ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification

Hao Chen, Benoit Lagadec, Francois F Bremond

ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification

Hao Chen[1,2,3] Benoit Lagadec[3] Francois Bremond[1,2]
[1]Inria [2]Université Côte d'Azur [3]European Systems Integration
{hao.chen, francois.bremond}@inria.fr benoit.lagadec@esifrance.net

Abstract

Unsupervised person re-identification (ReID) aims at learning discriminative identity features without annotations. Recently, self-supervised contrastive learning has gained increasing attention for its effectiveness in unsupervised representation learning. The main idea of instance contrastive learning is to match a same instance in different augmented views. However, the relationship between different instances has not been fully explored in previous contrastive methods, especially for instance-level contrastive loss. To address this issue, we propose Inter-instance Contrastive Encoding (ICE) that leverages inter-instance pairwise similarity scores to boost previous class-level contrastive ReID methods. We first use pairwise similarity ranking as one-hot hard pseudo labels for hard instance contrast, which aims at reducing intra-class variance. Then, we use similarity scores as soft pseudo labels to enhance the consistency between augmented and original views, which makes our model more robust to augmentation perturbations. Experiments on several large-scale person ReID datasets validate the effectiveness of our proposed unsupervised method ICE, which is competitive with even supervised methods. Code is made available at https://github.com/chenhao2345/ICE.

1. Introduction

Person re-identification (ReID) targets at retrieving an person of interest across non-overlapping cameras by comparing the similarity of appearance representations. Supervised ReID methods [28, 2, 22] use human-annotated labels to build discriminative appearance representations which are robust to pose, camera property and view-point variation. However, annotating cross-camera identity labels is a cumbersome task, which makes supervised methods less scalable in real-world deployments. Unsupervised methods [20, 21, 32] directly train a model on unlabeled data and thus have a better scalability.

Most of previous unsupervised ReID methods [27, 11, 41] are based on unsupervised domain adaptation (UDA). UDA methods adjust a model from a labeled source domain to an unlabeled target domain. The source domain provides a good starting point that facilitates target domain adaptation. With the help of a large-scale source dataset, state-of-the-art UDA methods [11, 41] significantly enhance the performance of unsupervised ReID. However, the performance of UDA methods is strongly influenced by source dataset's scale and quality. Moreover, a large-scale labeled dataset is not always available in the real world. In this case, fully unsupervised methods [20, 21] own more flexibility, as they do not require any identity annotation and directly learn from unlabeled data in a target domain.

Recently, contrastive learning has shown excellent performance in unsupervised representation learning. State-of-the-art contrastive methods [38, 5, 14] consider each image instance as a class and learns representations by matching augmented views of a same instance. As a class is usually composed of multiple positive instances, it hurts the performance of fine-grained ReID tasks when different images of a same identity are considered as different classes. Self-paced Contrastive Learning (SpCL) [13] alleviates this problem by matching an instance with the centroid of the multiple positives, where each positive converges to its centroid at a uniform pace. Although SpCL has achieved impressive performance, this method does not consider inter-instance affinities, which can be leveraged to reduce intra-class variance and make clusters more compact. In supervised ReID, state-of-the-art methods [2, 22] usually adopt a hard triplet loss [16] to lay more emphasis on hard samples inside a class, so that hard samples can get closer to normal samples. In this paper, we introduce Inter-instance Contrastive Encoding (ICE), in which we match an instance with its hardest positive in a mini-batch to make clusters more compact and improve pseudo label quality. Matching the hardest positive refers to using one-hot "hard" pseudo labels.

Since no ground truth is available, mining hardest positives within clusters is likely to introduce false positives into the training process. In addition, the one-hot label does not take the complex inter-instance relationship into consideration when multiple pseudo positives and negatives exist

in a mini-batch. Contrastive methods usually use data augmentation to mimic real-world distortions, *e.g*., occlusion, view-point and resolution variance. After data augmentation operations, certain pseudo positives may become less similar to an anchor, while certain pseudo negatives may become more similar. As a robust model should be invariant to distortions from data augmentation, we propose to use the inter-instance pairwise similarity as "soft" pseudo labels to enhance the consistency before and after augmentation.

