

Refining Pseudo Labels with Clustering Consensus over Generations for Unsupervised Object Re-identification

Xiao Zhang^{1*} Yixiao Ge^{1*} Yu Qiao^{2,3} Hongsheng Li^{1,4}

¹CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong

²SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

³Shanghai AI Laboratory

⁴School of CST, Xidian University

zhangx9411@gmail.com yxge@link.cuhk.edu.hk hsl@ee.cuhk.edu.hk

Abstract

Unsupervised object re-identification targets at learning discriminative representations for object retrieval without any annotations. Clustering-based methods [27, 46, 10] conduct training with the generated pseudo labels and currently dominate this research direction. However, they still suffer from the issue of pseudo label noise. To tackle the challenge, we propose to properly estimate pseudo label similarities between consecutive training generations with clustering consensus and refine pseudo labels with temporally propagated and ensembled pseudo labels. To the best of our knowledge, this is the first attempt to leverage the spirit of temporal ensembling [25] to improve classification with dynamically changing classes over generations. The proposed pseudo label refinery strategy is simple yet effective and can be seamlessly integrated into existing clustering-based unsupervised re-identification methods. With our proposed approach, state-of-the-art method [10] can be further boosted with up to 8.8% mAP improvements on the challenging MSMT17 [39] dataset.

1. Introduction

Recent years witnessed the remarkable progresses of employing unsupervised representation learning in various downstream visual recognition tasks, such as image classification [1, 13, 14, 20], object detection [23, 22, 17, 33], and object re-identification (re-ID) [27, 28, 37, 46, 10]. Object re-ID aims at retrieving objects of interest in large-scale gallery images given an object's query images. The task of unsupervised object re-ID further requires learning discriminative representations to properly model inter/intra-identity affinities without any annotations, which is a more

*The first two authors contribute equally.

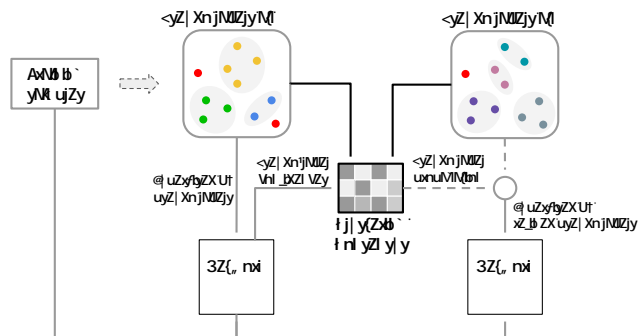


Figure 1: Illustration of the proposed Refining pseudo Label with Clustering Consensus (RLCC) framework. Hard pseudo labels or soft pseudo-label confidences from the previous generation $t - 1$ can be temporally propagated to generation t to effectively refine the pseudo labels at generation t to significantly improve the performance of unsupervised object re-identification.

practical setup in real-world applications.

Pseudo-label-based methods with a clustering-based label generation scheme were found effective in state-of-the-art semi-supervised/unsupervised object re-ID approaches [27, 46, 10, 9, 47, 8, 49]. An iterative and alternative two-stage pipeline is adopted in each training generation (epoch), *i.e.*, creating pseudo labels and training the network with the generated pseudo labels. Although multiple attempts on improving the quality of the pseudo labels have been investigated, the training is still substantially hindered by the inevitable label noise, showing noticeable performance gaps compared to the oracle experiments with ground-truth identities [10]. We argue that properly refining the pseudo labels is at the core of further improving unsupervised re-ID algorithms.

To tackle the challenge, we propose a simple yet effective pseudo label refinery strategy following the similar

spirit of temporal ensembling [25], *i.e.*, the pseudo labels from the past generation (epoch) also carry valuable supervision information and can help mitigate the pseudo label noise by smoothing the pseudo label variations.

The temporal ensembling technique has been widely adopted in semi-supervised learning [25, 35, 9] and self-supervised learning [15, 12] tasks. It aims at generating more robust supervision signals via aggregating models or predictions with a moving average strategy over previous generations (epochs). However, it is non-trivial to improve the pseudo labels in unsupervised object re-ID tasks with off-the-shelf label temporal ensembling methods [25, 35], since they assume that the class definitions of the recognition tasks remain fixed over training generations. In contrast, pseudo labels in different training generations for unsupervised re-ID vary much as the pseudo labels are always updated after each generation.

Towards this end, we introduce **Refining pseudo Labels with Clustering Consensus over consecutive training generations (RLCC)**. Specifically, we estimate the pseudo-label similarities over two consecutive generations with an Intersection over Union (IoU) criterion over the sample-label assignments, where a larger value indicates higher consensus between two pseudo classes in two consecutive generations. To exploit the valuable temporal knowledge encoded by the pseudo labels, we propose to propagate hard or soft pseudo labels from the previous generation to the current generation. The propagation is conducted via a random walk over the pseudo labels, guided by the cross-generation pseudo-label similarities. Given the temporally propagated labels, the noisy pseudo labels at the current generation can be properly refined via a momentum averaging formulation. Our proposed refined pseudo labels can be readily integrated into existing clustering-based unsupervised re-ID approaches [27, 46, 10] with marginal modifications, *i.e.* replacing the conventional hard pseudo labels with the proposed temporally propagated and ensembled soft pseudo labels.

