	66.1	-	0	27.4	-	50.7	-	0	10.2	-	24.4	2.9	0
	ECN[48]	CVPR'19	75.1	87.6	91.6	43.0	0	63.3	75.8	80.4	40.4	0	25.3	36.3	42.1	8.5	0
	SSG[8]	ICCV'19	80.0	90.0	92.4	58.3	0	73.0	80.6	83.2	53.4	0	32.2	-	51.2	13.3	0
	MMCL$_{trans}$[33]	CVPR'20	84.4	92.8	95.0	60.4	0	72.4	82.9	85.0	51.4	0	43.6	54.3	58.9	16.2	0
	SPCL[10]	NIPS'20	89.7	96.1	97.6	77.5	0	-	-	-	-	-	53.7	65.0	69.8	26.8	0
semi	EUG[37]	CVPR'18	49.8	66.4	72.7	22.5	-	45.2	59.2	63.4	24.5	-	-	-	-	-	-
	TAUDL[17]	ECCV'18	63.7	-	-	41.2	-	61.7	-	-	43.5	-	28.4	-	-	12.5	-
	UTAL[18]	PAMI'19	69.2	-	-	46.2	-	62.3	-	-	44.6	-	31.4	-	-	13.1	-
	SSG++[8]	ICCV'19	86.2	94.6	96.5	68.7	-	76.0	85.8	89.3	60.3	-	41.6	-	62.2	18.3	-
active	TMA[26]	ECCV'16	47.9	-	-	22.3	13.58	-	-	-	-	-	-	-	-	-	-
	DRAL[25]	ICCV'19	**84.2**	**94.3**	**96.6**	**66.3**	**0.15**	**74.3**	**84.8**	**88.4**	**56.0**	**0.12**	-	-	-	-	-
	DRAL$_{upper}$[25]	ICCV'19	88.0	95.3	96.8	73.3	100	78.0	88.7	91.6	60.9	100	-	-	-	-	-
ours	Baseline	this paper	87.1	95.0	96.5	69.9	0	77.4	87.1	90.7	60.5	0	44.9	57.9	63.1	20.3	0
	Supervised	this paper	94.2	98.2	98.9	83.4	100	84.6	92.2	94.1	71.1	100	71.3	84.2	88.0	45.5	100
	UCAL	this paper	**91.8**	**96.8**	**97.9**	**78.2**	**0.08**	**81.2**	**89.7**	**92.5**	**66.3**	**0.06**	**63.2**	**75.1**	**79.4**	**35.7**	**0.03**

4.3 Comparisons with State-Of-The-Art Methods

In this section, we compare the proposed UCAL model with sixteen state-of-the-art re-id methods, including unsupervised learning (OIM[39], BUC[20], SSL[21], MMCL[33], HCT[42]), unsupervised domain adaptation (PTGAN[36], ECN[48], SSG[8], MMCL$_{trans}$[33], SPCL[10]), semi-supervised learning (EUG[37], TAUDL[17], UTAL[18], SSG++[8]), and active learning (TMA[26], DRAL[25]) models. The rank-(1,5,10) matching accuracy(%) and mAP(%) performance evaluated on Market-1501 [45], DukeMTMC-ReID [46], MSMT17 [36] are showed in Table 1. Moreover, the cost of human labeling effort computed by Eq.(7) is also given. The baseline model is trained by the unsupervised clustering method (Sec. 3.1) without any labeling cost. The supervised model adopt the same loss function (Eq. (1)) for training but using ground truth label instead of pseudo label. The experimental results show three observations as follows.

(1) The proposed UCAL model outperforms all competitors of active learning models with significant margins both on performance and cost. For example, the mAP margin is 11.9%(78.2-66.3) on Market1501 and 10.3%(66.3-56.0) on DukeMTMC-ReID. More importantly, the labeling cost of UCAL model is only 0.08% on Market1501 while TMA[26]'s cost is 13.58% and DRAL[25]'s cost is 0.15%, and 0.06% on DukeMTMC-ReID while DRAL[25]'s cost is 0.12%.