Our proposed ICE incorporates class-level label (centroid contrast), instance pairwise hard label (hardest positive contrast) and instance pairwise soft label (augmentation consistency) into one fully unsupervised person ReID framework. Without any identity annotation, ICE significantly outperforms state-of-the-art UDA and fully unsupervised methods on main-stream person ReID datasets.

To summarize, our contributions are: (1) We propose to use pairwise similarity ranking to mine hardest samples as one-hot hard pseudo labels for hard instance contrast, which reduces intra-class variance. (2) We propose to use pairwise similarity scores as soft pseudo labels to enhance the consistency between augmented and original instances, which alleviates label noise and makes our model more robust to augmentation perturbation. (3) Extensive experiments highlight the importance of inter-instance pairwise similarity in contrastive learning. Our proposed method ICE outperforms state-of-the-art methods by a considerable margin, significantly pushing unsupervised ReID to real-world deployment.

2. Related Work

Unsupervised person ReID. Recent unsupervised person ReID methods can be roughly categorized into unsupervised domain adaptation (UDA) and fully unsupervised methods. Among UDA-based methods, several works [33, 19] leverage semantic attributes to reduce the domain gap between source and target domains. Several works [37, 48, 8, 49, 51, 4] use generative networks to transfer labeled source domain images into the style of target domain. Another possibility is to assign pseudo labels to unlabeled images, where pseudo labels are obtained from clustering [27, 10, 42, 3] or reference data [39]. Pseudo label noise can be reduced by selecting credible samples [1] or using a teacher network to assign soft labels [11]. All these UDA-based methods require a labeled source dataset. Fully unsupervised methods have a better flexibility for deployment. BUC [20] first treats each image as a cluster and progressively merge clusters. Lin *et al*. [21] replace clustering-based pseudo labels with similarity-based softened labels. Hierarchical Clustering is proposed in [40] to improve the quality of pseudo labels. Since each identity usually has multiple positive instances, MMCL [32] introduces a memory-based multi-label classification loss into unsupervised ReID. JVTC [18] and CycAs [35] explore temporal information to refine visual similarity. SpCL [13] considers each cluster and outlier as a single class and then conduct instance-to-centroid contrastive learning. CAP [34] calculates identity centroids for each camera and conducts intra- and inter-camera centroid contrastive learning. Both SpCL and CAP focus on instance-to-centroid contrast, but neglect inter-instance affinities.

Contrastive Learning. Recent contrastive learning methods [38, 14, 5] consider unsupervised representation learning as a dictionary look-up problem. Wu *et al*. [38] retrieve a target representation from a memory bank that stores representations of all the images in a dataset. MoCo [14] introduces a momentum encoder and a queue-like memory bank to dynamically update negatives for contrastive learning. In SimCLR [5], authors directly retrieve representations within a large batch. However, all these methods consider different instances of a same class as different classes, which is not suitable in a fine-grained ReID task. These methods learn invariance from augmented views, which can be regarded as a form of consistency regularization.

Consistency regularization. Consistency regularization refers to an assumption that model predictions should be consistent when fed perturbed versions of the same image, which is widely considered in recent semi-supervised learning [29, 26, 6]. The perturbation can come from data augmentation [26], temporal ensembling [29, 17, 12] and shallow-deep features [45, 6]. Artificial perturbations are applied in contrastive learning as strong augmentation [7, 36] and momentum encoder [14] to make a model robust to data variance. Based on temporal ensembling, Ge *et al*. [12] use inter-instance similarity to mitigate pseudo label noise between different training epochs for image localization. Wei *et al*. [36] propose to regularize inter-instance consistency between two sets of augmented views, which neglects intra-class variance problem. We simultaneously reduce intra-class variance and regularize consistency between augmented and original views, which is more suitable for fine-grained ReID tasks.

