Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID

Yixiao Ge **Dapeng Chen** **Feng Zhu** **Rui Zhao** **Hongsheng Li**
Multimedia Laboratory (MMLAB)
The Chinese University of Hong Kong
{yxge@link,hsli@ee}.cuhk.edu.hk

Abstract

Domain adaptive object re-ID aims to transfer the learned knowledge from the labeled source domain to the unlabeled target domain to tackle the open-class re-identification problems. Although state-of-the-art pseudo-label-based methods [9, 50, 47, 51, 10] have achieved great success, they did not make full use of all valuable information because of the domain gap and unsatisfying clustering performance. To solve these problems, we propose a novel self-paced contrastive learning framework with hybrid memory. The hybrid memory dynamically generates source-domain class-level, target-domain cluster-level and un-clustered instance-level supervisory signals for learning feature representations. Different from the conventional contrastive learning strategy, the proposed framework jointly distinguishes source-domain classes, and target-domain clusters and un-clustered instances. Most importantly, the proposed self-paced method gradually creates more reliable clusters to refine the hybrid memory and learning targets, and is shown to be the key to our outstanding performance. Our method outperforms state-of-the-arts on multiple domain adaptation tasks of object re-ID and even boosts the performance on the source domain without any extra annotations. Our generalized version on unsupervised person re-ID surpasses state-of-the-art algorithms by considerable **16.2%** and **14.6%** on Market-1501 and DukeMTMC-reID benchmarks. Code is available at https://github.com/yxgeee/SpCL.

1 Introduction

Unsupervised domain adaptation (UDA) for object re-identification (re-ID) aims at transferring the learned knowledge from the labeled source domain (dataset) to properly measure the inter-instance affinities in the unlabeled target domain (dataset). Common object re-ID problems include person re-ID and vehicle re-ID, where the source-domain and target-domain data do not share the same identities (classes). Existing UDA methods on object re-ID [36, 9, 50, 47, 51, 42] generally tackled this problem following a two-stage training scheme: (1) supervised pre-training on the source domain, and (2) unsupervised fine-tuning on the target domain. For stage-2 unsupervised fine-tuning, a pseudo-label-based strategy was found effective in state-of-the-art methods [9, 50, 47, 51], which alternates between generating pseudo classes by clustering target-domain instances and training the network with generated pseudo classes. In this way, the source-domain pre-trained network can be adapted to capture the inter-sample relations in the target domain with noisy pseudo-class labels.

Although the pseudo-label-based methods have led to great performance advances, we argue that there exist two major limitations that hinder their further improvements (Figure 1 (a)). (1) During the target-domain fine-tuning, the source-domain images were either not considered [9, 50, 47, 51] or were even found harmful to the final performance [10] because of the limitations of their methodology designs. The accurate source-domain ground-truth labels are valuable but were ignored during target-domain training. (2) Since the clustering process might result in individual outliers, to ensure

Preprint. Under review.

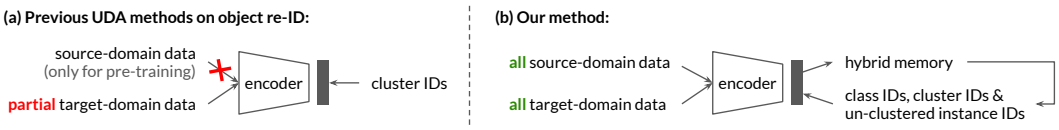

Figure 1: State-of-the-arts [9, 47, 51, 50] on UDA object re-ID discarded both the source-domain data and target-domain un-clustered data for training, while our proposed self-paced contrastive learning framework fully exploits all available data with hybrid memory for joint feature learning.

the reliability of the generated pseudo labels, existing methods [9, 47, 51, 10] simply discarded the outliers from being used for training. However, such outliers might actually be difficult but valuable samples in the target domain. Simply abandoning them might critically hurt the final performance.

To overcome the problems, we propose a *hybrid memory* to encode all available information from both source and target domains for feature learning. For the source-domain data, their ground-truth class labels can naturally provide valuable supervisions. For the target-domain data, clustering can be conducted to obtain relatively confident clusters as well as un-clustered outliers. All the source-domain class centroids, target-domain cluster centroids, and target-domain un-clustered instance features from the hybrid memory can provide supervisory signals for jointly learning discriminative feature representations across the two domains (Figure 1 (b)). A unified framework is developed for dynamically updating and distinguishing different entries in the proposed hybrid memory.

Specifically, since all the target-domain clusters and un-clustered instances are equally treated as independent classes, the clustering reliability would significantly impact the learned representations. We thus propose a *self-paced contrastive learning* strategy, which initializes the learning process by using the hybrid memory with the most reliable target-domain clusters. Trained with such reliable clusters, the discriminativeness of feature representations can be gradually improved and additional reliable clusters can be formed by incorporating more un-clustered instances into the new clusters. Such a strategy can effectively mitigate the effects of noisy pseudo labels and boost the feature learning process. To properly measure the cluster reliability, a novel multi-scale clustering reliability criterion is proposed, based on which only reliable clusters are preserved and other confusing clusters are disassembled back to un-clustered instances. In this way, our self-paced learning strategy gradually creates more reliable clusters to dynamically refine the hybrid memory and learning targets.

Our contributions are summarized as three-fold. (1) We propose a unified contrastive learning framework to incorporate all available information from both source and target domains for joint feature learning. It dynamically updates the hybrid memory to provide class-level, cluster-level and instance-level supervisions. (2) We design a self-paced contrastive learning strategy with a novel clustering reliability criterion to prevent training error amplification caused by noisy pseudo-class labels. It gradually generates more reliable target-domain clusters for learning better features in the hybrid memory, which in turn, improves clustering. (3) Our method significantly outperforms state-of-the-arts [9, 50, 47, 51, 42] on multiple domain adaptation tasks of object re-ID with up to **7.1%** mAP gains. The proposed unified framework could even boost the performance on the source domain by jointly training with un-annotated target-domain data, while most existing UDA methods "forget" the source domain after fine-tuning on the target domain. Our unsupervised version without labeled source-domain data on person re-ID task significantly outperforms state-of-the-arts [25, 42, 18] by **16.2%** and **14.6%** in terms of mAP on Martket-1501 and DukeMTMC-reID benchmarks.