Our contributions can be summarized into three-fold. (1) We introduce to leverage the spirit of temporal ensembling to regularize the noisy pseudo labels in unsupervised object re-ID. Note that existing temporal ensembling techniques [25, 35] are all designed for close-set classification models, which are not applicable in our task. (2) We propose a simple yet effective pseudo label refinery strategy: refining pseudo labels with clustering consensus over training generations (epochs). Our proposed strategy is well compatible with existing pseudo-label-based methods [27, 46, 10] and leads to further improvements on the already high-performance baseline. (3) Our method outperforms state-of-the-arts on multiple benchmarks for unsupervised object re-ID, surpassing state-of-the-art unsupervised method SpCL [10] with up to **8.8%** mAP improvements.

2. Related Works

Unsupervised object re-identification requires to learn discriminative representations for object retrieval without any labeled data. Existing methods [27, 28, 37, 46, 10] mainly focused on training the network with pseudo labels. A mainstream of methods [27, 46, 10] adopted clustering algorithms (*e.g.* DBSCAN [7]) to estimate pseudo labels and were proven effective to achieve satisfactory performance. BUC [27] introduced a bottom-up scheme to gradually incorporate more samples in the clusters for training. HCT [46] encouraged more accurate pseudo labels with a hierarchical clustering algorithm. More recently, SpCL [10] applied a self-paced learning scheme to progressively generate more reliable clusters. Although in different ways, they are all devoted to improving the pseudo label quality, which is shown to be the premise of the success of training.

Unsupervised domain adaptation (UDA) for object re-identification aims to transfer learned knowledge from the labeled source domain to the unlabeled target domain. Existing UDA methods for re-ID can be summarized into two main categories: pseudo-label-based methods [8, 9, 10, 34, 37, 45, 47, 49, 53] and domain translation-based methods [3, 6, 11, 39, 44, 43], pseudo-label-based methods are more effective to capture the target-domain distributions. SSG [8] estimated multi-scale pseudo labels by leveraging human part features. MMT [9] proposed to refine the soft pseudo labels via a mutual learning scheme. AD-Cluster [47] refined the clusters with augmented and generated images. Facing the same problem in unsupervised object re-identification, the major challenge is still on how to provide more reliable pseudo labels and mitigate the noise of the pseudo labels.

Unsupervised representation learning. In real-world applications, it is infeasible to annotate a large amount of training data. Therefore, unsupervised representation learning has been widely studied in many computer vision tasks like image classification [1, 13, 14, 20, 38, 26, 48], image retrieval [21], and object detection [19, 31, 4, 40]. Recently, self-supervised learning methods [2, 15, 18, 30, 36, 41, 54] were in favor for unsupervised pre-training tasks, where a contrastive loss was adopted to learn instance discriminative representations. However, networks trained by these methods need to be fine-tuned with ground-truth labels on down-stream tasks, which are not applicable in our unsupervised re-identification tasks.

Temporal ensembling was first introduced in semi-supervised learning tasks, forming consensus predictions over training generations. Laine *et al.* [25] proposed to use

the temporally ensembled predictions as the training targets for unlabeled samples. Instead of label ensembling, mean-teacher [35] proposed to utilize a temporally ensembled model to predict robust supervision signals. A moving average strategy with momentum was widely used for both model and label aggregating. The idea of temporal ensembling has also been exploited in self-supervised learning [15, 12]. Unfortunately, existing label ensembling techniques cannot be directly employed to improve the pseudo label quality for the unsupervised re-identification tasks, since they focused on problems with fixed class definitions over training generations.

3. Method

State-of-the-art unsupervised object re-ID algorithms [46, 10] are based on pseudo labels. Although a new set of pseudo labels can be generated before each training generation, such pseudo labels have inevitable noise and show large temporal variations over generations, which hinder effective optimization of the re-ID models.

Inspired by temporal ensembling [25], the pseudo labels and pseudo-label confidences of the training samples from the previous generation can still provide valuable supervision information and also help smooth the pseudo label variations over generations. The key innovation of our method lies on effectively propagating pseudo labels and confidences from the previous generation to the current one, refining the noisy pseudo labels. Our proposed label refinery strategy and loss function can be seamlessly integrated into the training of existing approaches without modifying their frameworks or architectures.

3.1. Revisit of Clustering-based Unsupervised Re-identification

Pseudo labels were found effective in unsupervised re-ID tasks to provide plausible supervisions on inter-sample affinities for training, where clustering-based label generation schemes dominated recent methods [27, 46, 10] for achieving state-of-the-art performance. Given N unlabeled data X , a two-stage training scheme is alternately adopted in each training generation: (1) generating pseudo labels Y via clustering the features of the unlabeled training instances, and (2) training the network f with the pseudo cluster labels. Density-based clustering algorithms (e.g., DBSCAN [7]) would result in outliers, which may serve as distinct instance-level classes as suggested in [10]. Note that we denote a training *epoch* as a *generation* in this paper as existing methods iteratively perform the above two-stage scheme with epoch as a cycle.

Naturally, the quality of pseudo labels Y largely impacts the network capability. Although pseudo labels can be gradually improved with the proceed of iteratively re-clustering,

we argue that past pseudo labels carry valuable supervision information but were totally ignored by previous methods.

3.2. Refined Pseudo Labels with Clustering Consensus over Generations

We propose to regularize the noisy pseudo labels $Y^{(t)}$ at the current training generation by exploiting the pseudo labels $Y^{(t-1)}$ from the past generation. There are overall $M^{(t)}$ pseudo classes in the current generation and $M^{(t-1)}$ pseudo classes in the previous generation. In general, $M^{(t)} = M^{(t-1)}$.

