Discrepant and Multi-instance Proxies for Unsupervised Person Re-identification

Chang Zou[1] Zeqi Chen[2] Zhichao Cui[3] Yuehu Liu[2*] Chi Zhang[2]
[1]School of Software Engineering, Xi'an Jiaotong University
[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University [3]Chang'an University

Abstract

Most recent unsupervised person re-identification methods maintain a cluster uni-proxy for contrastive learning. However, due to the intra-class variance and inter-class similarity, the cluster uni-proxy is prone to be biased and confused with similar classes, resulting in the learned features lacking intra-class compactness and inter-class separation in the embedding space. To completely and accurately represent the information contained in a cluster and learn discriminative features, we propose to maintain discrepant cluster proxies and multi-instance proxies for a cluster. Each cluster proxy focuses on representing a part of the information, and several discrepant proxies collaborate to represent the entire cluster completely. As a complement to the overall representation, multi-instance proxies are used to accurately represent the fine-grained information contained in the instances of the cluster. Based on the proposed discrepant cluster proxies, we construct cluster contrastive loss to use the proxies as hard positive samples to pull instances of a cluster closer and reduce intra-class variance. Meanwhile, instance contrastive loss is constructed by global hard negative sample mining in multi-instance proxies to push away the truly indistinguishable classes and decrease inter-class similarity. Extensive experiments on Market-1501 and MSMT17 demonstrate that the proposed method outperforms state-of-the-art approaches.

1. Introduction

Unsupervised person re-identification (Re-ID) aims to retrieve images of a particular person across camera views and scenes without annotations [35, 48]. Most unsupervised methods adopt a two-step alternating training scheme: 1) generating pseudo labels by k-nearest neighbor search [34, 42] or clustering [15, 13, 27, 43, 8]; 2) training the model based on a uni-proxy (*i.e.*, cluster centroid [9] or learnable weight [13]) of each cluster. However, due to the intra-class variance and inter-class similarity caused by the changeable human pose, illumination, and camera views [54], a uni-proxy /is often biased and confused, failing to fully and accurately describe the information of a cluster. As a result, the features learned based on the uni-proxy are not compact and have unclear cluster boundaries in the embedding space, which in turn affects the quality of clustering. In order to learn discriminative features, CAP [36] subdivides each cluster to obtain multiple camera-aware proxies, pulling an instance (*i.e.*, sample) closer to all proxies in the cluster to alleviate intra-class variance. The later works ICE [2] and PPLR [7] adopt the same strategy. Although these methods improve the compactness of clusters, they depend on extra labels and ignore the intra-class variance caused by factors other than camera views. On the other hand, several works [46, 14, 7] focus on reducing inter-class similarity to learn discriminative features. They consider performing batch hard negative sample mining [20] to promote inter-class separation. However, as shown in Figure 1, due to the randomness of sampling, the negative samples selected for a query from the mini-batch may not be true hard negatives in the global embedding space, and therefore cannot enlarge the inter-class separation of actual indistinguishable classes.

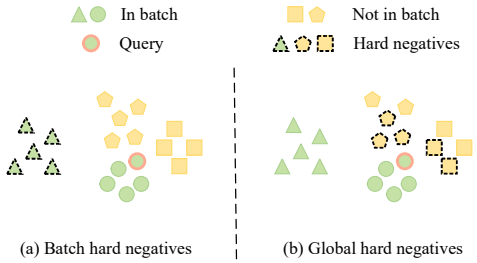
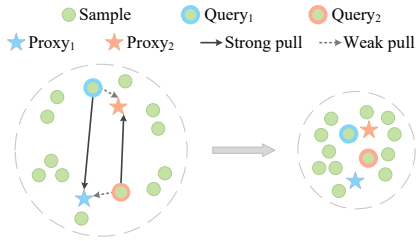


Figure 1. An illustration demonstrating that global hard negatives are more effective than batch hard in promoting inter-class separation. Different shapes represent different classes. (a) The batch hard negatives are the ones easy to distinguish. b) Our global hard negatives are the truly hardest and most informative samples of indistinguishable classes.

To reduce intra-class variance without relying on additional annotations, we propose to use several discrepant cluster proxies to complementarily represent a cluster. Each

*Corresponding author



(a) Without intra-class compactness (b)With intra-class compactness

Figure 2. An illustration showing that discrepant cluster proxies make a loosely distributed cluster gain intra-class compactness. Through the strong pull generated by each proxy for its low-similarity queries and the collaboration of two proxies, the whole cluster eventually gets more compact.

proxy concentrates on representing a portion of the information and the whole cluster is fully represented by several discrepant proxies. We obtain discrepant cluster proxies simply by updating the same cluster centroid with different update designs. Based on the cluster proxies, we propose cluster contrastive loss to increase the compactness of the clusters. As shown in Figure 2, Proxy_1 and Proxy_2 are the corresponding hard positive sample and easy positive sample for Query_1 according to pairwise similarity. Thus, contrastive loss enables Proxy_1 to generate a strong pull on Query_1 and Proxy_2 to generate a weak pull, resulting in Query_1 being closer to Proxy_1 after model optimization. Similarly, Query_2 will be closer to Proxy_2. As a result, Query_1 and Query_2 will become closer. Since the proxies are updated by these closer queries, Proxy_1 and Proxy_2 will also approach with training. Through the collaboration of two discrepant proxies, the cluster gradually gains intra-class compactness.

On the other hand, to further effectively decrease inter-class similarity while reducing intra-class variance, we propose to maintain finer-grained and more accurate multi-instance proxies through instance features of a cluster as the supplement of coarse-grained cluster proxies. Distinguished from the previous batch hard sample mining, the hard negative samples of a query are selected among the multi-instance proxies of all other classes with a global view. Then we exploit the true hard negatives to construct instance contrastive loss and purposefully increase the inter-class variance of indistinguishable classes.