2 Related Works

Unsupervised domain adaptation (UDA) for object re-ID. Existing UDA methods for object re-ID can be divided into two main categories, including pseudo-label-based methods [36, 47, 51, 9, 50, 57, 49, 42] and domain translation-based methods [6, 44, 3, 10]. This paper follows the former one since the pseudo labels were found more effective to capture the target-domain distributions. Though driven by different motivations, previous pseudo-label-based methods generally adopted a two-stage training scheme: (1) pre-training on the source domain with ground-truth IDs, and (2) adapting the target domain with pseudo labels. The pseudo labels can be generated by either clustering instance features [36, 47, 51, 9, 50] or measuring similarities with exemplar features [57, 49, 42], where the clustering-based pipeline maintains state-of-the-art performance to date. The major challenges faced by clustering-based methods is how to improve the precision of pseudo labels and how to mitigate the effects caused by noisy pseudo labels. SSG [47] adopted human local features to assign multi-scale

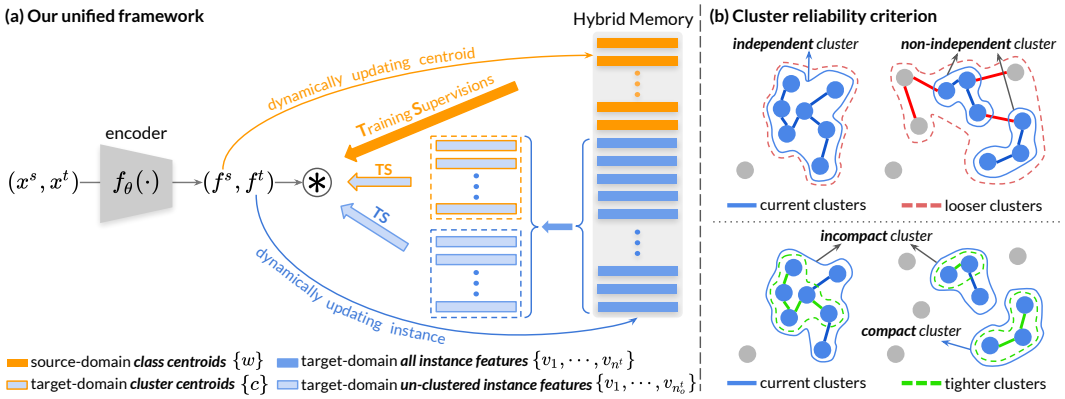

Figure 2: (a) The illustration of the proposed unified framework with a novel hybrid memory. (b) The proposed reliability criterion for measuring the cluster independence and compactness.

pseudo labels. PAST [51] introduced to utilize multiple regularizations alternately. MMT [9] proposed to generate more robust soft labels via the mutual mean-teaching. AD-Cluster [50] incorporated style-translated images to improve the discriminativeness of instance features. Although various attempts along this direction have led to great performance advances, they ignored to fully exploit all valuable information across the two domains which limits their further improvements, *i.e.*, they simply discarded both the source-domain labeled images and target-domain un-clustered outliers when fine-tuning the model on the target domain with pseudo labels.

Contrastive learning. State-of-the-art methods on unsupervised visual representation learning [30, 45, 15, 41, 59, 13, 2] are based on the contrastive learning. Being cast as either the dictionary look-up task [45, 13] or the consistent learning task [41, 2], a contrastive loss was adopted to learn instance discriminative representations by treating each unlabeled sample as a distinct class. Although the instance-level contrastive loss could be used to train embeddings that can be generalized well to downstream tasks with fine-tuning, it does not perform well on the domain adaptive object re-ID tasks which require to correctly measure the inter-class affinities on the unsupervised target domain.

Self-paced learning. The "easy-to-hard" training scheme is at the core of self-paced learning [19], which was originally found effective in supervised learning methods, especially with noisy labels [11, 16, 23]. Recently, some methods [39, 12, 4, 52, 60] incorporated the conception of self-paced learning into unsupervised learning tasks by starting the training process with the most confident pseudo labels. However, the self-paced policies designed in these methods were all based on the close-set problems with pre-defined classes, which cannot be generalized to our open-set object re-ID task with completely unknown classes on the target domain. Moreover, they did not consider how to plausibly train with hard samples that cannot be assigned confident pseudo labels all the time.

3 Methodology

To tackle the challenges in unsupervised domain adaptation (UDA) on object re-ID, we propose a self-paced contrastive learning framework (Figure 2 (a)), which consists of a CNN [20]-based encoder f_θ and a novel hybrid memory. The key innovation of the proposed framework lies in jointly training the encoder with all the source-domain class-level, target-domain cluster-level and target-domain un-clustered instance-level supervisions, which are dynamically updated in the hybrid memory to gradually provide more confident learning targets. In order to avoid training error amplification caused by noisy clusters, the self-paced learning strategy initializes the training process with the most reliable clusters and gradually incorporates more un-clustered instances to form new reliable clusters. A novel reliability criterion is introduced to measure the quality of clusters (Figure 2 (b)).

Our training scheme alternates between two steps: (1) grouping the target-domain samples into clusters and un-clustered instances by clustering the target-domain instance features in the hybrid memory with the self-paced strategy (Section 3.2), and (2) optimizing the encoder f_θ with a unified contrastive loss and dynamically updating the hybrid memory with encoded features (Section 3.1).

3.1 Constructing and Updating Hybrid Memory for Contrastive Learning

Given the target-domain training samples $\mathbb{X}^t$ without any ground-truth label, we employ the self-paced clustering strategy (Section 3.2) to group the samples into clusters and the un-clustered outliers.

The whole training set of both domains can therefore be divided into three parts, including the source-domain samples $\mathbb{X}^s$ with ground-truth identity labels, the target-domain pseudo-labeled data $\mathbb{X}_c^t$ within clusters and the target-domain instances $\mathbb{X}_o^t$ not belonging to any cluster, *i.e.*, $\mathbb{X}^t = \mathbb{X}_c^t \cup \mathbb{X}_o^t$. State-of-the-art UDA methods [9, 50, 47, 51] simply abandon all source-domain data and target-domain un-clustered instances, and utilize only the target-domain pseudo labels for adapting the network to the target domain, which, in our opinion, is a sub-optimal solution. Instead, we design a novel contrastive loss to fully exploit available data by treating all the source-domain classes, target-domain clusters and target-domain un-clustered instances as independent classes.

3.1.1 Unified Contrastive Learning

Given a general feature vector $\boldsymbol{f} = f_\theta(x), x \in \mathbb{X}^s \cup \mathbb{X}_c^t \cup \mathbb{X}_o^t$, our unified contrastive loss is

$$\mathcal{L}_{\boldsymbol{f}} = -\log \frac{\exp\left(\langle \boldsymbol{f}, \boldsymbol{z}^+ \rangle / \tau\right)}{\sum_{k=1}^{n^s} \exp\left(\langle \boldsymbol{f}, \boldsymbol{w}_k \rangle / \tau\right) + \sum_{k=1}^{n_c^t} \exp\left(\langle \boldsymbol{f}, \boldsymbol{c}_k \rangle / \tau\right) + \sum_{k=1}^{n_o^t} \exp\left(\langle \boldsymbol{f}, \boldsymbol{v}_k \rangle / \tau\right)}, \tag{1}$$

where $\boldsymbol{z}^+$ indicates the positive class prototype corresponding to $\boldsymbol{f}$, the temperature τ is empirically set as 0.05 and $\langle \cdot, \cdot \rangle$ denotes the inner product between two feature vectors to measure their similarity. n^s is the number of source-domain classes, n_c^t is the number of target-domain clusters and n_o^t is the number of target-domain un-clustered instances. More specifically, if $\boldsymbol{f}$ is a source-domain feature, $\boldsymbol{z}^+ = \boldsymbol{w}_k$ is the centroid of the source-domain class k that $\boldsymbol{f}$ belongs to. If $\boldsymbol{f}$ belongs to the k-th target-domain cluster, $\boldsymbol{z}^+ = \boldsymbol{c}_k$ is the k-th cluster centroid. If $\boldsymbol{f}$ is a target-domain un-clustered outlier, we would have $\boldsymbol{z}^+ = \boldsymbol{v}_k$ as the outlier instance feature corresponding to $\boldsymbol{f}$. Intuitively, the above joint contrastive loss encourages the encoded feature vector to approach its assigned classes, clusters or instances. Note that we utilize class centroids $\{\boldsymbol{w}\}$ instead of learnable class weights for encoding source-domain classes to match their semantics to those of the clusters' or outliers' centroids. Our experiments (Section 4.4) show that, if the semantics of class-level, cluster-level and instance-level supervisions do not match, the performance has significant drops.