Clustering consensus over generations. Since the label sets over training generations do not overlap, we cannot propagate and aggregate the pseudo labels from the previous generation to the current generation with the off-the-shelf temporal ensembling techniques [25].

We therefore propose to first establish the similarities between the pseudo labels of the consecutive two generations, $Y^{(t-1)}$ and $Y^{(t)}$, via clustering consensus. Specifically, we denote the sample set with a pseudo label of i as $I^{(t-1)}(i)$ at the previous generation $t-1$, where $i \in [1, M^{(t-1)}]$. Similarly, samples with a pseudo label of j at the current generation t are denoted as $I^{(t)}(j)$, where $j \in [1, M^{(t)}]$. The clustering consensus matrix $C \in \mathbb{R}^{M^{(t-1)} \times M^{(t)}}$ is therefore adopted to store Intersection over Union (IoU) criterion between pairs of sample sets from two consecutive generations,

$$C(i, j) = \frac{|I^{(t-1)}(i) \cap I^{(t)}(j)|}{|I^{(t-1)}(i) \cup I^{(t)}(j)|} \in [0, 1], \quad (1)$$

where $|\cdot|$ counts the number of samples of a set. Intuitively, $C(i, j)$ measures the consensus or similarity between the pseudo class i in the previous generation $t-1$ and the pseudo class j in the current generation t . After IoU calculation, we normalize each row of the original consensus matrix C to fulfill the constraints that $\sum_j \hat{C}(i, j) = 1$ for all i . The normalization function can be formulated as

$$\hat{C}(i, j) = \frac{C(i, j)}{\sum_{j=1}^{M^{(t)}} C(i, j)}, \quad (2)$$

where the original C will be replaced by \hat{C} after the normalization, i.e. $C \leftarrow \hat{C}$.

Pseudo label propagation. Given the estimated pseudo label similarities between consecutive training generations, the pseudo label information from generation $t-1$ can be propagated to generation t to refine the current pseudo labels. We investigate propagating two types of pseudo label information from generation $t-1$, (1) hard pseudo labels and (2) soft pseudo-label confidences, for refining pseudo labels $Y^{(t)}$ at generation t .

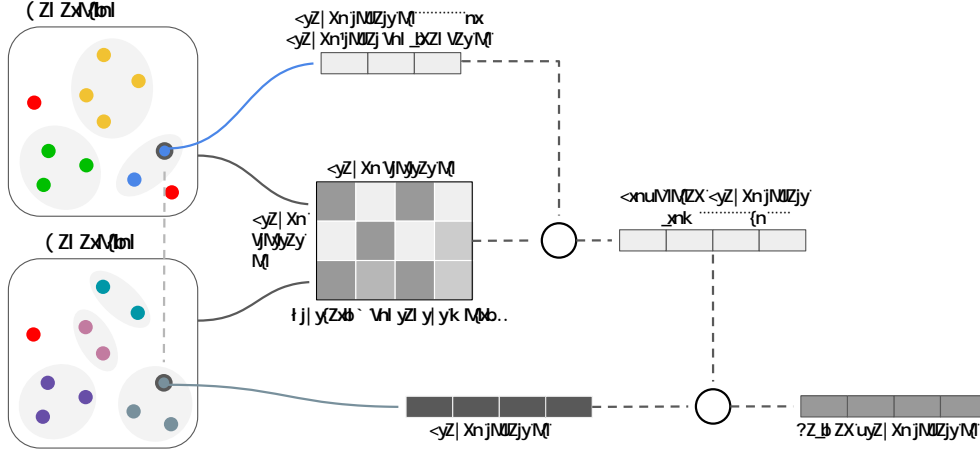


Figure 2: Illustration of the overall framework of our proposed Refining pseudo Labels with Clustering Consensus (RLCC) over training generations. The similarities between pseudo labels at generations $t - 1$ and t are estimated via their clustering consensus. For each sample, either its hard pseudo labels or its pseudo-label confidences at generation $t - 1$ can be propagated to generation t according to the cross-generation pseudo-label similarities. The propagated pseudo labels from generation $t - 1$ can effectively refine the pseudo labels at generation t to boost the performance of unsupervised object re-identification.

(1) *Hard pseudo label propagation.* The hard pseudo labels $\mathbf{Y}^{(t-1)}$ encode much information on inter-sample similarities based on the network trained from former generation $t - 2$. Given an one-hot hard pseudo label $\mathbf{y}_k^{(t-1)} \in \mathbf{Y}^{(t-1)}$ of the k th sample from generation $t - 1$, we propose to propagate its previous label to the current generation as

$$\hat{\mathbf{y}}_k^{(t)} = \mathbf{C} \mathbf{y}_k^{(t-1)}, \quad \text{where } \mathbf{y}_k^{(t-1)} \in \mathbf{R}^{M^{(t-1)}}. \quad (3)$$

The propagated label $\hat{\mathbf{y}}_k^{(t)} \in \mathbf{R}^{M^{(t)}}$ has the same dimension as the number of pseudo classes at generation t . If the “ground-truth” pseudo label for sample k is i , *i.e.* the i th entry of the one-hot vector is 1 as $\mathbf{y}_k^{(t-1)}(i) = 1$. The above equation would yield $\hat{\mathbf{y}}_k^{(t)}(j) = \mathbf{C}(i, j) \mathbf{y}_k^{(t-1)}(i)$. In other words, the propagated sample k ’s pseudo label to the current class j is determined by the cross-generation pseudo-label similarity $\mathbf{C}(i, j)$ between the pseudo class i at generation $t - 1$ and the pseudo class j at generation t . In addition, since each row of the \mathbf{C} matrix sums up to 1, the propagated labels to the current generation would also sum up to 1 to ensure they represent a valid confidence vector for supervision, *i.e.* $\sum_j \hat{\mathbf{y}}_k^{(t)}(j) = 1$. With the proposed propagation scheme, although the two sets of labels have different class definitions (clusters), the pseudo labels can still be successfully propagated across different generations to refine the pseudo labels.