Our contributions can be summarized as follows:

- We propose contrastive learning based on discrepant cluster proxies, which complementarily represent a cluster and collaboratively reduce intra-class variation.
- We propose global hard negative sample mining based on multi-instance proxies to select truly hard and informative negative samples to purposefully increase the inter-class variance of indistinguishable classes.
- Extensive experimental results with superior performance against the state-of-the-art methods demonstrate the effectiveness of the proposed method.

2. Related Work

Unsupervised Person Re-ID. Existing unsupervised methods can be roughly categorized into unsupervised domain adaptive (UDA) methods and purely unsupervised learning (USL) methods. UDA methods [13, 15, 14, 31, 44, 26, 35, 11, 52, 1, 21] transfer the knowledge learned from the labeled source domain to the unlabeled target domain. In contrast, USL methods [28, 34, 27, 43, 36, 41, 7, 46, 25, 45] is trained directly on unlabeled target datasets. Our method meets the more challenging USL setting. Recently, USL methods that generate pseudo labels by clustering and perform contrastive learning on cluster proxies have made great progress. SpCL [15] averages the instance features of a class in the memory bank as a uni-proxy for the class. Cluster-Contrast [9] directly stores a uni-proxy for each cluster to maintain the updating consistency. However, the cluster uni-proxy cannot effectively reduce the existing intra-class variance. Thus, CAP [36] forms multiple camera-aware proxies for each cluster to alleviate the camera domain gap. MCRN [39] stores multi-centroid representations for a cluster but only selects one as the proxy for a query to mitigate the effects of mixed clusters. Unlike these methods, we obtain several discrepant cluster proxies to completely represent a cluster and serve as hard positive samples to collaboratively enhance intra-class compactness.

Hard Sample Mining. Hard sample mining can improve training speed and performance [49]. Many recent unsupervised Re-ID methods utilize hard batch sample mining [20] to increase intra-class compactness and inter-class separation. MMT [14] and PPLR [7] learn hard samples by constructing softmax-triplet loss on the hardest positive and negative sample pairs. ICE [2] mines the hardest positive sample in the mini-batch and takes all samples of other identities as negatives to reduce intra-class variance. ISE [46] explores the hardest positive and negative samples among the original and generated samples within a batch. However, hard sample mining in the mini-batch does not consider global information of all classes. Therefore, we propose global hard negative sample mining based on multi-instance proxies to effectively enhance the inter-class variance among classes hard to discriminate.

Contrastive Learning. Contrastive learning [17, 6, 5, 32, 40, 16, 37] aims at maximizing the similarity of representations obtained from different distorted versions of a sample [16]. MoCo [17] builds a queued dictionary to keep an abundance of negative samples and introduces a momentum encoder to ensure their consistency. We perform both cluster-level and instance-level contrastive learning based on discrepant cluster proxies and multi-instance proxies. Like MoCo, We use a momentum encoder to keep the consistency of negative samples.

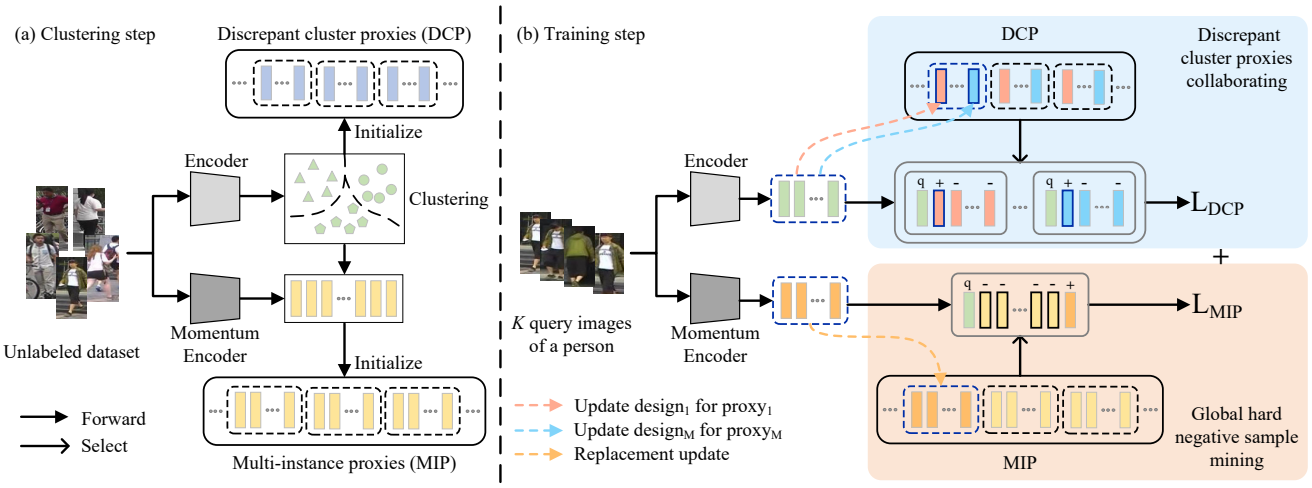


Figure 3. An overview framework of the proposed DCMIP. It alternates two steps. (a) In the clustering step, we cluster encoder-encoded features to generate pseudo labels. We then initialize the discrepant cluster proxies (DCP) with cluster centroids of these features and initialize the multi-instance proxies (MIP) with randomly selected K instance features encoded by the momentum encoder f_{θ_m} for each cluster. (b) In the training step, we exploit the hard positive proxies in DCP and hard negative proxies in MIP by $\mathcal{L}_{DCP}$ and $\mathcal{L}_{MIP}$ to perform discrepant cluster proxies collaborating and global hard negative sample mining, respectively. Then, different update designs are applied to encoder-encoded features to update DCP, and all instance features encoded by f_{θ_m} are used to update MIP.