3.1.2 Hybrid Memory

As the cluster number n_c^t and outlier instance number n_o^t may change during training with the alternate clustering strategy, the class prototypes for the unified contrastive loss (Eq. (1)) are built in a non-parametric and dynamic manner. We propose a novel hybrid memory to provide the source-domain class centroids $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{n^s}\}$, target-domain cluster centroids $\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{n_c^t}\}$ and target-domain un-clustered instance features $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{n_o^t}\}$. For continuously storing and updating the above three types of entries, we propose to cache source-domain *class* centroids $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{n^s}\}$ and all the target-domain *instance* features $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{n^t}\}$ simultaneously in the hybrid memory, where n^t is the number of all the target-domain instances and $n^t \neq n_c^t + n_o^t$. Without loss of generality, we assume that un-clustered features in $\{\boldsymbol{v}\}$ have indices $\{1, \cdots, n_o^t\}$, while other clustered features in $\{\boldsymbol{v}\}$ have indices from $n_o^t + 1$ to n^t. In other words, $\{\boldsymbol{v}_{n_o^t+1}, \cdots, \boldsymbol{v}_{n^t}\}$ dynamically form the cluster centroids $\{\boldsymbol{c}\}$ while $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{n_o^t}\}$ remain un-clustered instances.

Memory initialization. The hybrid memory is initialized with the extracted features by performing forward computation of f_θ: the initial source-domain class centroids $\{\boldsymbol{w}\}$ can be obtained as the mean feature vectors of each class, while the initial target-domain instance features $\{\boldsymbol{v}\}$ are directly encoded by f_θ. After that, the target-domain cluster centroids $\{\boldsymbol{c}\}$ are initialized with the mean feature vectors of each cluster from $\{\boldsymbol{v}\}$, *i.e.*,

$$\boldsymbol{c}_k = \frac{1}{|\mathcal{I}_k|} \sum_{\boldsymbol{v}_i \in \mathcal{I}_k} \boldsymbol{v}_i, \tag{2}$$

where $\mathcal{I}_k$ denotes the k-th cluster set that contains all the feature vectors within cluster k and $|\cdot|$ denotes the number of features in the set. Note that the source-domain class centroids $\{\boldsymbol{w}\}$ and the target-domain instance features $\{\boldsymbol{v}\}$ are only initialized once at the beginning of the whole learning algorithm, while the target-domain cluster centroids $\{\boldsymbol{c}\}$ are re-calculated if any cluster or its instance is updated.

Memory update. At each iteration, the encoded feature vectors in each mini-batch would be involved in hybrid memory updating. For the source-domain class centroids $\{\boldsymbol{w}\}$, the k-th centroid $\boldsymbol{w}_k$ is updated by the mean of the encoded features belonging to class k in the mini-batch as

$$\boldsymbol{w}_k \leftarrow m^s \boldsymbol{w}_k + (1 - m^s) \cdot \frac{1}{|\mathcal{B}_k|} \sum_{\boldsymbol{f}_i^s \in \mathcal{B}_k} \boldsymbol{f}_i^s, \tag{3}$$

where $\mathcal{B}_k$ denotes the feature set belonging to source-domain class k in the current mini-batch and $m^s \in [0, 1]$ is a momentum coefficient for updating source-domain class centroids. m^s is empirically set as 0.2.

The target-domain cluster centroids cannot be stored and updated in the same way as the source-domain class centroids, since the clustered set $\mathbb{X}_c^t$ and un-clustered set $\mathbb{X}_o^t$ are constantly changing. As the hybrid memory caches all the target-domain features $\{\boldsymbol{v}\}$, each encoded feature vector $\boldsymbol{f}_i^t$ in the mini-batch is utilized to update its corresponding instance entry $\boldsymbol{v}_i$ by

$$\boldsymbol{v}_i \leftarrow m^t \boldsymbol{v}_i + (1 - m^t)\boldsymbol{f}_i^t, \tag{4}$$

where $m^t \in [0, 1]$ is the momentum coefficient for update target-domain instance features and is set as 0.2 in our experiments. Given the updated instance memory $\boldsymbol{v}_i$, if $\boldsymbol{f}_i^t$ belongs to the cluster k, the corresponding centroid $\boldsymbol{c}_k$ needs to be updated with Eq. (2).

3.2 Self-paced Learning with Reliable Clusters

A simple way to split the target-domain data into clusters $\mathbb{X}_c^t$ and un-clustered outliers $\mathbb{X}_o^t$ is to cluster the target-domain instance features $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{n^t}\}$ from the hybrid memory by a certain algorithm (*e.g.*, DBSCAN [7]). Since all the target-domain clusters and un-clustered outlier instances are treated as distinct classes in Eq. (1), the clustering reliability would significantly impact the learned representations. If the clustering is perfect, merging all the instances into their true clusters would no doubt improve the final performance (denotes as "oracle" in Table 5). However, in practice, merging an instance into a wrong cluster does more harm than good. A self-paced learning strategy is therefore introduced, where in the re-clustering step before each epoch, only the most reliable clusters are preserved and the unreliable clusters are disassembled back to un-clustered instances. A reliability criterion is proposed to identify unreliable clusters by measuring the independence and compactness.

Independence of clusters. A reliable cluster should be independent from other clusters and individual samples. Intuitively, if a cluster is far away from other samples, it can be considered as highly independent. However, due to the uneven density in the latent space, we cannot naïvely use the distances between the cluster centroid and outside-cluster samples to measure the cluster independence. Generally, the clustering results can be tuned by altering certain hyper-parameters of the clustering criterion. One can *loosen* the clustering criterion to possibly include *more* samples in each cluster or *tighten* the clustering criterion to possibly include *fewer* samples in each cluster. We denote the samples within the same cluster of $\boldsymbol{f}_i^t$ as $\mathcal{I}(\boldsymbol{f}_i^t)$. We propose the following metric to measure the cluster independence, which is formulated as an intersection-over-union (IoU) score,

$$\mathcal{R}_{\text{indep}}(\boldsymbol{f}_i^t) = \frac{|\mathcal{I}(\boldsymbol{f}_i^t) \cap \mathcal{I}_{\text{loose}}(\boldsymbol{f}_i^t)|}{|\mathcal{I}(\boldsymbol{f}_i^t) \cup \mathcal{I}_{\text{loose}}(\boldsymbol{f}_i^t)|} \in [0, 1], \tag{5}$$

where $\mathcal{I}_{\text{loose}}(\boldsymbol{f}_i^t)$ is the cluster set containing $\boldsymbol{f}_i^t$ when the clustering criterion becomes looser. Larger $\mathcal{R}_{\text{indep}}(\boldsymbol{f}_i^t)$ indicates a more independent cluster for $\boldsymbol{f}_i^t$, *i.e.*, even one looses the clustering criterion, there would be no more sample to be included into the new cluster $\mathcal{I}_{\text{loose}}(\boldsymbol{f}_i^t)$. Samples within the same cluster set (*e.g.*, $\mathcal{I}(\boldsymbol{f}_i^t)$) generally have the same independence score.