(2) *Soft pseudo-label confidence propagation.* Although the hard pseudo labels carry some useful information about the feature distributions from network, the hard assignments of the samples to the pseudo labels make them less robust against label noise. Existing temporal ensembling methods [25] have shown that the samples’ class confidence vectors

from the previous generations can also act as informative training supervisions or regularization for the later generations. We take advantages of the key insight and investigate propagating soft pseudo-label confidences from the previous generation $t - 1$ to the current generation t . For the k th sample $\mathbf{x}_k \in \mathbf{X}$, given the network from the previous generation $\mathbf{f}^{(t-1)}$, the sample’s classification confidences to pseudo labels at generation $t - 1$ can be obtained as $\mathbf{f}^{(t-1)}(\mathbf{x}_k) \in \mathbf{R}^{M^{(t-1)}}$, where the output dimension of the model matches the number of pseudo labels $M^{(t-1)}$ at generation $t - 1$. Similar to the hard pseudo label propagation, sample k ’s soft pseudo-label confidences at generation $t - 1$ can also be propagated to generation t based on the proposed clustering consensus matrix,

$$\hat{\mathbf{y}}_k^{(t)} = \mathbf{C} \mathbf{f}^{(t-1)}(\mathbf{x}_k), \quad \text{where } \mathbf{f}^{(t-1)}(\mathbf{x}_k) \in \mathbf{R}^{M^{(t-1)}}. \quad (4)$$

Here $\hat{\mathbf{y}}_k^{(t)} \in \mathbf{R}^{M^{(t)}}$ denotes the propagated soft pseudo labels from generations $t - 1$ to t . The intuition of the propagation is similar to that of hard pseudo label propagation. The soft pseudo-label confidences at generation $t - 1$ can be propagated to the current generation t according to the cross-generation pseudo-label similarities by \mathbf{C} . The row-wise normalization property of \mathbf{C} ensures that the summation of the propagated labels is always up to 1. The key difference compared with hard pseudo label propagation is that the model from the previous generation $t - 1$ should be kept and used to generate the soft pseudo label for propagation on-the-fly with some extra computational cost.

Pseudo label refinery at generation t . The propagated pseudo labels $\hat{y}_k^{(t)}$ from generation $t - 1$ can be integrated into the current pseudo labels $y_k^{(t)}$ via the momentum averaging formulation

$$\tilde{y}_k^{(t)} = \alpha \cdot y_k^{(t)} + (1 - \alpha) \cdot \hat{y}_k^{(t)}, \quad (5)$$

where $\alpha \in [0, 1]$ is a momentum coefficient for ensembling. By properly estimating the cross-generation pseudo-label similarities in \mathbf{C} , we can refine the original hard pseudo labels with the propagated pseudo labels or pseudo-label confidences from the past generation, *i.e.* pseudo classes that are consistent over the generations would be more confident. Moreover, the pseudo label variations over consecutive generations can be well smoothed via Eq. (5), leading to more stable training behavior.

Training objective. Our proposed pseudo label refinery strategy is well compatible with existing methods, and can be readily integrated by replacing the hard pseudo labels $y_k^{(t)} \rightarrow \tilde{y}_k^{(t)}$ with the introduced temporally ensembled soft pseudo labels $\tilde{y}_k^{(t)}$ in the training objective, *i.e.*

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^N \text{ce}(\mathbf{f}^{(t)}(\mathbf{x}_k), \tilde{y}_k^{(t)}), \quad (6)$$

where $\text{ce}(\mathbf{p}, \mathbf{q}) = -\mathbf{q} \log \mathbf{p}$ is a cross-entropy loss that uses the refined pseudo labels $\tilde{y}_k^{(t)}$ as the training supervisions.

3.3. Generalization to State-of-the-art SpCL [10] Framework

The state-of-the-art unsupervised re-ID method, SpCL [10], has a major difference from conventional clustering-based methods, *i.e.* no classification head is included in \mathbf{f} and $\mathbf{f}(\mathbf{x}_k) \in \mathbb{R}^L$ indicates the encoded L -dimensional feature for \mathbf{x}_k with normalization. A non-parametric memory module with dynamic class prototypes $\mathbf{W} \in \mathbb{R}^{M \times L}$ is adopted in SpCL to replace the classifier to output the class (or pseudo class) logits.