3. Method

3.1. Overview

Given an unlabeled person Re-ID dataset $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}$, where x_i is the i-th image and N_D is the number of images. For the USL Re-ID task, the objective is to train a robust network f_θ to project a sample x_i in the data space $\boldsymbol{\mathcal{D}}$ to a feature $f_\theta(x_i)$ in the embedding space $\boldsymbol{\mathcal{F}}$.

Recently, most unsupervised Re-ID methods [15, 9, 46, 36, 2] generate pseudo labels by DBSCAN [12] algorithm. After DBSCAN clustering, the unlabeled dataset $\mathcal{D}$ becomes $\mathcal{D}' = \{(x_i, y_i)\}_{i=1}^{N'_\mathcal{D}}$, where $y_i \in \{1, 2, \ldots, C\}$ is the pseudo label of the i-th image. N'_D is the number of images after discarding outliers and C is the number of clusters. Then a memory bank $\mathcal{M}$ is constructed to store proxies for clusters. Since the cluster centroid contains average information, recent methods [9, 46] simply use it as the uni-proxy for a cluster. Based on the proxies, the InfoNCE loss function [32] is applied for model optimization. Despite there are also different variants of proxies [15, 53, 36], we summarize their general formulation as follows:

$$\mathcal{L}_{Info} = -\log \frac{\exp(\boldsymbol{q} \cdot \boldsymbol{p}^+/\tau)}{\sum_{i=1}^{N} \exp(\boldsymbol{q} \cdot \boldsymbol{p}_i/\tau)}, \tag{1}$$

where $\boldsymbol{q}$ is a query instance feature extracted by f_θ. $\boldsymbol{p}_i$ is the i-th proxy of selected N proxies from the memory bank $\mathcal{M}$. Among the N proxies, $\boldsymbol{p}^+$ shares the same pseudo label with $\boldsymbol{q}$. τ is a temperature factor. Since both $\boldsymbol{q}$ and $\boldsymbol{p}_i$ are L_2-normalized, the cosine similarity $\boldsymbol{q} \cdot \boldsymbol{p}_i$ is used as the similarity score between features.

When the model parameters are updated by gradient descent, the proxy $\boldsymbol{p}^+$ are also updated by the query $\boldsymbol{q}$:

$$\boldsymbol{p}^+ \leftarrow \mu \cdot \boldsymbol{p}^+ + (1-\mu) \cdot \boldsymbol{q}, \tag{2}$$

where μ is a momentum factor.

In this paper, we propose a contrastive learning framework based on discrepant cluster proxies and multi-instance proxies (DCMIP) as shown in Figure 3. As above, we extract the features of the training set by encoder f_θ and generate pseudo labels through DBSCAN. The difference is that we simultaneously maintain cluster proxies and multi-instance proxies for a cluster, and construct contrastive loss at both the cluster and instance levels.

Due to the large number of instance proxies, we introduce a momentum encoder f_{θ_m} following MoCo [17] to maintain the consistency of negative instance proxies. The update of the momentum encoder is formulated as follows:

$$\theta_m^t = \alpha \theta_m^{t-1} + (1-\alpha)\theta^t, \tag{3}$$

where α is the momentum coefficient that controls the updated speed and is set to 0.999. The momentum encoder f_{θ_m} evolves more smoothly, so the instance features encoded by f_{θ_m} are more consistent. Note that, the cluster proxies are initialized and updated with the encoder-encoded features, while the instance proxies are initialized and updated with instance features encoded by f_{θ_m}.

3.2. Discrepant Cluster Proxies

We argue that the cluster uni-proxy tends to focus on the common information of a class and fails to reflect the intra-

class variance that exists. To solve this problem, we propose to maintain *discrepant cluster proxies* (DCP) to complementarily represent a cluster and improve the compactness of the cluster based on these discrepant proxies.

Memory initialization. For each cluster, we maintain M cluster proxies in the memory bank $\mathcal{M}$. For all proxies of the *j*-th cluster, we initialize them with the cluster centroid $c_j = \frac{1}{|\mathcal{H}_j|}\sum_{x_i \in \mathcal{H}_j} x_i$, where $\mathcal{H}_j$ denotes the *j*-th cluster and $|\cdot|$ denotes the number of instances in it. Thus, the memory bank $\mathcal{M} \in \mathbb{R}^{C\times M\times d}$ has $C \times M$ entries, and d is the dimension of the features.

Memory update. Previous studies [22, 55] found that the hardness of positive and negative samples is crucial for contrastive learning. The gradient of InfoNCE loss (Eq. 1) corresponding to query $\boldsymbol{q}$ is:

$$\frac{\partial \mathcal{L}_{Info}}{\partial \boldsymbol{q}} = -\frac{1}{\tau}\left(\left(1-\mathcal{P}^{+}\right)\cdot \boldsymbol{p}^{+} - \sum_{\boldsymbol{p}^{-}\in\mathcal{N}_q} \mathcal{P}^{-}\cdot \boldsymbol{p}^{-}\right), \tag{4}$$

where $\mathcal{P}^{+/-} \in [0,1]$ is the matching probability distribution between query $\boldsymbol{q}$ and the positive/negative proxy $\boldsymbol{p}^{+}/\boldsymbol{p}^{-}$, *i.e.*, $\mathcal{P}^{+/-} = \frac{\exp(\boldsymbol{q}\cdot\boldsymbol{p}^{+/-}/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{q}\cdot\boldsymbol{p}_i/\tau)}$. $\mathcal{N}_q$ denotes the set of $N-1$ negative proxies other than positive $\boldsymbol{p}^{+}$. We can find that a hard positive sample with low similarity to the query tends to produce a larger gradient, generating a stronger pull to draw the query closer. But only employing such a single proxy to represent a cluster is biased and may affect the learning of inter-class relationships. Therefore, we propose to use several discrepant proxies for a cluster.