Compactness of clusters. A reliable cluster should also be compact, *i.e.*, the samples within the same cluster should have small inter-sample distances. In an extreme case, when a cluster is most compact, all the samples in the cluster have zero inter-sample distances. Its samples would not be split into different clusters even when the clustering criterion is tightened. Based on this assumption, we can define the following metric to determine the compactness of the clustered point $\boldsymbol{f}_i^t$ as

$$\mathcal{R}_{\text{comp}}(\boldsymbol{f}_i^t) = \frac{|\mathcal{I}(\boldsymbol{f}_i^t) \cap \mathcal{I}_{\text{tight}}(\boldsymbol{f}_i^t)|}{|\mathcal{I}(\boldsymbol{f}_i^t) \cup \mathcal{I}_{\text{tight}}(\boldsymbol{f}_i^t)|} \in [0, 1], \tag{6}$$

where $\mathcal{I}_{\text{tight}}(\boldsymbol{f}_i^t)$ is the cluster set containing $\boldsymbol{f}_i^t$ when tightening the criterion. Larger $\mathcal{R}_{\text{comp}}(\boldsymbol{f}_i^t)$ indicates smaller inter-sample distances around $\boldsymbol{f}_i^t$ within $\mathcal{I}(\boldsymbol{f}_i^t)$, since a cluster with larger inter-sample distances is more likely to include fewer points when a tightened criterion is adopted. The same cluster's data points may have different compactness scores due to the uneven density.

Given the above metrics for measuring the cluster reliability, we could compute the independence and compactness scores for each data point within clusters. We set up $\alpha, \beta \in [0, 1]$ as independence and compactness thresholds for determining reliable clusters. Specifically, we preserve independent clusters with compact data points whose $\mathcal{R}_{\text{indep}} > \alpha$ and $\mathcal{R}_{\text{comp}} > \beta$, while the remaining data

are treated as un-clustered outlier instances. With the update of the encoder f_θ and target-domain instance features $\{\boldsymbol{v}\}$ from the hybrid memory, more reliable clusters can be gradually created to further improve the feature learning. The overall algorithm is detailed in Appendix A.

4 Experiments

4.1 Datasets and Evaluation Protocol

Table 1: Statistics of the datasets used for training and evaluation. (*) denotes the synthetic datasets.

Dataset	# train IDs	# train images	# test IDs	# query images	# cameras	# total images
Market-1501 [53]	751	12,936	750	3,368	6	32,217
DukeMTMC-reID [35]	702	16,522	702	2,228	8	36,411
MSMT17 [44]	1,041	32,621	3,060	11,659	15	126,441
PersonX [37]*	410	9,840	856	5,136	6	45,792
VeRi-776 [27]	575	37,746	200	1,678	20	51,003
VehicleID [26]	13,164	113,346	800	5,693	-	221,763
VehicleX [29]*	1,362	192,150	-	-	11	192,150

We evaluate our proposed method on both the mainstream real→real adaptation tasks and the more challenging synthetic→real adaptation tasks in person re-ID and vehicle re-ID problems. As shown in Table 1, three real-world person datasets and one synthetic person dataset, as well as two real-world vehicle datasets and one synthetic vehicle dataset, are adopted in our experiments.

Person re-ID datasets. The Market-1501, DukeMTMC-reID and MSMT17 are widely used real-world person image datasets in domain adaptive tasks, among which, MSMT17 has the most images and is most challenging. The synthetic PersonX [37] is generated based on Unity [34] with manually designed obstacles, *e.g.*, random occlusion, resolution and illumination differences, *etc*.

Vehicle re-ID datasets. Although domain adaptive person re-ID has been long studied, the same task on the vehicle has not been fully explored. We conduct experiments with the real-world VeRi-776, VehicleID and the synthetic VehicleX datasets. VehicleX [29] is also generated by the Unity engine [48, 40] and further translated to have the real-world style by SPGAN [6].

Evaluation protocol. In the experiments, only ground-truth IDs on the source-domain datasets are provided for training. Mean average precision (mAP) and cumulative matching characteristic (CMC), proposed in [53], are adopted to evaluate the methods' performances on the target-domain datasets. No post-processing technique, *e.g.*, re-ranking [54] or multi-query fusion [53], is adopted.

4.2 Implementation Details

We adopt an ImageNet-pretrained [5] ResNet-50 [14] as the backbone for the encoder f_θ. Following the clustering-based UDA methods [9, 47, 36], we use DBSCAN [7] for clustering before each epoch. The maximum distance between neighbor points, which is the most important parameter in DBSCAN, is tuned to loosen or tighten the clustering in our proposed self-paced learning strategy. We use a constant threshold α and dynamic threshold β for identifying independent clusters with the most compact points by the reliability criterion. More details can be found in Appendix C.

4.3 Comparison with State-of-the-arts

UDA performance on the target domain. We compare our proposed framework with state-of-the-art UDA methods on multiple domain adaptation tasks in Table 2, including five real→real and three synthetic→real tasks. The tasks in Tables 2 (b) & (c) were not surveyed by previous methods, so we implement state-of-the-art MMT [9] on these datasets for comparison. Our method significantly outperforms all state-of-the-arts on both person and vehicle datasets with a plain ResNet-50 backbone, achieving 2-4% improvements in terms of mAP on the common real→real tasks and up to 7.1% increases on the challenging synthetic→real tasks. An inspiring discovery is that the synthetic→real task could achieve competitive performance as the real→real task with the same target-domain dataset (*e.g.*, DukeMTMC-reID, VeRi-776), which indicates that we are one more step closer towards no longer needing any manually annotated real-world images in the future.

Further improvements on the source domain. State-of-the-art UDA methods inevitably forget the source-domain knowledge after fine-tuning the pretrained networks on the target domain, as demonstrated by MMT [9] in Table 3. In contrast, our proposed unified framework could effectively model complex inter-sample relations across the two domains, boosting the source-domain performance by 6-7% mAP. Our method also outperforms state-of-the-art supervised re-ID methods [28, 38] on the source domain without either using multiple losses or extra training tricks. Such a phenomenon indicates that our method could be applied to improve the supervised training by incorporating unlabeled data without extra human labor.

Table 2: Comparison with state-of-the-art methods on domain adaptive object re-ID. (*) the implementation is based on the authors' code.

(a) *Real→real* unsupervised domain adaptation on person re-ID datasets.