To generalize and integrate our proposed approach into the SpCL framework, we need to first compute the clustering consensus matrix \mathbf{C} by Eq. (1). The hard pseudo labels or the soft pseudo-label confidences can be propagated from the previous generation $t - 1$ to the current generation t guided by \mathbf{C} . Specifically, the propagation of hard pseudo labels can be computed by Eq. (3). To propagate soft confidences, as the network \mathbf{f} cannot directly output class confidences, we estimate the confidences with the class prototypes in the memory as $\mathbf{W}^{(t-1)} \mathbf{f}^{(t-1)}(\mathbf{x}_k)$. Eq. (4) then becomes

$$\hat{y}_k^{(t)} = \mathbf{C} \cdot \mathbf{W}^{(t-1)} \mathbf{f}^{(t-1)}(\mathbf{x}_k)_{\text{softmax}}, \quad (7)$$

where $\mathbf{W}^{(t-1)} \in \mathbb{R}^{M^{(t-1)} \times L}$ denotes the normalized class prototypes at the generation $t - 1$ and $\mathbf{W}^{(t-1)} \mathbf{f}^{(t-1)}(\mathbf{x}_k) \in \mathbb{R}^{M^{(t-1)}}$ is a temperature hyper-parameter for sharpening the class confidences. Given the propagated label $\hat{y}_k^{(t)}$, we can refine the noisy pseudo label $y_k^{(t)}$ via Eq. (5). SpCL adopts a unified contrastive loss as the training objective, which can be treated as a variant of the cross-entropy loss. To integrate the temporally ensembled soft pseudo labels $\tilde{y}_k^{(t)}$ into the unified contrastive loss,

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^N \text{ce}(\mathbf{W}^{(t)} \mathbf{f}^{(t)}(\mathbf{x}_k), \tilde{y}_k^{(t)}), \quad (8)$$

where $\mathbf{W}^{(t)} \in \mathbb{R}^{M^{(t)} \times L}$. Although SpCL itself aims at improving pseudo label quality by robustly identifying outliers, our proposed pseudo label refinery strategy with temporal ensembling is well complementary with it, further improving the already strong SpCL (see Sec. 4.4).

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our proposed pseudo label refinery strategy on three widely-used person re-ID datasets and a vehicle re-ID dataset. **Market-1501** [50] contains 751 identities for training and 750 identities for testing, captured by 6 cameras. There are 12,936 training images, 19,732 gallery images, and 3,368 query images. **DukeMTMC-reID** [32] contains 16,522 images of 702 identities for training, and the remaining images out of another 702 identities for testing. All images are collected from 8 cameras. **MSMT17** [39] is a newly released person re-ID dataset with the most images. It is composed of 126,411 person images from 4,101 identities collected by 15 cameras. **VeRi-776** [29] collects vehicle images in the real-world urban surveillance scenario. The training set has 575 vehicles with 37,746 images and the testing set has 200 vehicles with 11,579 images, captured by 20 cameras.

Evaluation metrics. In all the experiments, no ground-truth identities are provided for training. Mean average precision (mAP) and cumulative matching characteristic (CMC) [50] are adopted to evaluate the methods. No post-processing technique (*e.g.* re-ranking [51], multi-query fusion [50]), is adopted for inference.

4.2. Implementation Details

Our proposed pseudo label refinery strategy can be readily integrated into existing clustering-based unsupervised re-ID methods. To fully verify the effectiveness of our refined pseudo labels, we implement all the experiments based on the state-of-the-art unsupervised re-ID framework

Propagation		Market-1501 [50]			
		mAP	top-1	top-5	top-10
Hard Pseudo Label	1.00	74.1	88.9	95.2	96.8
	0.95	75.4	89.5	95.6	97.2
	0.90	75.2	89.4	95.2	96.9
	0.85	76.3	90.1	95.8	97.2
	0.80	75.1	89.0	95.3	96.9
Soft Pseudo-label Confidence	1.00	74.1	88.9	95.2	96.8
	0.95	77.6	90.9	95.8	97.0
	0.90	77.7	90.8	96.3	97.5
	0.85	76.7	90.1	96.3	97.0
	0.80	75.1	89.9	95.3	96.7

Table 1: Comparison between the hard pseudo label propagation and the soft pseudo-label confidence propagation in our RLCC strategies. The momentum coefficient varies in $[0.8, 1.0]$ for label ensembling in Eq. (5).

SpCL [10], which is treated as our baseline. DBSCAN [7] clustering followed by a self-paced strategy is utilized for generating hard pseudo labels after each epoch. The further improvements based on our approach on the already very strong baseline are convincing to show the superiority of our method. We adopt the same settings as used in [10] except for the training objective as described in Sec. 3.3, *i.e.*, we use our refined soft pseudo labels to replace the noisy hard pseudo labels in its unified contrastive loss. Two hyper-parameters are required in our proposed RLCC strategy, *i.e.* in Eq. (5) and in Eq. (7). Optimal performances are achieved when adopting the soft pseudo-label confidences propagation with $\alpha = 0.9$ and $\beta = 30$.

The person images are resized to 256×128 and the vehicle images are resized to 224×224 . Several data augmentation techniques, such as randomly flipping, erasing [52], and cropping, are applied to the training samples. Each mini-batch consists of 64 images belonging to 16 pseudo classes are sampled. An ImageNet [5] pre-trained ResNet-50 [16] is adopted as the backbone for f . During training, Adam [24] is adopted to optimize the backbone with a weight decay of 0.0005. The initial learning rate is set to 3.5×10^{-4} and is decreased to 0.1 of its previous value every 20 epochs in the total 50 epochs.

4.3. Ablation Studies

In this section, we investigate different designs of our proposed temporally pseudo label refinery, as well as the factors of the hyper-parameters on the Market-1501 dataset [50]. In each experiment, all settings are kept the same except for the mentioned one. Unless otherwise specified, all the experiments in the remaining parts of this section set $\beta = 30$ and $\alpha = 0.9$, respectively.