To obtain discrepant cluster proxies, we momentum update each of the M identically initialized proxies of a cluster as Eq. 2 by different feature vectors from the current mini-batch. For the *m*-th proxy $\boldsymbol{p}_{i,m}$ of the *i*-th cluster, the feature vector can be obtained in several ways:

$$\boldsymbol{q}_{mean} \leftarrow \frac{1}{K}\sum_{\boldsymbol{q}\in Q^i}\boldsymbol{q}, \tag{5}$$

$$\boldsymbol{q}_{rand} \leftarrow \boldsymbol{q}_j, \boldsymbol{q}_j \in Q^i, \tag{6}$$

$$\boldsymbol{q}_{hard} \leftarrow \arg\min_{\boldsymbol{q}} \boldsymbol{q}\cdot\boldsymbol{p}_{i,m}, \boldsymbol{q}\in Q^i, \tag{7}$$

where Q^i is the sample feature set of the *i*-th cluster in current mini-batch. $\boldsymbol{q}_{mean}$ is the average feature of the set. $\boldsymbol{q}_{rand}$ is a randomly selected sample feature from Q^i. The selection probability is $\mathcal{P}_{(\boldsymbol{q}_{rand}=\boldsymbol{q}_j)} = \frac{1}{K}, j = 1,2,\ldots,K$, where K denotes the number of samples for an identity in the batch. $\boldsymbol{q}_{hard}$ is the sample feature which has lowest similarity with proxy $\boldsymbol{p}_{i,m}$. The three different vectors correspond to three different update designs for cluster proxies, which we name "Mean", "Rand" and "Hard", respectively.

In our experiments, we find that the optimum cluster proxies obtained by different update designs should not only be discrepant but also stable. The discrepancy of proxies ensures the hardness of positive samples, *i.e.*, the strength of the generated pull to the queries. The stability ensures that the pull direction of a proxy does not change drastically, otherwise, a proxy cannot form a stable pull, and a stable collaboration cannot be formed among several proxies. According to experimental results, maintaining two cluster proxies with update designs of "Mean"+"Hard" and "Mean"+"Rand" for Market-1501 and MSMT17 delivers the best performance by making a trading-off between high discrepancy and high stability. We further discuss the discrepancy and stability in Sec. 4.4.

Cluster contrastive loss. With M discrepant cluster proxies, we form a cluster contrastive loss as follows:

$$\mathcal{L}_{DCP} = -\frac{1}{M}\sum_{j=1}^{M}\log\frac{\exp\left(\boldsymbol{q}\cdot\boldsymbol{p}_j^{+}/\tau\right)}{\sum_{i=1}^{C}\exp\left(\boldsymbol{q}\cdot\boldsymbol{p}_{i,j}/\tau\right)}, \tag{8}$$

where $\boldsymbol{p}_{i,j}$ is the *j*-th proxy of the *i*-th cluster. $\boldsymbol{p}_j^{+}$ shares the same label with the query $\boldsymbol{q}$ and is the *j*-th proxy for that cluster. Note that the same update design is adopted for the *j*-th proxy of all clusters.

Several discrepant proxies complementarily represent a cluster, and collaboratively reduce intra-class variance to make the cluster compact.

3.3. Multi-Instance Proxies

Considering that the discrepant cluster proxies cannot reflect the valuable fine-grained information contained in the hard instances of the cluster, we further maintain multi-instance proxies (MIP) for each cluster to perform global hard negative sample mining.

Memory initialization. We randomly select K instance features encoded by the momentum encoder f_{θ_m} to initialize multi-instance proxies for each cluster. Note that K equals the number of images sampled for an identity in a mini-batch. Combining cluster proxies and instance proxies, the memory bank $\mathcal{M} \in \mathbb{R}^{C\times(M+K)\times d}$ has $C \times (M+K)$ entries in total.

Memory update. While updating the model parameters, the instance features of the current mini-batch are used to update the instance proxies as follows:

$$P^i \leftarrow Q_m^i, \tag{9}$$

where Q_m^i is the instance feature set of the *i*-th cluster in the mini-batch encoded by f_{θ_m} and P^i is the set of instance proxies of that cluster in the memory bank $\mathcal{M}$. Unlike the momentum update of the cluster proxies, the instance proxies are directly replaced by the K instances with the same label in the current mini-batch. This allows us to keep as many up-to-date instance proxies as possible to represent the fine-grained information of a cluster.

Instance contrastive loss. We compute the pairwise similarity of an input query to all instance proxies of other classes in the memory bank and rank them in descending order. We select the top-$\mathcal{N}$ most similar instance proxies as the global hardest negatives. Considering that the instance features in the current mini-batch are more up-to-date than those in the memory bank $\mathcal{M}$, and that the momentum encoder f_{θ_m} is more stable and more robust to label noise, we choose the instance feature in the batch encoded by f_{θ_m} with the lowest similarity to the query as the hard positive. Based on the hard positive and the $\mathcal{N}$ global hardest negatives, the following instance contrastive loss is constructed:

$$\mathcal{L}_{MIP} = -\log \frac{\exp\left(\boldsymbol{q}\cdot\boldsymbol{m}^{+}\right)/\tau}{\exp\left(\boldsymbol{q}\cdot\boldsymbol{m}^{+}\right)/\tau + \sum_{i=1}^{\mathcal{N}} \exp\left(\boldsymbol{q}\cdot\boldsymbol{p}_n^i\right)/\tau}, \tag{10}$$

where $\boldsymbol{m}^{+}$ is the hard positive and p_n^i is the i-th hard negative instance proxies. These hard negatives accurately increase the inter-class variance of indistinguishable classes in the global embedding space from the perspective of inter-instance relationships.