Methods		DukeMTMC-reID→Market-1501				Market-1501→DukeMTMC-reID			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
PUL [8]	TOMM'18	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
TJ-AIDL [43]	CVPR'18	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
SPGAN [6]	CVPR'18	22.8	51.5	70.1	76.8	22.3	41.1	56.6	63.0
HHL [56]	ECCV'18	31.4	62.2	78.8	84.0	27.2	46.9	61.0	66.7
ARN [22]	CVPRW'18	39.4	70.3	80.4	86.3	33.4	60.2	73.9	79.5
ECN [57]	CVPR'19	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
UCDA [33]	ICCV'19	30.9	60.4	-	-	31.0	47.7	-	-
PDA-Net [21]	ICCV'19	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
CR-GAN [3]	ICCV'19	54.0	77.7	89.7	92.7	48.6	68.9	80.2	84.7
PCB-PAST [51]	ICCV'19	54.6	78.4	-	-	54.3	72.4	-	-
SSG [47]	ICCV'19	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
ECN++ [58]	TPAMI'20	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4
MMCL [42]	CVPR'20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
SNR [17]	CVPR'20	61.7	82.8	-	-	58.1	76.3	-	-
AD-Cluster [50]	CVPR'20	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
MMT [9] (*k*-means)	ICLR'20	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5
MMT [9] (DBSCAN)*	ICLR'20	73.8	89.5	96.0	97.6	62.3	76.3	87.7	91.2
Ours		**76.7**	**90.3**	**96.2**	**97.7**	**68.8**	**82.9**	**90.1**	**92.5**

Methods		Market-1501→MSMT17				DukeMTMC-reID→MSMT17			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
PTGAN [44]	CVPR'18	2.9	10.2	-	24.4	3.3	11.8	-	27.4
ECN [57]	CVPR'19	8.5	25.3	36.3	42.1	10.2	30.2	41.5	46.8
SSG [47]	ICCV'19	13.2	31.6	-	49.6	13.3	32.2	-	51.2
ECN++ [58]	TPAMI'20	15.2	40.4	53.1	58.7	16.0	42.5	55.9	61.5
MMCL [42]	CVPR'20	15.1	40.8	51.8	56.7	16.2	43.6	54.3	58.9
MMT [9] (*k*-means)	ICLR'20	22.9	49.2	63.1	68.8	23.3	50.1	63.9	69.8
MMT [9] (DBSCAN)*	ICLR'20	24.0	50.1	63.5	69.3	25.1	52.9	**66.3**	**71.3**
Ours		**25.4**	**51.6**	**64.3**	**69.7**	**26.5**	**53.1**	65.8	70.5

(b) *Synthetic→real* unsupervised domain adaptation on person re-ID datasets.

Methods		PersonX→Market-1501				PersonX→DukeMTMC-reID			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
MMT [9] (DBSCAN)*	ICLR'20	71.0	86.5	94.8	97.0	60.1	74.3	86.5	90.5
Ours		**73.1**	**87.3**	**95.0**	**97.0**	**67.2**	**81.8**	**90.2**	**92.6**

(c) *Real→real* and *synthetic→real* unsupervised domain adaptation on vehicle re-ID datasets.

Methods		VehicleID→VeRi-776				VehicleX→VeRi-776			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
MMT [9] (DBSCAN)*	ICLR'20	35.3	74.6	82.6	87.0	35.6	76.0	83.1	87.4
Ours		**38.4**	**79.9**	**86.2**	**89.3**	**38.3**	**82.1**	**87.8**	**90.2**

Table 3: Comparison with state-of-the-art UDA methods and supervised learning methods when evaluating on the labeled source domain. (*) the implementation is based on the authors' code.

Methods		DukeMTMC-reID→Market-1501				Market-1501→DukeMTMC-reID			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
MMT [9] (*k*-means)	ICLR'20	15.4	28.5	42.9	49.2	26.1	53.4	69.7	75.5
MMT [9] (DBSCAN)*	ICLR'20	18.4	31.6	47.6	53.9	27.5	56.0	72.3	78.0
Encoder train/test on the source domain		70.6	82.5	91.7	94.6	80.5	92.0	97.3	98.4
Ours test on the source domain		**77.9** (+7.3)	**88.2**	**94.4**	**96.0**	**86.9** (+6.4)	**93.8**	**98.0**	**98.7**
Supervised learning methods on source		DukeMTMC-reID				Market-1501			
Bags of tricks [28]	CVPRW'19	76.4	86.4	93.9	96.1	85.9	94.5	98.2	99.0
Circle loss [38]	CVPR'20	-	-	-	-	84.9	94.2	-	-

Unsupervised re-ID without any labeled training data. Another stream of research focuses on training the re-ID model without any labeled data, *i.e.*, excluding source-domain data from the training set. Our method can be easily generalized to such a setting by discarding the source-domain class centroids $\{\boldsymbol{w}\}$ from both the hybrid memory and training objective (Eq. (1)). As shown in Table 4, our method considerably outperforms state-of-the-arts by up to 16.2% improvements in terms of mAP. We also implement state-of-the-art unsupervised method MoCo [13], which adopts the conventional contrastive loss, and unfortunately, it is inapplicable on unsupervised re-ID tasks.

4.4 Ablation Studies

We analyse the effectiveness of our proposed unified contrastive loss with hybrid memory and self-paced learning strategy in Table 5. The "oracle" experiment adopts the target-domain ground-truth IDs as cluster labels for training, reflecting the maximal performance with our pipeline.

Unified contrastive learning mechanism. In order to verify the necessity of each type of classes in the unified contrastive loss (Eq. (1)), we conduct experiments when removing any one of the source-domain class-level, target-domain cluster-level or un-clustered instance-level supervisions. Baseline "*Src.* class" adopts only source-domain images with ground-truth IDs for training. "*Src.* class + *tgt.* instance" treats each target-domain sample as a distinct class. It totally fails with even worse results than the baseline "*Src.* class", showing that directly generalizing conventional contrastive loss to UDA

Table 4: Comparison with state-of-the-art methods on the unsupervised person re-ID task without the labeled source-domain data. (*) the implementation is based on the authors' code.

Methods		Market-1501				DukeMTMC-reID			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
OIM [46]	CVPR'17	14.0	38.0	58.0	66.3	11.3	24.5	38.8	46.0
BUC [24]	AAAI'19	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4
SSL [25]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9
MMCL [42]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0
HCT [18]	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4
MoCo [13]*	CVPR'20	6.1	12.8	27.1	35.7	5.6	10.7	22.0	27.8
Ours w/o source-domain data		**72.6**	**87.7**	**95.2**	**96.9**	**65.3**	**81.2**	**90.3**	**92.2**

Table 5: Ablation studies of our proposed method on individual components.

Methods	DukeMTMC-reID→Market-1501				Market-1501→DukeMTMC-reID			
	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
analysis of the unified contrastive learning mechanism:								
Src. class	19.6	44.4	62.0	69.4	15.7	29.4	44.6	51.8
Src. class + *tgt.* instance	6.0	14.8	27.3	36.0	4.7	10.2	21.1	28.3
Src. class + *tgt.* cluster (w/o self-paced)	23.3	45.0	58.6	64.8	38.8	58.3	69.2	72.8
Src. class + *tgt.* cluster (w/ self-paced)	66.2	83.5	92.2	94.6	60.3	75.9	84.0	86.3
Src. class → *Src.* learnable weights	72.9	88.0	95.1	96.5	64.3	79.8	88.9	91.3
Ours w/o unified contrast	66.6	83.8	93.5	95.5	61.7	76.5	86.4	88.6
analysis of the self-paced learning strategy:								
Ours w/o self-paced $\mathcal{R}_{\text{comp}}$&$\mathcal{R}_{\text{indep}}$	74.5	88.8	95.5	97.1	66.7	80.9	89.7	92.0
Ours w/o self-paced $\mathcal{R}_{\text{comp}}$	75.2	89.8	95.7	97.1	67.5	82.0	90.0	92.3
Ours w/o self-paced $\mathcal{R}_{\text{indep}}$	76.3	89.8	95.9	97.4	68.0	82.5	**90.1**	92.3
Oracle	84.4	93.5	97.4	98.5	74.6	86.7	92.9	95.1
Ours (full)	**76.7**	**90.3**	**96.2**	**97.7**	**68.8**	**82.9**	**90.1**	**92.5**

tasks is inapplicable. "*Src.* class + *tgt.* cluster" follows existing UDA methods [9, 47, 51, 10], by simply discarding un-clustered instances from training. Noticeable performance drops are observed, especially without the self-paced policy to constrain reliable clusters.