4.3.1 Hard v.s. Soft Pseudo Label Propagation

As introduced in Sec. 3.2, we provide two pseudo label propagation strategies, *i.e.* propagating hard pseudo labels via Eq. (3) and propagating soft pseudo-label confidences via Eq. (4). We evaluate both of the two propagation strategies, as illustrated in Table 1. Since different propagation strategies may have different optimal momentum coefficients for ensembling (Eq. (5)), we conduct experiments with the momentum varying from 0.80 to 1.00. The results reveal that soft pseudo-label confidences propagation surpasses hard pseudo label propagation consistently under all the hyper-parameter settings, indicating the superiority of propagating soft pseudo-label confidences in the proposed temporally pseudo label ensembling.

To better analyze the correlations between the hard and soft propagation, we define a general formulation that can reflect the intermediate status between two designs,

$$\hat{\mathbf{y}}_k^{(t)} = \alpha \mathbf{y}_k^{(t-1)} + (1 - \alpha) \mathbf{f}^{(t-1)}(\mathbf{x}_k), \quad (9)$$

where aggregating hyper-parameter α varies from 0 to 1. Eq. (9) is equivalent to the soft propagation Eq. (4) when $\alpha = 0$, and is equivalent to the hard propagation Eq. (3) when $\alpha = 1$. Note that actually, we use Eq. (7) as the soft propagation to align with the baseline framework of SpCL [10], but for simplicity, we write in the form of Eq. (4) without loss of generality. By using Eq. (9) for pseudo label propagation and manipulating the hyper-parameter $\alpha \in [0, 1]$, we achieve the results as shown in Fig. 3. We observe that the optimal performance is obtained when $\alpha = 0$, *i.e.* only soft pseudo-label confidences are used for label propagation. The performance generally decreases as α increases. This phenomenon implies that soft pseudo-label confidences indeed encode more informative supervisions than the hard pseudo labels.

4.3.2 Class Prototypes for Confidence Estimation

As introduced in Sec. 3.3, we use the class prototypes obtained from the non-parametric memory module of SpCL [10] to estimate the soft pseudo-class confidences via Eq. (7). Specifically, the class prototypes $\mathbf{W}^{(t-1)}$ from the previous generation $t - 1$ is adopted in the estimation of class confidences. We consider two options of $\mathbf{W}^{(t-1)}$, one can be cached at the beginning of generation $t - 1$ and the other can be computed at the end of generation $t - 1$. Intuitively, the former one carries the most information from the network of generation $t - 2$, showing larger temporal variations than the later one, whose supervision information has been mostly tuned at generation $t - 1$. The comparison results can be found in Table 2. We observe that caching the class prototypes $\mathbf{W}^{(t-1)}$ at the beginning of generation $t - 1$ and then using the kept $\mathbf{W}^{(t-1)}$ to estimate the pseudo-class confidences for propagation could better refine the pseudo labels,

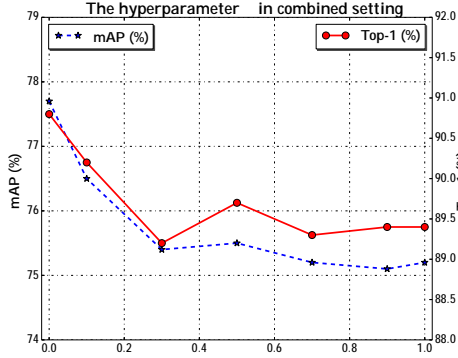


Figure 3: Performance on Market-1501 [50] when linearly combining the hard pseudo labels and soft pseudo-label confidences for label propagation, as specified in Eq. (9). α is used for adjusting the ratio between hard and soft propagation. Results show that soft pseudo-label confidences are more informative than hard pseudo labels.

$W^{(t-1)}$		Market-1501 [50]			
		mAP	top-1	top-5	top-10
Beginning of the generation	0.95	77.6	90.9	95.8	97.0
	0.90	77.7	90.8	96.3	97.5
	0.85	76.7	90.1	96.3	97.0
End of the generation	0.95	75.2	89.6	96.3	97.5
	0.90	75.6	89.2	96.2	97.6
	0.85	74.4	89.0	95.7	97.3

Table 2: Comparison of class prototypes cached at different training stages. The results exhibit that class prototypes cached at the beginning of generation $t - 1$ lead to better pseudo-class confidence estimation for the current generation.

leading to better performance. The phenomenon indicates that properly propagating and ensembling the supervision information is important to refine the pseudo label quality.

4.3.3 Hyper-parameter Analysis

As mentioned in Sec. 4.2, we have two hyper-parameters in our proposed pseudo label refinery strategy.

Momentum in Eq. (5). As illustrated in Fig. 4, we analyze the effects of the momentum $\alpha \in [0, 1]$ in the ensembling equation. Note the pseudo labels remain the same when $\alpha = 1$ and the pseudo labels are totally replaced with the propagated soft labels when $\alpha = 0$. Intuitively, the pseudo labels will be smoother with a smaller α . However, a too smooth label would raise the issue of information erasing, *e.g.* pseudo labels with uniform distributions contain no information at all. From the experiment results, we find that

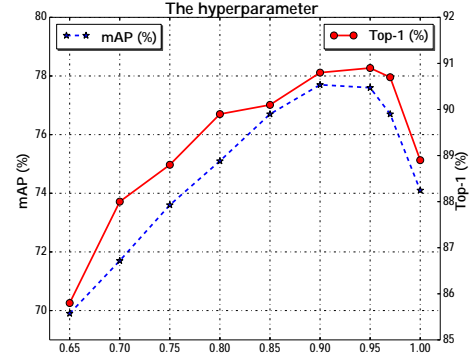


Figure 4: Performances on Market-1501 [50] with different values of the momentum α in Eq. (5) when fixing $\beta = 30$. The model achieves the best mAP=77.7% when $\alpha = 0.9$.