3. Proposed Method

3.1. Overview

Given a person ReID dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, our objective is to train a robust model on $\mathcal{X}$ without annotation. For inference, representations of a same person are supposed to be as close as possible. State-of-the-art contrastive methods [14, 5] consider each image as an individual class and maximize similarities between augmented views of a same instance with InfoNCE loss [30]:

$$\mathcal{L}_{InfoNCE} = \mathbb{E}[-\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}] \quad (1)$$

where q and k_+ are two augmented views of a same instance in a set of candidates k_i. τ is a temperature hyper-parameter

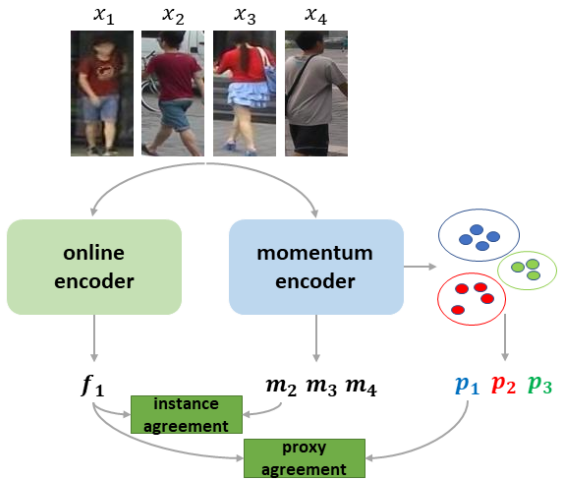

Figure 1: General architecture of ICE. We maximize the similarity between anchor and pseudo positives in both inter-class (proxy agreement between an instance representation f_1 and its cluster proxy p_1) and intra-class (instance agreement between f_1 and its pseudo positive m_2) manners.

that controls the scale of similarities.

Following MoCo [14], we design our proposed ICE with an online encoder and a momentum encoder as shown in Fig. 1. The online encoder is a regular network, *e.g.*, ResNet50 [15], which is updated by back-propagation. The momentum encoder (weights noted as θ_m) has the same structure as the online encoder, but updated by accumulated weights of the online encoder (weights noted as θ_o):

$$\theta_m^t = \alpha\theta_m^{t-1} + (1-\alpha)\theta_o^t \tag{2}$$

where α is a momentum coefficient that controls the update speed of the momentum encoder. t and $t-1$ refer respectively to the current and last iteration. The momentum encoder builds momentum representations with the moving averaged weights, which are more stable to label noise.

At the beginning of each training epoch, we use the momentum encoder to extract appearance representations $\mathcal{M} = \{m_1, m_2, ..., m_N\}$ of all the samples in the training set $\mathcal{X}$. We use a clustering algorithm DBSCAN [9] on these appearance representations to generate pseudo identity labels $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$. We only consider clustered inliers for contrastive learning, while un-clustered outliers are discarded. We calculate proxy centroids $p_1, p_2, ...$ and store them in a memory for a proxy contrastive loss $\mathcal{L}_{proxy}$ (see Sec. 3.2). Note that this proxy memory can be camera-agnostic [13] or camera-aware [34].

Then, we use a random identity sampler to split the training set into mini-batches where each mini-batch contains N_P pseudo identities and each identity has N_K instances. We train the whole network by combining the $\mathcal{L}_{proxy}$ (with class-level labels), a hard instance contrastive loss $\mathcal{L}_{h\_ins}$ (with hard instance pairwise labels, see Sec. 3.3) and a soft instance consistency loss $\mathcal{L}_{s\_ins}$ (with soft instance pairwise labels, see Sec. 3.4):

$$\mathcal{L}_{total} = \mathcal{L}_{proxy} + \lambda_h\mathcal{L}_{h\_ins} + \lambda_s\mathcal{L}_{s\_ins} \tag{3}$$

To increase the consistency before and after data augmentation, we use different augmentation settings for prediction and target representations in the three losses (see Tab. 1).

Loss	Predictions (augmentation)	Targets (augmentation)
$\mathcal{L}_{proxy}$	f (Strong)	p (None)
$\mathcal{L}_{h\_ins}$	f (Strong)	m (Strong)
$\mathcal{L}_{s\_ins}$	P (Strong)	Q (None)

Table 1: Augmentation settings for 3 losses.