3.4. Overall Loss

We name the contrast learning framework based on discrepant cluster proxies and multi-instance proxies DCMIP. The overall loss function of DCMIP is:

$$\mathcal{L}_{DCMIP} = \begin{cases} \mathcal{L}_{DCP}, & \text{if } epoch \leqslant E_{ins} \\ \lambda\mathcal{L}_{DCP} + (1-\lambda)\,\mathcal{L}_{MIP} & \text{else} \end{cases}, \tag{11}$$

where λ is the loss weight. For $\mathcal{L}_{MIP}$, due to the poor quality of the representations in the early training stage, the hard samples at this point may be meaningless. Using these hard samples may lead to the model being trained in the wrong direction from the beginning [49]. Therefore, we set $E_{ins} = 20$ to start the instance-level contrastive learning from the 21st epoch, and the parameters of f_{θ_m} are initialized with the parameters of current f_θ. We also report the results starting from other epochs in Appendix A.1.

DCMIP enhances the quality of representations from both intra-class and inter-class relationships. The intra-class variance is reduced by using the cluster proxies as hard positive samples in cluster contrastive loss (Eq. 8), and the inter-class variance is increased by using the instance proxies as hard negative samples in instance contrastive loss (Eq. 10). This allows the model to learn discriminative features and in turn improve the clustering quality.

4. Experiments

4.1. Datasets and Evaluation Protocols

We evaluate our method on Market-1501 [47] and MSMT17 [38]. Market-1501 is collected with 6 cameras on the Tsinghua University campus and consists of 32,668 images of 1,501 person identities, with a training set of 12,936 images of 751 identities and a test set of 19,732 images of 750 identities. MSMT17 is a more challenging dataset, using 15 cameras for data collection and consisting of 126,441 images of 4,101 identities, with a training set of 32,621 images of 1,041 identities and a test set of 93,820 images of 3,060 identities. Both Cumulative Matching Characteristics (CMC) Top-1, Top-5, Top-10 accuracies and mean Average Precision (mAP) are adopted in our experiments.

4.2. Implementation Details

We adopt ResNet50 [18] pre-trained on ImageNet [10] as the backbone. Following Cluster-Contrast [9], the generalized mean pooling [30] is used for the final pooling layer. The input image size is 320×128. At the beginning of each epoch, we use DBSCAN clustering to generate pseudo labels. The maximum distance between two samples in DBSCAN is set to 0.45 for Market-1501 and 0.7 for MSMT17. The mini-batch size is 256 consisting of 16 identities and 16 images for each identity. From the 21st epoch, we start instance-level contrastive learning and $K = 16$ instance proxies are maintained for each cluster. In instance contrastive loss (Eq. 10), we select $\mathcal{N} = 256$ negative instance proxies for each query and set the loss weight $\lambda = 0.5$ (Eq. 11). The update momentum μ of cluster proxies is set to 0.1 (Eq. 2). The temperature hyper-parameter τ in the two losses (Eq. 8, Eq. 10) is set to 0.05. We use an Adam [23] optimizer with weight decay of 5×10^{-4}. The initial learning rate is set to 3.5×10^{-5} and divided by 10 every 20 epochs. For both datasets, we train 50 epochs. After training, the momentum encoder f_{θ_m} is used for inference. We also provide the analysis for the maximum distance of DBSCAN and the loss weight λ in Appendix A.1.

4.3. Ablation Study

In this subsection, to analyze the effectiveness of the proposed components, we conduct extensive experiments on Market-1501 and MSMT17. We adopt the method that uses the cluster centroid as the uni-proxy of a cluster and updates the uni-proxy by the design of "Mean" as our *baseline*.

Effectiveness of the discrepant cluster proxies (DCP). Note that for Market-1501 and MSMT17, we maintain two discrepant cluster proxies and use the update designs of "Mean"+"Hard" (Eq. 5, Eq. 7) and "Mean"+"Rand" (Eq. 5, Eq. 6), respectively. As shown in Table 1, our DCP significantly exceeds the baseline using the uni-proxy, especially +4.8%/+1.5% mAP/top-1 improvement on Market-1501 and +3.2%/+2.7% mAP/top-1 improvement on MSMT17. It demonstrates that complementary and collaborative discrepant cluster proxies can describe clusters more comprehensively, therefore contributing to learning good sample representations more than the cluster uni-proxy.

Method	Market-1501		MSMT17	
	mAP	top-1	mAP	top-1
Baseline	81.0	92.8	35.2	66.1
Baseline+DCP	85.8	94.3	38.4	68.8
Baseline+MIP	84.5	93.9	39.2	68.3
DCMIP	**86.7**	**94.7**	**40.9**	**69.3**

Table 1. Ablation studies on proposed components of DCMIP.

Method	Market-1501		MSMT17	
	mAP	top-1	mAP	top-1
DCMIP w/o MIP	85.4	93.7	37.5	68.0
+ MIP	**86.7**	**94.7**	**40.9**	**69.3**
+ Batch hard triplet	86.1	93.9	37.9	66.0
+ Batch hard instance	86.0	94.4	38.4	66.2

Table 2. Ablation studies on different hard sample mining techniques. In all rows, we keep the weight λ of cluster contrastive loss in total loss (Eq. 11) as 0.5 from the 21st epoch.