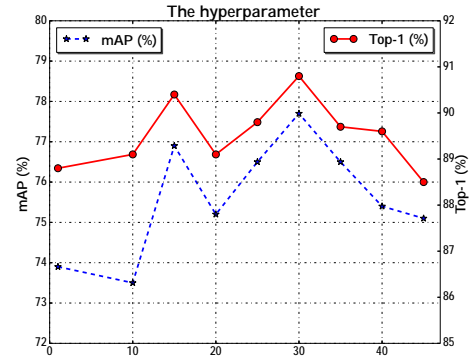


Figure 5: Performances on the unsupervised Market-1501 [50] with different temperature β values in Eq. (7) when fixing $\alpha = 0.9$. The model achieves the best mAP=77.7% when $\beta = 30$.

an optimal performance can be achieved when α is around 0.9. And the results is robust when α varies from $[0.8, 1.0)$, *i.e.* achieving performance gains over the baseline model ($\alpha = 1$).

Temperature in Eq. (7). The distribution of original class confidences $W^{(t-1)} f^{(t-1)}(x_k)$ estimated by the encoded features and class prototypes tend to be smooth and uniform especially when a large number of pseudo labels exist, since the similarity between the encoded feature and most class prototypes may always be small. We introduce to adopt a temperature hyper-parameter β to sharpen the confidence distribution, leading to more informative soft pseudo-class confidences. As shown in Fig. 5, the optimal performance is obtained when $\beta = 30$. The performances are robust when β varies from 15 to 45, showing that our method is not sensitive to either hyper-parameter.

Methods		Market-1501 [50]			
		mAP	top-1	top-5	top-10
OIM [42]	CVPR'17	14.0	38.0	58.0	66.3
BUC [27]	AAAI'19	38.3	66.2	79.6	84.5
SSL [28]	CVPR'20	37.8	71.7	83.8	87.4
MMCL [37]	CVPR'20	45.5	80.3	89.4	92.3
HCT [46]	CVPR'20	56.4	80.0	91.6	95.2
MMT [9]	ICLR'20	67.1	84.6	94.3	96.5
SpCL [10]	NeurIPS'20	73.1	88.1	95.1	97.0
RLCC		77.7	90.8	96.3	97.5

Table 3: Comparison with state-of-the-art unsupervised person re-ID methods on the Market-1501 dataset [50].

Methods		DukeMTMC-reID [32]			
		mAP	top-1	top-5	top-10
OIM [42]	CVPR'17	14.0	38.0	58.0	66.3
BUC [27]	AAAI'19	27.5	47.4	62.6	68.4
SSL [28]	CVPR'20	28.6	47.4	62.6	68.4
MMCL [37]	CVPR'20	40.2	65.2	75.9	80.0
HCT [46]	CVPR'20	50.7	69.6	83.4	87.4
MMT+ [9]	ICLR'20	60.3	75.6	86.0	89.2
SpCL [10]	NeurIPS'20	65.3	81.2	90.3	92.2
RLCC		69.2	83.2	91.6	93.8

Table 4: Comparison with state-of-the-art unsupervised person re-ID methods on the DukeMTMC-reID dataset [32].

Methods		MSMT17 [39]			
		mAP	top-1	top-5	top-10
MoCo [15]	CVPR'20	1.6	4.3	9.7	13.5
MMCL [37]	CVPR'20	11.2	35.4	44.8	49.8
SpCL [10]	NeurIPS'20	19.1	42.3	55.6	61.2
RLCC		27.9	56.5	68.4	73.1

Table 5: Comparison with state-of-the-art unsupervised person re-ID methods on the MSMT17 dataset [39].

4.4. Comparison with State-of-the-arts

We compare our proposed RLCC against state-of-the-art unsupervised re-ID methods on the aforementioned Market-1501 [50], DukeMTMC-reID [32], MSMT17 [39], and VeRi [29] datasets. The results exhibit that RLCC significantly surpasses all the listed state-of-the-arts in these re-ID evaluation datasets.

Unsupervised person re-identification. The comparisons with the state-of-the-art algorithms on person re-ID datasets including Market-1501 [50], DukeMTMC-reID [32] and MSMT17 [39] are shown in Tables 3-5, respectively. On Market-1501 [50], we obtain the best performance among all the compared methods with 77.7% mAP. Compared to the recent state-of-the-art unsupervised method SpCL [10], which is also the baseline model of

Methods		VeRi-776 [29]			
		mAP	top-1	top-5	top-10
MoCo [15]	CVPR'20	9.5	24.9	40.6	51.8
SpCL [10]	NeurIPS'20	36.9	79.9	86.8	89.9
RLCC		39.6	83.4	88.8	90.9

Table 6: Comparison with state-of-the-art unsupervised vehicle re-ID methods on the VeRi-776 dataset [39].

our method, we achieve a noticeable 4.6% mAP improvement. On DukeMTMC-reID [32], RLCC achieves an obvious 3.9% mAP improvement compared to SpCL [10]. For the most challenging MSMT17 [39] benchmark, RLCC achieves an impressive 27.9% mAP, which considerably outperforms state-of-the-art SpCL [10] by an 8.8% mAP improvement.