We adopt the non-parametric class centroids to supervise the source-domain feature learning, however, conventional methods generally adopt a learnable classifier for supervised learning. "*Src.* class → *Src.* learnable weights" is therefore conducted to verify the necessity of using source-domain class centroids for training to match the semantics of target-domain training supervisions. We also test the effect of not extending negative classes across different types of contrasts. For instance, source-domain samples only treat non-corresponding source-domain classes as their negative classes. "Ours w/o unified contrast" shows inferior performance in Table 5. This indicates the effectiveness of the unified contrastive learning between all types of classes in Eq. (1).

Self-paced learning strategy. We propose the self-paced learning strategy to preserve the most reliable clusters for providing stronger supervisions. $\mathcal{R}_{\text{indep}}$ and $\mathcal{R}_{\text{comp}}$ are proposed to measure the independence and compactness of clusters, respectively. To verify the effectiveness of such a strategy, we evaluate our framework when removing either $\mathcal{R}_{\text{indep}}$ or $\mathcal{R}_{\text{comp}}$, or both of them. Obvious performance drops are observed under all these settings. We illustrate the number of clusters during training in Figure 3. It can be observed that the number of clusters are closer to the ground-truth IDs with the proposed self-paced learning strategy regardless of the un-clustered instance-level contrast, indicating higher reliability of the clusters and the effectiveness of the self-paced strategy.

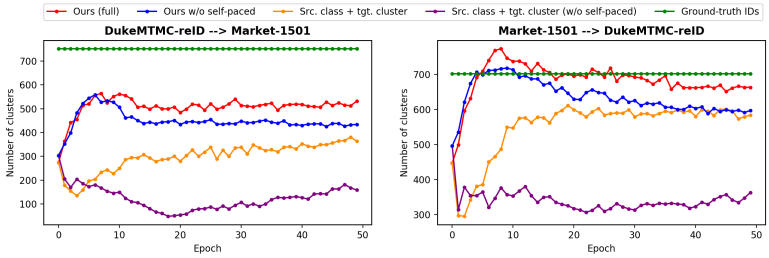

Figure 3: Ablation study by observing the dynamically changing cluster numbers during training.

5 Discussion and Conclusion

Our method has shown considerable improvements over a variety of unsupervised or domain adaptive object re-ID tasks. The supervised performance can also be promoted labor-free by incorporating unlabeled data for training in our framework. The core is at exploiting all available data for jointly training with hybrid supervision. Positive as the results are, there still exists a gap from the oracle, suggesting that the pseudo-class labels may not be satisfactory enough even with the proposed self-paced strategy. Further studies are called for. Beyond the object re-ID task, our method has great potential on other unsupervised learning tasks, which needs to be explored.

Broader Impact

Our method can help to identify and track different types of objects (*e.g.*, vehicles, cyclists, pedestrians, *etc*.) across different cameras (domains), thus boosting the development of smart retail, smart transportation, and smart security systems in the future metropolises. In addition, our proposed self-paced contrastive learning is quite general and not limited to the specific research field of object re-ID. It can be well extended to broader research areas, including unsupervised and semi-supervised representation learning.

However, object re-ID systems, when applied to identify pedestrians and vehicles, might give rise to the infringement of people's privacy. Therefore, governments and officials need to carefully establish strict regulations and laws to control the usage of re-ID technologies. Furthermore, we should be cautious of the misidentification of the re-ID systems to avoid possible disturbance.

References

[1] Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR (June 2019)

[2] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)

[3] Chen, Y., Zhu, X., Gong, S.: Instance-guided context rendering for cross-domain person re-identification. In: ICCV. pp. 232–242 (2019)

[4] Choi, J., Jeong, M., Kim, T., Kim, C.: Pseudo-labeling curriculum for unsupervised domain adaptation. arXiv preprint arXiv:1908.00262 (2019)

[5] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database (2009)

[6] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)

[7] Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. vol. 96, pp. 226–231 (1996)

[8] Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning (2018)

[9] Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: ICLR (2020)

[10] Ge, Y., Zhu, F., Zhao, R., Li, H.: Structured domain adaptation for unsupervised person re-identification. arXiv preprint arXiv:2003.06650 (2020)

[11] Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D.: Curriculumnet: Weakly supervised learning from large-scale web images. In: ECCV. pp. 135–150 (2018)

[12] Guo, X., Liu, X., Zhu, E., Zhu, X., Li, M., Xu, X., Yin, J.: Adaptive self-paced deep clustering with data augmentation. TKDE (2019)

[13] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)

[14] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[15] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. ICLR (2019)

[16] Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. ICML (2018)

[17] Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. CVPR (2020)

[18] Kaiwei Zeng, Munan Ning, Y.W.Y.G.: Hierarchical clustering with hard-batch triplet loss for person re-identification. In: CVPR (2020)

[19] Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: NeurIPS. pp. 1189–1197 (2010)

[20] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)

[21] Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C.F.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. ICCV (2019)

[22] Li, Y.J., Yang, F.E., Liu, Y.C., Yeh, Y.Y., Du, X., Frank Wang, Y.C.: Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In: CVPRW (2018)

[23] Lin, L., Wang, K., Meng, D., Zuo, W., Zhang, L.: Active self-paced learning for cost-effective and progressive face identification. TPAMI **40**(1), 7–19 (2017)

[24] Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI (2019)

[25] Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: CVPR (2020)

[26] Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: CVPR. pp. 2167–2175 (2016)

[27] Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: ECCV. pp. 869–884. Springer (2016)

[28] Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW (2019)

[29] Naphade, M., Wang, S., Anastasiu, D., Tang, Z., Chang, M.C., Yang, X., Zheng, L., Sharma, A., Chellappa, R., Chakraborty, P.: The 4th ai city challenge (2020)

[30] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

[31] Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)

[32] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS, pp. 8026–8037 (2019)

[33] Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., Gao, Y.: A novel unsupervised camera-aware domain adaptation framework for person re-identification. ICCV (2019)

[34] Riccitiello, J.: John riccitiello sets out to identify the engine of growth for unity technologies (interview). VentureBeat. Interview with Dean Takahashi. Retrieved January **18**, 3 (2015)

[35] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCVW (2016)

[36] Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. arXiv preprint arXiv:1807.11334 (2018)

[37] Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: CVPR (2019)

[38] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: CVPR (2020)

[39] Tang, K., Ramanathan, V., Fei-Fei, L., Koller, D.: Shifting weights: Adapting object detectors from image to video. In: NeurIPS. pp. 638–646 (2012)

[40] Tang, Z., Naphade, M., Birchfield, S., Tremblay, J., Hodge, W., Kumar, R., Wang, S., Yang, X.: Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: ICCV. pp. 211–220 (2019)