3.2. Proxy Centroid Contrastive Baseline

For a camera-agnostic memory, the proxy of cluster a is defined as the averaged momentum representations of all the instances belonging to this cluster:

$$p_a = \frac{1}{N_a}\sum_{m_i \in y_a} m_i \tag{4}$$

where N_a is the number of instances belonging to the cluster a.

We apply a set of data augmentation on $\mathcal{X}$ and feed them to the online encoder. For an online representation f_a belonging to the cluster a, the camera-agnostic proxy contrastive loss is a softmax log loss with one positive proxy p_a and all the negatives in the memory:

$$\mathcal{L}_{agnostic} = \mathbb{E}[-\log\frac{\exp\left(f_a \cdot p_a/\tau_a\right)}{\sum_{i=1}^{|p|}\exp\left(f_a \cdot p_i/\tau_a\right)}] \tag{5}$$

where $|p|$ is the number of clusters in a training epoch and τ_a is a temperature hyper-parameter. Different from unified contrastive loss [11], outliers are not considered as single instance clusters. In such way, outliers are not pushed away from clustered instances, which allows us to mine more hard samples for our proposed hard instance contrast. As shown in Fig. 2, all the clustered instances converge to a common cluster proxy centroid. However, images inside a cluster are prone to be affected by camera styles, leading to high intra-class variance. This problem can be alleviated by adding a cross-camera proxy contrastive loss [34].

For a camera-aware memory, if we have $\mathcal{C} = \{c_1, c_2, ...\}$ cameras, a camera proxy p_{ab} is defined as the averaged momentum representations of all the instances belonging to the cluster a in camera c_b:

$$p_{ab} = \frac{1}{N_{ab}}\sum_{m_i \in y_a \cap m_i \in c_b} m_i \tag{6}$$

where N_{ab} is the number of instances belonging to the cluster a captured by camera c_b.

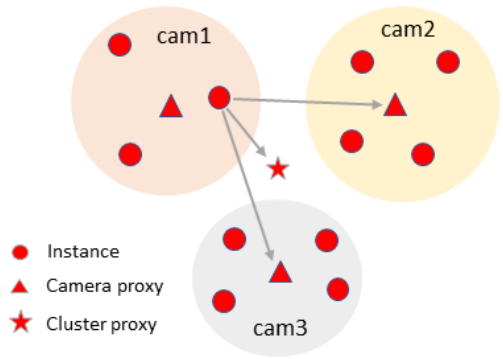

Figure 2: Proxy contrastive loss. Inside a cluster, an instance is pulled to a cluster centroid by $L_{agnostic}$ and to cross-camera centroids by L_{cross}.

Given an online representation f_{ab}, the cross-camera proxy contrastive loss is a softmax log loss with one positive cross-camera proxy p_{ai} and N_{neg} nearest negative proxies in the memory:

$$\mathcal{L}_{cross} = \mathbb{E}[-\frac{1}{|\mathcal{P}|}\sum_{i\neq b \cap i\in\mathcal{C}} \log \frac{\exp\left(< f_{ab}\cdot p_{ai} > /\tau_c\right)}{\sum_{j=1}^{N_{neg}+1}\exp\left(< f_{ab}\cdot p_j > /\tau_c\right)}] \tag{7}$$

where $< \cdot >$ denotes cosine similarity and τ_c is a cross-camera temperature hyper-parameter. $|\mathcal{P}|$ is the number of cross-camera positive proxies. Thanks to this cross-camera proxy contrastive loss, instances from one camera are pulled closer to proxies of other cameras, which reduces intra-class camera style variance.

We define a proxy contrastive loss by combining cluster and camera proxies with a weighting coefficient 0.5 from [34]:

$$\mathcal{L}_{proxy} = \mathcal{L}_{agnostic} + 0.5\mathcal{L}_{cross} \tag{8}$$

3.3. Hard Instance Contrastive Loss

Although intra-class variance can be alleviated by cross-camera contrastive loss, it has two drawbacks: 1) more memory space is needed to store camera-aware proxies, 2) impossible to use when camera ids are unavailable. We propose a camera-agnostic alternative by exploring inter-instance relationship instead of using camera labels. Along with training, the encoders become more and more strong, which helps outliers progressively enter clusters and become hard inliers. Pulling hard inliers closer to normal inliers effectively increases the compactness of clusters.