(2) Compared with state-of-the-art unsupervised/domain adaptation/semi-supervised learning models, the performance of our model is superior but with very low cost. Although there is no labeling cost under the above setting, the performance is limited especially on the large scale benchmark such as MSMT17. As discussed in Section 2, the labeling cost of semi-supervised learning is not able to estimated. Compared with the state-of-the-art unsupervised learning model, our UCAL model improves the performance of rank1 and mAP on MSMT17 by 9.5%(63.2-53.7) and 8.9%(35.7-26.8) with 0.03% cost.

(3) The proposed UCAL model narrows the gap between active learning and supervised learning models. Specifically, compared with the full labeling cost (100%) requirement of supervised learning model, by our UCAL model, the gap of rank1/mAP is narrowed to -2.4%(91.8-94.2) and -5.2%(78.2-83.4) with 0.08% cost on Market1501, -3.4%(81.2-84.6) and -4.8%(66.3-71.1) with 0.06% cost on DukeMTMC-ReID, -8.1%(63.2-71.3) and -9.8%(35.7-45.5) with 0.03% cost on MSMT17.

4.4 Component Analysis and Discussion

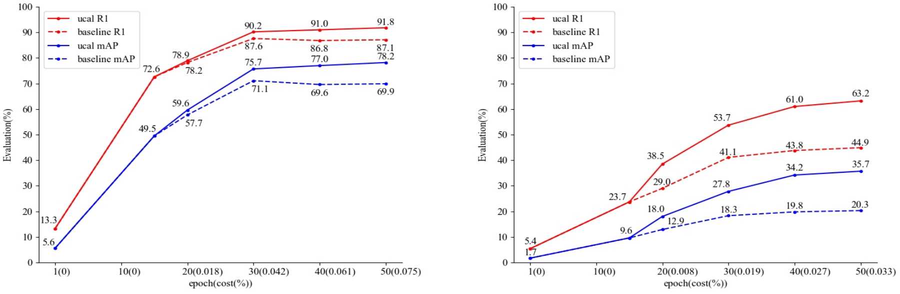
We conducted detailed UCAL model component analysis on two large person re-id datasets, Market1501 and MSMT17.

Effect of SNPS and MPPS Modules. We started by testing the performance impact of SNPS module and MPPS module. Based on the baseline model (Section 3.1), we firstly tested our SNPS component and MPPS component separately. Then, we combined these two components and tested the final performance. Table 2 shows that, the proposed SNPS and MPPS are both superior over the baseline model. After combining these two modules together to refine the clustering result, the UCAL model achieves mAP gain of 7.2 percent (78.2-71.0) and 13.6 percent (34.8-21.2) on Market1501 and MSMT17 respectively. This validates the proposed idea of split and merge modules, which refines the clusering structure by mining hard negative centroid-pairs and hard positive centroid-pairs.

Table 2: Effect of SNPS component and MPPS component.

Methods	Market-1501 [45]					MSMT17 [36]				
	R1	R5	R10	mAP	cost(%)	R1	R5	R10	mAP	cost(%)
Baseline	87.1	94.6	96.4	71.0	0	46.3	60.0	65.6	21.2	0
SNPS	90.5	96.4	97.7	76.8	0.027	53.4	66.1	71.0	26.7	0.010
MPPS	88.9	96.1	97.6	72.2	0.046	56.0	68.7	73.9	30.2	0.019
SNPS+MPPS	**91.8**	**96.8**	**97.9**	**78.2**	0.075	**63.2**	**75.1**	**79.4**	**35.7**	0.033

Effect with Labeling Cost. To further examine how well the proposed labeling strategy enables more discriminative re-id model learning, we tracked the improvement of UCAL compared with the base clustering model throughout the training. Fig. 2 shows that, along with the increase of the labeling cost, the improvement of UCAL becomes more and more obvious versus the base clustering model.



(a) Market1501

(b) MSMT17

Figure 2: The improvement of ucal compared with the base clustering model during training on (a) Market1501 and (b) MSMT17 benchmarks. The labeling cost of UCAL is depicted in the in brackets. Best viewed in colour.