Effectiveness of the multi-instance proxies (MIP). To demonstrate the effectiveness of MIP, we combine MIP with the baseline and DCP respectively. In Table 1, comparing to the baseline, mAP/top-1 of Baseline+MIP is improved by 3.5%/1.1% on Market-1501 and 4.0%/2.2% on MSMT17. DCMIP (DCP+MIP) increases mAP/top-1 of Baseline+DCP by 0.9%/0.4% and 2.5%/0.5% on Market-1501 and MSMT17 severally. This demonstrates that for both cluster uni-proxy and multi-proxies, global hard negative mining based on MIP can capture the fine-grained information contained in truly hard instances in the global embedding space. In Table 2, we compare MIP with two batch hard sample mining techniques. One way is hard batch triplet mining [20], which forms a triplet with the anchor, the hardest positive, and the hardest negative in the mini-batch. The other way is batch hard instance mining [2], which uses the most similar instance of the same class and all instances of other classes in the mini-batch as the positive and the negatives. As the results show, global hard negative sample mining based on MIP outperforms the two techniques. This demonstrates that our MIP overcomes the limitation of batch hard sample mining by exploiting the hardest negative instance proxies to purposefully increase the inter-class variance of indistinguishable classes.

DCMIP combines DCP and MIP for contrastive learning based on both cluster proxies and instance proxies. Compared with the cluster uni-proxy baseline, our method improves mAP/top-1 by 5.7%/1.9% on Market-1501 and 5.7%/3.2% on MSMT17 by a large margin. We believe that DCMIP can reduce intra-class variance through the collaboration of discrepant cluster proxies and increase inter-class variance through global hard negative mining based on multi-instance proxies.

Clustering quality. To intuitively demonstrate the ability of our method to reduce intra-class variation and inter-class similarity, we visualized randomly selected samples of 20 classes by t-SNE [33]. As shown in Figure 4, the compactness of all classes is significantly improved in DCMIP compared to the baseline. For several classes that are too close to distinguish in the baseline, our method increases their inter-class distances. Moreover, for the two classes with mixed features, DCMIP successfully separates them. We also report the results of cluster quality measured with four cluster evaluation metrics on Market-1501 and MSMT17 in Appendix A.2.

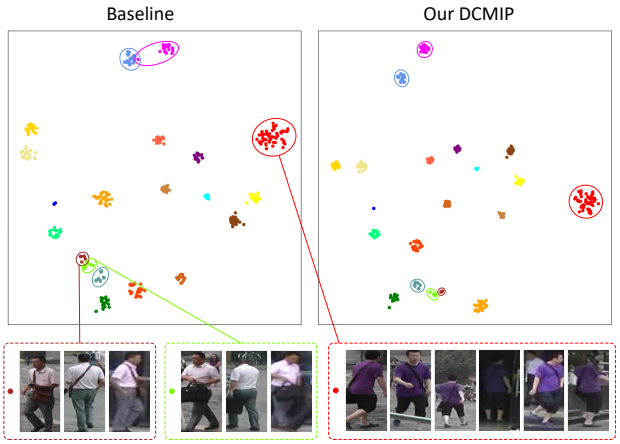


Figure 4. The t-SNE visualization of 20 random classes in Market-1501 between baseline and our DCMIP. Different colors represent different IDs. At the bottom, we show the case of small inter-class variance and large intra-class variance through partial person images of several identities.

Update policy	Market-1501		MSMT17	
	mAP	top-1	mAP	top-1
Mean	81.0	92.8	35.2	66.1
Rand	81.6	92.4	35.8	64.6
Hard	82.6	92.3	18.5	39.8
Mean+Hard	**85.8**	**94.3**	33.9	61.0
Mean+Rand	83.1	93.1	**38.4**	**68.8**
Rand+Hard	84.7	93.3	31.8	58.4
Mean+Rand+Hard	85.5	94.2	37.5	65.4

Table 3. Comparison of different update policies for cluster proxies.

4.4. Parameter Analysis

Different update policies for cluster proxies. We defined three update designs in Sec. 3.2 to update the cluster proxies: "Mean", "Rand", and "Hard", as shown in Eq. 5, Eq. 6, and Eq. 7, respectively. Several cluster proxies are obtained by applying different update designs to the same initial cluster centroid. The three designs can form seven different update policies by combination: "Mean", "Rand", "Hard", "Mean+Hard", "Mean+Rand", "Rand+Hard", and "Mean+Rand+Hard". As shown in Table 3, discrepant cluster proxies obtained by appropriate update policies outperform the uni-proxy, but the number of cluster proxies

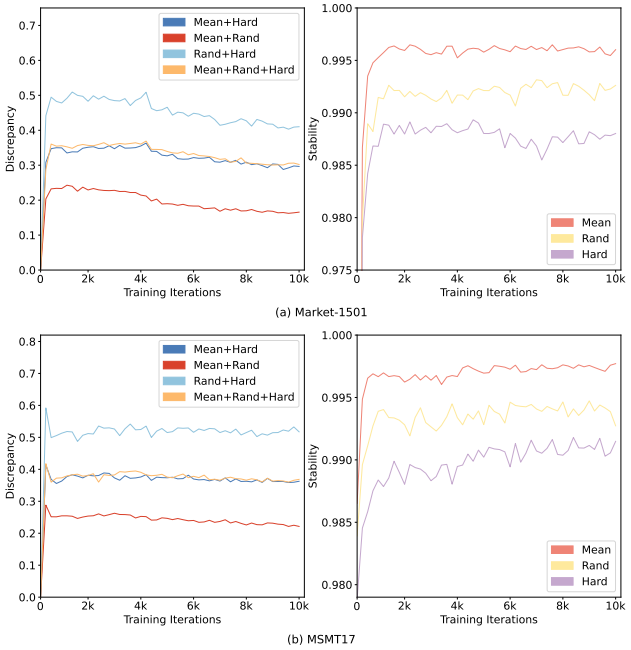


Figure 5. The discrepancy of cluster multi-proxies and the stability of uni-proxy on Market-1501 and MSMT17. The discrepancy is calculated by taking the mean cosine similarity of proxy pairs inside a cluster, averaging it over all clusters, and then subtracting that value from 1. The stability is computed by averaging the cosine similarity of a uni-proxy before and after being updated over all clusters, reflecting the smoothness of the update.

is not always the more the better. According to the results, Market-1501 and MSMT17 achieve the optimum with "Mean+Hard" and "Mean+Rand", respectively. Without specification, we use the two policies by default.