A mini-batch is composed of N_P identities, where each identity has N_K positive instances. Given an anchor instance f^i belonging to the ith class, we sample the hardest positive momentum representation m_k^i that has the lowest cosine similarity with f^i, see Fig. 4. For the same anchor, we have $J = (N_P - 1) \times N_K$ negative instances that do not belong to the ith class. The hard instance contrastive loss for f^i is a softmax log loss of $J + 1$ (1 positive and J negative) pairs, which is defined as:

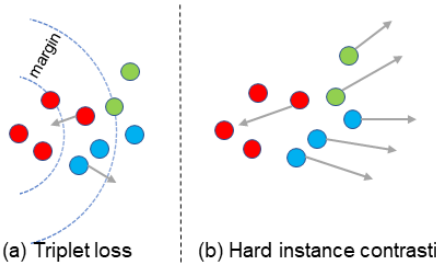

Figure 3: Comparison between triplet and hard instance contrastive loss.

$$\mathcal{L}_{h\_ins} = \mathbb{E}[-\log \frac{\exp\left(< f^i \cdot m_k^i > /\tau_{h\_ins}\right)}{\sum_{j=1}^{J+1}\exp\left(< f^i \cdot m_j > /\tau_{h\_ins}\right)}] \tag{9}$$

where $k = \arg\min_{k=1,..,N_K}(< f^i \cdot m_k^i >)$ and τ_{h_ins} is the hard instance temperature hyper-parameter. By minimizing the distance between the anchor and the hardest positive and maximizing the distance between the anchor and all negatives, $\mathcal{L}_{h\_ins}$ increases intra-class compactness and inter-class separability.

Relation with triplet loss. Both $\mathcal{L}_{h\_ins}$ and triplet loss [16] pull an anchor closer to positive instances and away from negative instances. As shown in Fig. 3, the traditional triplet loss pushes away a negative pair from a positive pair by a margin. Differently, the proposed $\mathcal{L}_{h\_ins}$ pushes away all the negative instances as far as it could with a softmax. If we select one negative instance, the $\mathcal{L}_{h\_ins}$ can be transformed into the triplet loss. If we calculate pairwise distance within a mini-batch to select the hardest positive and the hardest negative instances, the $\mathcal{L}_{h\_ins}$ is equivalent to the batch-hard triplet loss[16]. We compare hard triplet loss (hardest negative) with the proposed $\mathcal{L}_{h\_ins}$ (all negatives). in Tab. 2.

Negative in $\mathcal{L}_{h\_ins}$	Market1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
hardest	80.1	92.8	68.2	82.5
all	**82.3**	**93.8**	**69.9**	**83.3**

Table 2: Comparison between using the hardest negative and all negatives in the denominator of $\mathcal{L}_{h\_ins}$.

3.4. Soft Instance Consistency Loss

Both proxy and hard instance contrastive losses are trained with one-hot hard pseudo labels, which can not capture the complex inter-instance similarity relationship between multiple pseudo positives and negatives. Especially, inter-instance similarity may change after data augmentation. As shown in Fig. 4, the anchor A becomes less similar to pseudo positives (P_1, P_2, P_3), because of the visual distortions. Meanwhile, the anchor A becomes more similar to pseudo negatives (N_1, N_2), since both of them have red shirts. By maintaining the consistency before and after