5 Conclusion

We presented a novel *Unsupervised Clustering Active Learning* (UCAL) model for active learning based person re-identification. The model improves the feature representation ability of deep model dramatically with low labeling cost. This is achieved optimising jointly both Split by hard Negative centroid-Pair Selection (SNPS) module and Merge by hard Positive centroid-Pair Selection (MPPS) module by in a unified architecture. Extensive evaluations were conducted on three large-scale person re-id benchmarks to validate the advantages of the proposed UCAL model over state-of-the-art active learning re-id methods.

References

[1] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2990–2999, 2017.

[2] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. *Proc. Bri. Mach. Vis. Conf.*, 2018.

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 232–242, 2019.

[4] Abir Das, Rameswar Panda, and Amit Roy-Chowdhury. Active image pair selection for continuous person re-identification. In *IEEE Int. Conf. on Img. Proc.*, pages 4263–4267, 2015.

[5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 994–1003, 2018.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[7] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–18, 2018.

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6112–6121, 2019.

[9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *Proc. Int. Conf. on Learn. Rep.*, 2020.

[10] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Proc. Neur. Info. Proc. Sys.*, 2020.

[11] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 9527–9536, 2019.

[15] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[16] Changsheng Li, Handong Ma, Zhao Kang, Ye Yuan, Xiao-Yu Zhang, and Guoren Wang. On deep unsupervised active learning. *Proc. Int. Jo. Conf. of Artif. Intell.*, 2020.

[17] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proc. Eur. Conf. Comput. Vis.*, pages 737–753, 2018.

[18] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(7):1770–1782, 2019.

[19] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 152–159, 2014.

[20] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conf. on Art. Intel.*, 2019.

[21] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3390–3399, 2020.

[22] Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. Early active learning with pairwise constraint for person re-identification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 103–118, 2017.

[23] Wenhe Liu, Xiaojun Chang, Ling Chen, Dinh Phung, Xiaoqin Zhang, Yi Yang, and Alexander G Hauptmann. Pair-based uncertainty and diversity promoting early active learning for person re-identification. *ACM Transactions on Intelligent Systems and Technology*, 11(2):1–15, 2020.

[24] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3550–3557, 2014.

[25] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6122–6131, 2019.

[26] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification. In *Proc. Eur. Conf. Comput. Vis.*, pages 858–877, 2016.

[27] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Early active learning via robust representation and structured sparsity. In *Proc. Int. Jo. Conf. of Artif. Intell.*, pages 1572–1578, 2013.

[28] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1306–1315, 2016.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshop of Eur. Conf. Comput. Vis.*, pages 17–35, 2016.

[30] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496, 2014.

[31] Sourya Roy, Sujoy Paul, Neal E Young, and Amit K Roy-Chowdhury. Exploiting transitivity for learning person re-identification models on a budget. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7064–7072, 2018.

[32] Yingzhi Tang, Yang Xi, Nannan Wang, Bin Song, and Xinbo Gao. Cgan-tm: A novel domain-to-domain transferring method for person re-identification. *IEEE Trans. Img. Proc.*, 29:5641–5651, 2020.

[33] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10981–10990, 2020.

[34] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Highly efficient regression for scalable person re-identification. In *Proc. Bri. Mach. Vis. Conf.*, pages 1–8, 2016.

[35] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *Proc. Eur. Conf. Comput. Vis.*, pages 405–422, 2016.

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 79–88, 2018.

[37] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5177–5186, 2018.

[38] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3733–3742, 2018.

[39] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3376–3385. IEEE, 2017.

[40] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *AAAI Conf. on Art. Intel.*, volume 34, pages 12597–12604, 2020.

[41] Mang Ye, Xiangyuan Lan, and Pong C Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proc. Eur. Conf. Comput. Vis.*, pages 170–186, 2018.

[42] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 13657–13665, 2020.

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

[44] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 8222–8231, 2019.

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1116–1124, 2015.

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3754–3762, 2017.

[47] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proc. Eur. Conf. Comput. Vis.*, pages 172–188, 2018.

[48] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 598–607, 2019.