Unsupervised vehicle re-identification. We also evaluate our proposed RLCC approach on a popular vehicle re-ID benchmark VeRi-776 [29]. The comparisons with state-of-the-art unsupervised algorithms on VeRi-776 [29] are shown in Table 6. Compared to SpCL [10], we achieve a 2.7% mAP improvement.

The stable performance gains by our RLCC over SpCL [10] reveals that our method can consistently improve the baseline model on various object re-ID task under the unsupervised setting.

5. Conclusions

Pseudo label noise is one of the most significant factors that hinder the further improvements of clustering-based unsupervised object re-ID methods [27, 46, 10]. To solve this issue, we introduce to refine noisy pseudo label with temporally propagated and aggregated soft labels, which can be readily integrated into existing methods with marginal modifications. As the label sets vary in different training generations, we propose to estimate clustering consensus to encourage label propagation via a random walk over consecutive generations. Our success suggests that temporal ensembling with the proposed pseudo-label confidence propagation can effectively mitigate pseudo label noise to achieve higher performance. Further studies on more properly leveraging the temporal knowledge over more generations are called for.

Acknowledgements. This work is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14208417, 14207319, 14202217, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18, in part by CUHK Strategic Fund, in part by the Joint Lab of CAS-HK, and in part by the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800).

References

- [1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Eur. Conf. Comput. Vis.*, 2018. 1, 2
- [2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020. 2
- [3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *Int. Conf. Comput. Vis.*, pages 232–242, 2019. 2
- [4] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2
- [5] Jia Deng, Wei Dong, Richard Socher, Li-JiWa Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 6
- [6] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 994–1003, 2018. 2
- [7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, page 226–231. AAAI Press, 1996. 2, 3, 6
- [8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Int. Conf. Comput. Vis.*, pages 6112–6121, 2019. 1, 2
- [9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *Int. Conf. Learn. Represent.*, pages 1–15, 2020. 1, 2, 8
- [10] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2, 3, 5, 6, 8
- [11] Yixiao Ge, Feng Zhu, Rui Zhao, and Hongsheng Li. Structured domain adaptation with online relation regularization for unsupervised person re-id, 2020. 2
- [12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Adv. Neural Inform. Process. Syst.*, volume 33, 2020. 2, 3
- [13] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Int. Conf. Learn. Represent.*, 2020. 1, 2
- [14] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Int. Conf. Comput. Vis.*, pages 8401–8409, 2019. 1, 2
- [15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020. 2, 3, 8
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 6
- [17] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Int. Conf. Comput. Vis.*, pages 6668–6677, 2019. 1
- [18] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Int. Conf. Learn. Represent.*, pages 1–15, 2019. 2
- [19] Fa-Ting Hong, Wei-Hong Li, and Wei-Shi Zheng. Learning to detect important people in unlabelled images for semi-supervised important people detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2
- [20] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *Int. Conf. Learn. Represent.*, 2019. 1, 2
- [21] Young Kyun Jang and Nam Ik Cho. Generalized product quantization network for semi-supervised image retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2
- [22] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Int. Conf. Comput. Vis.*, pages 480–490, 2019. 1
- [23] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Chang-ick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Int. Conf. Comput. Vis.*, pages 6092–6101, 2019. 1
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [25] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Int. Conf. Learn. Represent.*, pages 1–13, 2017. 1, 2, 3, 4
- [26] Wanyu Lin, Zhaolin Gao, and Baochun Li. Shoestring: Graph-based semi-supervised classification with severely limited labeled data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2
- [27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, volume 33, pages 8738–8745, 2019. 1, 2, 3, 8
- [28] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3390–3399, 2020. 1, 2, 8

- [29] Xincheng Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Eur. Conf. Comput. Vis.*, pages 869–884. Springer, 2016. **5, 8**
- [30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *Adv. Neural Inform. Process. Syst.*, 2018. **2**
- [31] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. **2**
- [32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis.*, pages 17–35, 2016. **5, 8**
- [33] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 780–790, 2019. **1**
- [34] 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**
- [35] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inform. Process. Syst.*, pages 1195–1204, 2017. **2, 3**
- [36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Eur. Conf. Comput. Vis.*, pages 1–14, 2020. **2**
- [37] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10981–10990, 2020. **1, 2, 8**
- [38] Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, and Zezhou Li. Weakly supervised fine-grained image classification via gaussian mixture model oriented discriminative learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. **2**
- [39] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 79–88, 2018. **1, 2, 5, 8**
- [40] Zhonghua Wu, Qingyi Tao, Guosheng Lin, and Jianfei Cai. Exploring bottom-up and top-down cues with attentive learning for weakly supervised object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. **2**
- [41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3733–3742, 2018. **2**
- [42] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3415–3424, 2017. **8**
- [43] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1435, 2019. **2**
- [44] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12613–12620, 2020. **2**
- [45] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2148–2157, 2019. **2**
- [46] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13657–13665, 2020. **1, 2, 3, 8**
- [47] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9021–9030, 2020. **1, 2**
- [48] Manyuan Zhang, Guanglu Song, Hang Zhou, and Yu Liu. Discriminability distillation in group representation learning. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020. **2**
- [49] Xinyu Zhang, Jiawei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Int. Conf. Comput. Vis.*, pages 8222–8231, 2019. **1, 2**
- [50] Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, pages 1116–1124, 2015. **5, 6, 7, 8**
- [51] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1318–1327, 2017. **5**
- [52] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. **6**
- [53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 598–607, 2019. **2**
- [54] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Int. Conf. Comput. Vis.*, pages 6002–6012, 2019. **2**