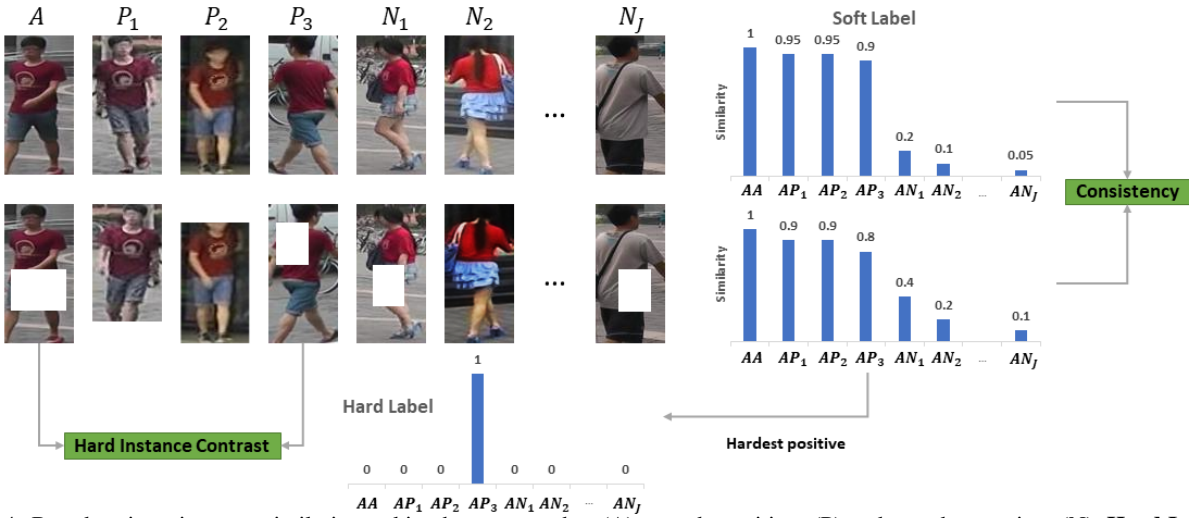

Figure 4: Based on inter-instance similarity ranking between anchor (A), pseudo positives (P) and pseudo negatives (N), **Hard Instance Contrastive Loss** matches an anchor with its hardest positive in a mini-batch. **Soft Instance Consistency Loss** regularizes the inter-instance similarity before and after data augmentation.

augmentation, a model is supposed to be more invariant to augmentation perturbations. We use the inter-instance similarity scores without augmentation as soft labels to rectify those with augmentation.

For a batch of images after data augmentation, we measure the inter-instance similarity between an anchor f_A with all the mini-batch $N_K \times N_P$ instances, as shown in Fig. 4. Then, the inter-instance similarity is turned into a prediction distribution P by a softmax:

$$P = \frac{\exp\left(< f_A \cdot m > /\tau_{s\_ins}\right)}{\sum_{j=1}^{N_P \times N_K} \exp\left(< f_A \cdot m_j > /\tau_{s\_ins}\right)} \tag{10}$$

where τ_{s_ins} is the soft instance temperature hyper-parameter. f_A is an online representation of the anchor, while m is momentum representation of each instance in a mini-batch.

For the same batch without data augmentation, we measure the inter-instance similarity between momentum representations of the same anchor with all the mini-batch $N_K \times N_P$ instances, because the momentum encoder is more stable. We get a target distribution Q:

$$Q = \frac{\exp\left(< m_A \cdot m > /\tau_{s\_ins}\right)}{\sum_{j=1}^{N_P \times N_K} \exp\left(< m_A \cdot m_j > /\tau_{s\_ins}\right)} \tag{11}$$

The soft instance consistency loss is Kullback-Leibler Divergence between two distributions:

$$\mathcal{L}_{s\_ins} = \mathcal{D}_{KL}(P||Q) \tag{12}$$

In previous methods, consistency is regularized between weakly augmented and strongly augmented images [26] or two sets of differently strong augmented images [36]. Some methods [17, 29] also adopted mean square error (MSE) as their consistency loss function. We compare our setting with other possible settings in Tab. 3.

Consistency	Market1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
MSE	80.0	92.7	68.4	82.1
Strong-strong Aug	80.4	92.8	68.2	82.5
ours	**82.3**	**93.8**	**69.9**	**83.3**

Table 3: Comparison of consistency loss. Ours refers to KL divergence between images with and without data augmentation.