The discrepancy and stability of cluster proxies. As shown in Table 3, Market-1501 prefers the update design "Hard" in all policies, while MSMT17 does the opposite. The proxy updated by "Hard" has lower similarity to the instances of a cluster and can produce a larger gradient. However, due to the drastic updates, the design "Hard" is less stable than "Mean" and "Rand" in Figure 5. Considering that the clustering quality of MSMT17 is low (see Appendix A.2), the sample least similar to a proxy is very likely to be noise, updating with it may lead to incorrect learning direction. Conversely, the proxy of Market-1501 updated by "Hard" is more reliable due to the higher clustering quality. In addition, MSMT17 has a higher stability requirement than Market-1501 because it has about twice as many samples per class as Market-1501, which means that a proxy must be stable for a larger number of samples, and forming a stable collaboration is more difficult when using several proxies. Therefore, the design "Hard" behaves differently on two different-sized datasets. For Market-1501, "Mean+Hard" has a high discrepancy, and "Mean" complements the stability of "Hard", thus achieving the

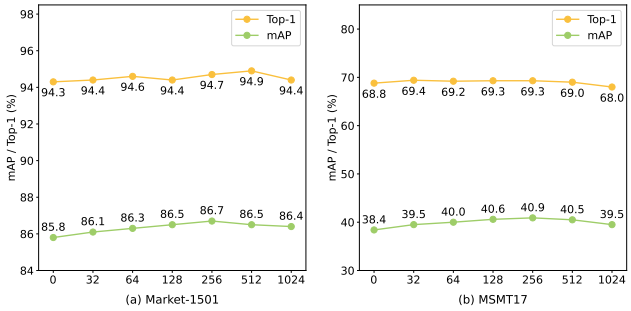


Figure 6. Parameter analysis of the number of hard negative instance proxies $\mathcal{N}$ on Market-1501 and MSMT17.

best balance between high discrepancy and high stability. For MSMT17, although "Mean+Rand" has a low discrepancy, it avoids the problem of "Hard" and forms discrepant proxies which can collaborate stably to achieve the best performance. We conjecture that since three dynamically changing proxies are more difficult to form a stable collaboration than two proxies, despite the high discrepancy and the ability to represent more information, the policy "Mean+Rand+Hard" is not optimal.

The number of hard negative instance proxies. We analyze the number of hard negative instance proxies $\mathcal{N}$ selected by global hard negative mining. In Figure 6, we can see that the performance of Market-1501 and MSMT17 firstly increases and then decreases as $\mathcal{N}$ raises. $\mathcal{N} = 0$ indicates the case of cluster-level contrastive learning based on DCP only. Both datasets achieve the best mAP when we set $\mathcal{N}$ to 256. We speculate that when $\mathcal{N} > 0$, we can effectively increase distances between classes hard to distinguish with the valuable globally hardest negative samples. However, as $\mathcal{N}$ grows, meaningless easy samples may be selected, instead reducing the matching probability of meaningful samples and affecting the gradient thus causing a decrease in performance. Therefore, we set $\mathcal{N}$ = 256.

4.5. Comparison with State-of-the-Arts

In Table 4, we compare our DCMIP with state-of-the-art Re-ID methods on Market-1501 and MSMT17. In an unsupervised setting, our DCMIP significantly outperforms previous methods. We achieve mAP/top-1 of 86.7%/94.7% and 40.9%/69.3% on Market-1501 and MSMT17, respectively. Compared to unsupervised methods without any labels, our discrepant cluster proxies and multi-instance proxies substantially improve mAP by 3.7% and 7.9% on Market-1501 and MSMT17 than the uni-proxy method Cluster-Contrast [9]. Moreover, our DCMIP surpasses the second-best method ISE [46] on Market-1501 and MSMT17 by 1.4% and 3.9% in mAP and outperforms ICE [2] and PPLR [7] on MSMT17 by a remarkable margin. Compared to unsupervised methods with camera labels, our DCMIP without any camera knowledge outperforms four

Method		Market-1501				MSMT17			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
Unsupervised methods with camera labels									
IICS [41]	CVPR'21	72.9	89.5	95.2	97.0	26.9	56.4	68.8	73.4
CAP [36]	AAAI'21	79.2	91.4	96.3	97.7	36.9	67.4	78.0	81.4
ICE [2]	ICCV'21	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4
PPLR [7]	CVPR'22	84.4	94.3	97.8	98.6	42.2	73.3	83.5	86.5
Unsupervised methods without any labels									
BUC [27]	AAAI'19	29.6	61.9	73.5	78.2	-	-	-	-
JVTC [24]	ECCV'20	41.8	72.9	84.2	88.7	15.1	39.0	50.9	56.8
MMCL [34]	CVPR'20	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8
HCT [43]	CVPR'20	56.4	80.0	91.6	95.2	-	-	-	-
SpCL [15]	NeurIPS'20	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61.2
JVTC+* [3]	CVPR'21	75.4	90.5	96.2	97.1	29.7	54.4	68.2	74.2
OPLG-HCD [50]	ICCV'21	78.1	91.1	96.4	97.7	26.9	53.7	65.3	70.2
ICE [2]	ICCV'21	79.5	92.0	97.0	98.1	29.8	59.0	71.7	77.0
MCRN [39]	AAAI'22	80.8	92.5	-	-	31.2	63.6	-	-
SECRET [19]	AAAI'22	81.0	92.6	-	-	31.3	60.4	-	-
PPLR [7]	CVPR'22	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
RMCL [29]	KBS'23	81.7	93.0	97.6	98.4	32.5	62.3	73.6	78.0
Cluster-Contrast [9]	ACCV'22	83.0	92.9	97.2	98.0	33.0	62.0	71.8	76.7
ISE [46]	CVPR'22	85.3	94.3	**98.0**	**98.8**	37.0	67.6	77.5	81.0
DCMIP (320×128)	This paper	**86.7**	**94.7**	**98.0**	**98.8**	**40.9**	**69.3**	**79.7**	**83.6**
Supervised methods									
DG-Net [51]	CVPR'19	86.0	94.8	-	-	52.3	77.2	-	-
ABD-Net [4]	ICCV'19	88.3	95.6	-	-	60.8	82.3	90.6	-
ISE (w/ ground truth) [46]	CVPR'22	87.8	95.6	98.5	99.2	51.0	76.8	87.1	90.6
DCMIP (w/ ground truth)	This paper	89.2	96.2	98.5	99.0	62.8	83.9	91.6	93.8