[41] Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)

[42] Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: CVPR (2020)

[43] Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)

[44] Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018)

[45] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR. pp. 3733–3742 (2018)

[46] Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. pp. 3415–3424 (2017)

[47] Yang, F., Yunchao, W., Guanshuo, W., Yuqian, Z., Honghui, S., Thomas, H.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. ICCV (2019)

[48] Yao, Y., Zheng, L., Yang, X., Naphade, M., Gedeon, T.: Simulating content consistent vehicle datasets with attribute descent. arXiv preprint arXiv:1912.08855 (2019)

[49] Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)

[50] Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y.: Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: CVPR (2020)

[51] Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. ICCV (2019)

[52] Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. pp. 2020–2030 (2017)

[53] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)

[54] Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)

[55] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)

[56] Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero-and homogeneously. In: ECCV (2018)

[57] Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR (2019)

[58] Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Learning to adapt invariance in memory for person re-identification. TPAMI (2020)

[59] Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: ICCV. pp. 6002–6012 (2019)

[60] Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018)

A Algorithm Details

Algorithm 1 Self-paced contrastive learning algorithm on domain adaptive object re-ID

Require: Source-domain labeled data $\mathbb{X}^s$ and target-domain unlabeled data $\mathbb{X}^t$;
Require: Initialize the backbone encoder f_θ with ImageNet-pretrained ResNet-50;
Require: Initialize the hybrid memory with features extracted by f_θ;
Require: Temperature τ for Eq. (1), momentum m^s for Eq. (3), momentum m^t for Eq. (4);
for n in [1, num_epochs] **do**
  Group $\mathbb{X}^t$ into $\mathbb{X}^t_c$ and $\mathbb{X}^t_o$ by clustering $\{\boldsymbol{v}\}$ from the hybrid memory with the independence Eq. (5) and compactness Eq. (6) criterion;
  Initialize the cluster centroids $\{\boldsymbol{c}\}$ with Eq. (2) in the hybrid memory;
  for each mini-batch $\{x_i^s\} \subset \mathbb{X}^s$, $\{x_i^t\} \subset \mathbb{X}^t$ **do**
    1: Encode features $\{\boldsymbol{f}_i^s\}$, $\{\boldsymbol{f}_i^t\}$ for $\{x_i^s\}$, $\{x_i^t\}$ with f_θ;
    2: Compute the unified contrastive loss with $\{\boldsymbol{f}_i^s\}$, $\{\boldsymbol{f}_i^t\}$ by Eq. (1) and update the encoder f_θ by back-propagation;
    3: Update source-domain related class centroids $\{\boldsymbol{w}\}$ in the hybrid memory with $\{\boldsymbol{f}_i^s\}$ and momentum m^s (Eq. (3));
    4: Update target-domain related instance features $\{\boldsymbol{v}\}$ in the hybrid memory with $\{\boldsymbol{f}_i^t\}$ and momentum m^t (Eq. (4));
    5: Update target-domain related cluster centroids $\{\boldsymbol{c}\}$ with updated $\{\boldsymbol{v}\}$ in the hybrid memory (Eq. (2));
  end for
end for

B Discussions

Comparison with conventional contrastive loss. The proposed self-paced contrastive loss has two distinct differences with the conventional contrastive loss [45, 13, 2, 30]. (1) It incorporates the hybrid memory to jointly distinguish source-domain classes, target-domain clusters, and target-domain unclustered instances, while conventional contrastive loss only focuses on separating instances without considering any ground-truth classes or pseudo-class labels as our method does. (2) Our self-paced learning strategy gradually creates more reliable clusters to dynamically update the hybrid memory and refine the learning targets, while existing contrastive losses utilize fixed instance-level classes.

Comparison with ECN [57, 58]. There is an existing work, ECN [57] with its extension version [58], which also adopts a feature memory for the domain adaptive person re-ID task. Comparison results in Table 2 demonstrate the superiority of our proposed method, and there are three main differences between our method and ECN. (1) Our proposed hybrid memory dynamically provides all the source-domain class-level, target-domain cluster-level and un-clustered instance-level supervisory signals, while the memory used in ECN only provides instance-level supervisions on the target domain. (2) Our proposed unified contrastive learning framework exploits all available information across two domains for joint feature learning, while ECN ignores the relations between images from different domains. (3) We propose a self-paced learning strategy to gradually refine the learning targets on both clusters and un-clustered instances, while ECN adopts noisy k-nearest neighbors as learning targets for all the samples without consideration of uneven density in the latent space.

C More Implementation Details

We implement our framework in PyTorch [32] and adopt 4 GTX-1080TI GPUs for training. The domain adaptation task with both source-domain and target-domain data takes $\sim$ 3 hours for training, and the unsupervised learning task with only target-domain data takes $\sim$ 2 hours for training on DukeMTMC-reID, Market-1501 and PersonX datasets. When training on MSMT17, VehicleID, VeRi-776 and VehicleX datasets, time needs to be doubled due to over $2\times$ images in the training set.

C.1 Network Optimization

We adopt an ImageNet [5]-pretrained ResNet-50 [14] up to the global average pooling layer, followed by a 1D BatchNorm layer and an L_2-normalization layer, as the backbone for the encoder f_θ. Domain-specific BNs [1] are used in f_θ for narrowing domain gaps. Adam optimizer is adopted to optimize f_θ with a weight decay of 0.0005. The initial learning rate is set to 0.00035 and is decreased to 1/10 of its previous value every 20 epochs in the total 50 epochs. The temperature τ in Eq. (1) is empirically set as 0.05. The hybrid memory is initialized by extracting the whole training set with the ImageNet-pretrained encoder f_θ, and is then dynamically updated with $m^s = m^t = 0.2$ in Eq. (3)&(4) at each iteration.

C.2 Training Data Organization

During training, each mini-batch contains 64 source-domain images of 16 ground-truth classes (4 images for each class) and 64 target-domain images of *at least* 16 pseudo classes, where target-domain clusters and un-clustered instances are all treated as independent pseudo classes (4 images for each cluster or 1 image for each un-clustered instance). The person images are resized to 256×128 and the vehicle images are resized to 224×224. Random data augmentation is applied to each image before it is fed into the network, including randomly flipping, cropping and erasing [55].

C.3 Target-domain Clustering

Following the clustering-based UDA methods [9, 47, 36], we use DBSCAN [7] and Jaccard distance [54] with k-reciprocal nearest neighbors for clustering before each epoch, where $k = 30$. For DBSCAN, the maximum distance between neighbors is set as $d = 0.6$ and the minimal number of neighbors for a dense point is set as 4. In our proposed self-paced learning strategy described in Section 3.2, we tune the value of d to loosen or tighten the clustering criterion. Specifically, we adopt $d = 0.62$ to form the looser criterion and $d = 0.58$ for the tighter criterion, denoted as $\Delta d = 0.02$. The constant threshold α for identifying independent clusters is defined by the top-90% $\mathcal{R}_{\text{indep}}$ before the first epoch and remains the same for all the training process. The dynamic threshold β for identifying compact clusters is defined by the maximum $\mathcal{R}_{\text{comp}}$ in each cluster on-the-fly, *i.e.*, we preserve the most compact points in each cluster.