4. Experiments

4.1. Datasets and Evaluation Protocols

Market-1501 [43], DukeMTMC-reID[24] and MSMT17 [37] datasets are used to evaluate our proposed method. Market-1501 dataset is collected in front of a supermarket in Tsinghua University from 6 cameras. It contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for test. DukeMTMC-reID is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons for training, 2,228 query images and 17,661 gallery images of 702 persons for test from 8 cameras. MSMT17 is a large-scale Re-ID dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras. Both Cumulative Matching Characteristics (CMC) Rank1, Rank5, Rank10 accuracies and mean Average Precision (mAP) are used in our experiments.

4.2. Implementation details

General training settings. To conduct a fair comparison with state-of-the-art methods, we use an ImageNet [25] pre-trained ResNet50 [15] as our backbone network. We report results of IBN-ResNet50 [23] in Appendix. An Adam optimizer with a weight decay rate of 0.0005 is used to optimize our networks. The learning rate is set to 0.00035 with a warm-up scheme in the first 10 epochs. No learning rate decay is used in the training. The momentum encoder is up-

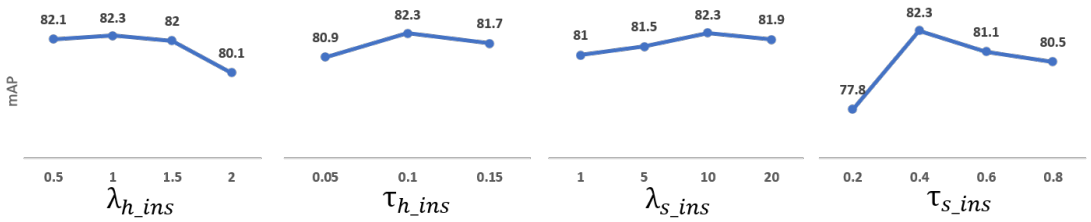

Figure 5: Parameter analysis on Market-1501 dataset.

dated with a momentum coefficient $\alpha = 0.999$. We renew pseudo labels every 400 iterations and repeat this process for 40 epochs. We use a batchsize of 32 where $N_P = 8$ and $N_K = 4$. We set $\tau_a = 0.5$, $\tau_c = 0.07$ and $N_{neg} = 50$ in the proxy contrastive baseline. Our network is trained on 4 Nvidia 1080 GPUs under Pytorch framework. The total training time is around 2 hours on Market-1501. After training, only the momentum encoder is used for the inference.

Clustering settings. We calculate k-reciprocal Jaccard distance [46] for clustering, where k is set to 30. We set a minimum cluster samples to 4 and a distance threshold to 0.55 for DBSCAN. We also report results of a smaller threshold 0.5 (more appropriate for the smaller dataset Market1501) and a larger threshold 0.6 (more appropriate for the larger dataset MSMT17) in Appendix.

Data augmentation. All images are resized to 256×128. The strong data augmentation refers to random horizontal flipping, cropping, Gaussian blurring and erasing [47].

4.3. Parameter analysis

Compared to the proxy contrastive baseline, ICE brings in four more hyper-parameters, including λ_{h_ins}, τ_{h_ins} for hard instance contrastive loss and λ_{s_ins}, τ_{s_ins} for soft instance consistency loss. We analyze the sensitivity of each hyper-parameter on the Market-1501 dataset. The mAP results are illustrated in Fig. 5. As hardest positives are likely to be false positives, an overlarge λ_{h_ins} or undersized τ_{h_ins} introduce more noise. λ_{h_ins} and λ_{s_ins} balance the weight of each loss in Eq. (3). Given the results, we set $\lambda_{h\_ins} = 1$ and $\lambda_{s\_ins} = 10$. τ_{h_ins} and τ_{s_ins} control the similarity scale in hard instance contrastive loss and soft instance consistency loss. We finally set $\tau_{h\_ins} = 0.1$ and $\tau_{s\_ins} = 0.4$. Our hyper-parameters are tuned on Market-1501 and kept same for DukeMTMC-reID and MSMT17. Achieving state-of-the-art results simultaneously on the three datasets can validate the generalizability of these hyper-parameters.