Table 4. Comparison with state-of-the-art methods on Market-1501 and MSMT17. The best results of unsupervised methods without any labels are marked in **bold**. Note that the input image size of DCMIP is 320×128.

methods (*i.e.*, IICS [41], CAP [36], ICE [2], PPLR [7]) on Market-1501 and three (*i.e.*, IICS [41], CAP [36], ICE [2]) on MSMT17 in mAP. In addition, under the supervised setting, our DCMIP achieves competitive performance to the well-known supervised method DG-Net [51] and ADB-Net [4]. It is worth noting that DCMIP with ground truth scores higher in mAP and top-1 than ISE [46] by 11.8% and 7.1% on MSMT17, which demonstrates the superiority and potential of our approach on large datasets.

5. Discussion

Our DCMIP reduces intra-class variation by two discrepant cluster proxies for all clusters, but this may not be the optimal solution. For clusters with high intra-class compactness, further reducing the intra-class variation is not necessary, as it may impair generalizability. For clusters with low intra-class compactness, more cluster proxies are required to represent diverse subsets and lower intra-class variance. We will explore additional strategies to obtain discrepant proxies as well as a dynamic cluster proxy number for different clusters in the future study.

6. Conclusion

In this paper, we propose a contrastive learning framework based on discrepant cluster proxies and multi-instance proxies for unsupervised person re-identification. We maintain two discrepant cluster proxies by different update designs to complementarily represent a cluster and act as hard positive samples in the cluster contrastive loss to collaboratively reduce intra-class variance. We also maintain multi-instance proxies for a cluster to accurately represent the fine-grained instance information. Then global hard negative sample mining is performed among the instance proxies to increase the inter-class variance of indistinguishable classes through the instance contrastive loss. Comprehensive experiments have shown that our framework outperforms prior state-of-art methods on two prevalent datasets.

Acknowledgment

This work was supported by the National Key Research and Development Project of New Generation Artificial Intelligence of China under Grant 2018AAA0102504. We also thank Xueming Qian for valuable advice.

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 189–205, 2018. 2

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14960–14969, 2021. 1, 2, 3, 6, 7, 8

[3] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2004–2013, 2021. 8

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361, 2019. 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[7] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based Pseudo Label Refinement for Unsupervised Person Re-identification, Mar. 2022. 1, 2, 7, 8

[8] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Transactions on Image Processing*, 30:7815–7829, 2021. 1

[9] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pages 1142–1160, 2022. 1, 2, 3, 5, 7, 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018. 2

[12] 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. 3

[13] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019. 1, 2

[14] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 1, 2

[15] Yixiao Ge, Feng Zhu, Dapeng Chen, and Rui Zhao. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020. 1, 2, 3, 8

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 879–887, 2022. 8

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2, 6

[21] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8526–8536, 2021. 2

[22] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 4

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *European Conference on Computer Vision*, pages 483–499. Springer, 2020. 8

[25] Mingkun Li, Chun-Guang Li, and Jun Guo. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 2022. 2

[26] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*, 2018. 2

[27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8738–8745, 2019. 1, 2, 8

[28] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3390–3399, 2020. 2

[29] Zhiqi Pang, Chunyu Wang, Junjie Wang, and Lingling Zhao. Reliability modeling and contrastive learning for unsupervised person re-identification. *Knowledge-Based Systems*, page 110263, 2023. 8

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 5

[31] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102:107173, 2020. 2

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3

[33] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014. 6

[34] Dongkai Wang and Shiliang Zhang. Unsupervised Person Re-Identification via Multi-Label Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10978–10987, Seattle, WA, USA, June 2020. IEEE. 1, 2, 8

[35] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 1, 2

[36] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, volume 2, page 4, 2021. 1, 2, 3, 8

[37] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 5

[39] Yuhang Wu, Tengteng Huang, Haotian Yao, Chi Zhang, Yuanjie Shao, Chuchu Han, Changxin Gao, and Nong Sang. Multi-centroid representation network for domain adaptive person re-id. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2750–2758, 2022. 2, 8

[40] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, Salt Lake City, UT, June 2018. IEEE. 2

[41] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11926–11935, 2021. 2, 8

[42] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2148–2157, 2019. 1

[43] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13654–13662, Seattle, WA, USA, June 2020. IEEE. 1, 2, 8

[44] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *European Conference on Computer Vision*, pages 594–611. Springer, 2020. 2

[45] Guoqing Zhang, Hongwei Zhang, Weisi Lin, Arun Kumar Chandran, and Xuan Jing. Camera contrast learning for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[46] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit Sample Extension for Unsupervised Person Re-Identification, Apr. 2022. 1, 2, 3, 7, 8

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[48] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[49] Wenzhao Zheng, Jiwen Lu, and Jie Zhou. Hardness-Aware Deep Metric Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(09):3214–3228, Sept. 2021. 2, 5

[50] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8371–8381, 2021. 8

[51] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 8

[52] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–188, 2018. 2

[53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person

re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2723–2738, 2020. 3

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 1

[55] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021. 4