D Additional Experimental Results

D.1 Performance with IBN-ResNet [31]

Instance-batch normalization (IBN) [31] has been proved effective in object re-ID methods in either unsupervised [9] or supervised [28] learning tasks. We evaluate our framework with IBN-ResNet as the backbone of the encoder, which is formed by replacing all BN layers in ResNet-50 [14] with IBN

Table 6: Comparison of different backbones in our framework, *i.e.*, ResNet-50 and IBN-ResNet.

Source	Target	Ours w/ ResNet-50				Ours w/ IBN-ResNet			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
DukeMTMC-reID	Market-1501	76.7	90.3	96.2	97.7	**79.2**	**91.5**	**96.9**	**98.0**
Market-1501	DukeMTMC-reID	68.8	82.9	90.1	92.5	**69.9**	**83.4**	**91.0**	**93.1**
DukeMTMC-reID	MSMT17	26.5	53.1	65.8	70.5	**31.8**	**58.9**	**70.4**	**75.2**
Market-1501	MSMT17	25.4	51.6	64.3	69.7	**31.0**	**58.1**	**69.6**	**74.1**
PersonX	Market-1501	73.1	87.3	95.0	97.0	**77.9**	**90.5**	**96.1**	**97.7**
PersonX	DukeMTMC-reID	67.2	81.8	90.2	92.6	**68.8**	**81.9**	**90.6**	**92.7**
VehicleID	VeRi-776	**38.4**	**79.9**	**86.2**	**89.3**	38.0	79.7	85.8	88.4
VehicleX	VeRi-776	**38.3**	**82.1**	**87.8**	**90.2**	37.8	80.7	86.1	89.2
None	Market-1501	72.6	87.7	95.2	96.9	**73.8**	**88.4**	**95.3**	**97.3**
None	DukeMTMC-reID	65.3	81.2	**90.3**	92.2	**66.7**	**82.1**	90.0	**92.4**

layers. As shown in Table 6, the performance can be further improved with IBN-ResNet except for the vehicle datasets.

E Parameter Analysis

We tune the hyper-parameters on the task of Market-1501→DukeMTMC-reID, and the chosen hyper-parameters are directly applied to all the other tasks.

E.1 Temperature τ for Contrastive Loss

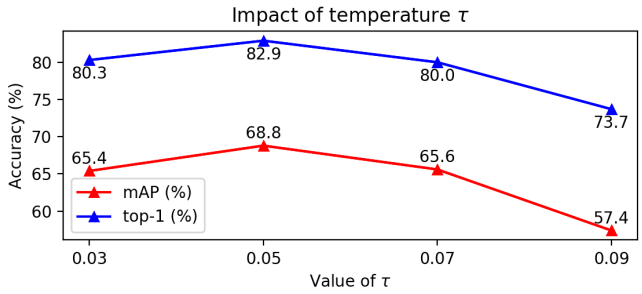

Figure 4: Performance of our framework with different values of temperature τ.

As demonstrated in Figure 4, our framework achieves the optimal performance when setting the temperature τ as 0.05 in Eq. (1) on the task of Market-1501→DukeMTMC-reID.

E.2 Momentum Coefficients m^s, m^t for Hybrid Memory

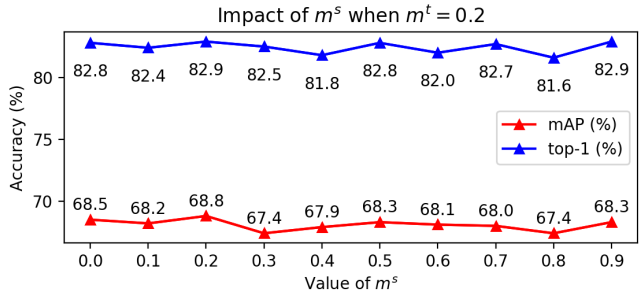

Figure 5: Performance of our framework with different values of m^s when $m^t = 0.2$.

Our proposed hybrid memory simultaneously stores and updates the source-domain class centroids with momentum m^s in Eq. (3) and the target-domain instance features with momentum m^t in Eq. (4). We adopt $m^s = m^t = 0.2$ in our experiments by tuning such hyper-parameter on the task of Market-1501→DukeMTMC-reID.

We find that the value of m^t is critical to the optimal performance (Figure 6) while our framework is not sensitive to the value of m^s (Figure 5), so we adopt the same momentum coefficient on two

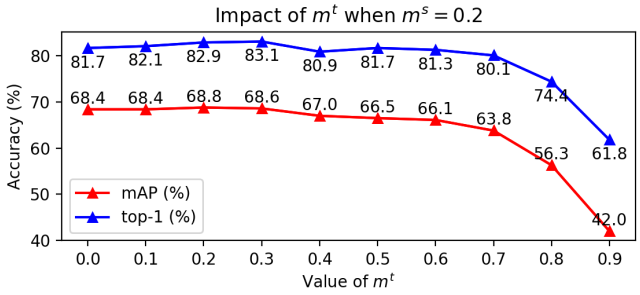

Figure 6: Performance of our framework with different values of m^t when $m^s = 0.2$.

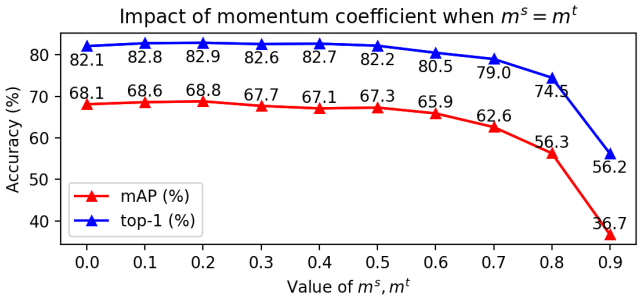

Figure 7: Performance of our framework with different values of m^s, m^t when $m^s = m^t$.

domains for convenience, *i.e.*, $m^s = m^t$. Despite the value of m^t affects the final performance, the results of our framework are robust when m^t changes within a large range, *i.e.*, $[0.0, 0.5]$ in Figure 7.

E.3 Residual Δd for Cluster Reliability Criterion

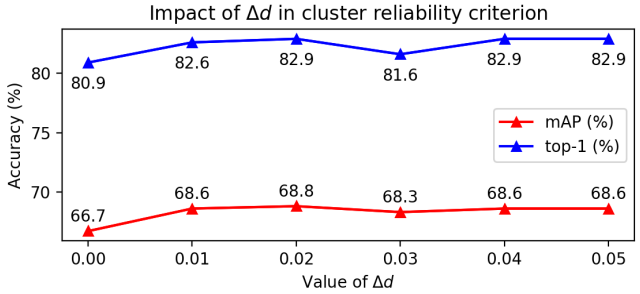

Figure 8: Performance of our framework with different values of Δd in the cluster reliability criterion.

As described in Section C.3, we tune the value of the maximum neighbor distance d with a residual $\Delta d = 0.02$ to measure the cluster reliability in our self-paced learning strategy. As shown in Figure 8, $\Delta d = 0.00$ can be thought of as removing the self-paced strategy from training, which is the same as "Ours w/o self-paced $\mathcal{R}_{\text{comp}}\&\mathcal{R}_{\text{indep}}$" in Table 5. Our method could achieve similar performance when Δd changes within $[0.01, 0.05]$, which indicates that our proposed reliability criterion is not sensitive to the hyper-parameter Δd.