4.4. Ablation study

The performance boost of ICE in unsupervised ReID mainly comes from the proposed hard instance contrastive loss and soft instance consistency loss. We conduct ablation experiments to validate the effectiveness of each loss, which is reported in Tab. 4. We illustrate the number of clusters during the training in Fig. 6 and t-SNE [31] after training in Fig. 8 to evaluate the compactness of clusters. We also illustrate the dynamic KL divergence of Eq. (12) to measure representation sensitivity to augmentation perturbation in Fig. 7 .

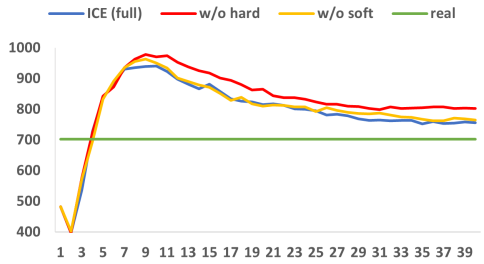

Figure 6: Dynamic cluster numbers during 40 training epochs on DukeMTMC-reID. "hard" and "soft" respectively denote L_{h_ins} and L_{s_ins}. A lower number denotes that clusters are more compact.

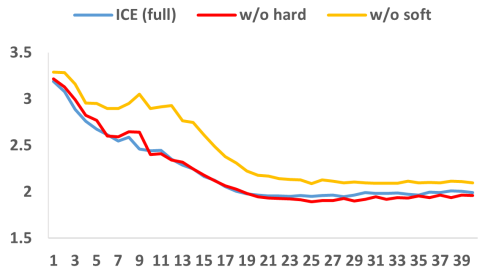

Figure 7: Dynamic KL divergence during 40 training epochs on DukeMTMC-reID. Lower KL divergence denotes that a model is more robust to augmentation perturbation.

Hard instance contrastive loss. Our proposed $\mathcal{L}_{h\_ins}$ reduces the intra-class variance in a camera-agnostic manner, which increases the quality of pseudo labels. By reducing intra-class variance, a cluster is supposed to be more compact. With a same clustering algorithm, we expect to have less clusters when clusters are more compact. As shown in Fig. 6, DBSCAN generated more clusters during the training without our proposed $\mathcal{L}_{h\_ins}$. The full ICE framework has less clusters, which are closer to the real number of identities in the training set. On the other hand, as shown in Fig. 8, the full ICE framework has a better intra-class compactness and inter-class separability than the camera-aware

Camera-aware memory	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Baseline $\mathcal{L}_{proxy}$	79.3	91.5	96.8	97.6	67.3	81.4	90.8	92.9	36.4	67.8	78.7	82.5
$+\mathcal{L}_{h\_ins}$	80.5	92.6	97.3	98.4	68.8	82.4	90.4	93.6	38.0	69.1	79.9	83.4
$+\mathcal{L}_{s\_ins}$	81.1	93.2	97.5	98.5	68.4	82.0	91.0	93.2	38.1	68.7	79.8	83.7
$+\mathcal{L}_{h\_ins}+\mathcal{L}_{s\_ins}$	**82.3**	**93.8**	**97.6**	**98.4**	**69.9**	**83.3**	**91.5**	**94.1**	**38.9**	**70.2**	**80.5**	**84.4**
Camera-agnostic memory	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Baseline $\mathcal{L}_{agnostic}$	65.8	85.3	95.1	96.6	50.9	67.9	81.6	86.6	24.1	52.3	66.2	71.6
$+\mathcal{L}_{h\_ins}$	78.2	91.3	96.9	98.0	65.4	79.6	88.9	91.9	**30.3**	**60.8**	**72.9**	**77.6**
$+\mathcal{L}_{s\_ins}$	47.2	66.7	86.0	91.6	36.2	50.4	70.3	76.3	17.8	38.8	54.2	60.9
$+\mathcal{L}_{h\_ins}+\mathcal{L}_{s\_ins}$	**79.5**	**92.0**	**97.0**	**98.1**	**67.2**	**81.3**	**90.1**	**93.0**	29.8	59.0	71.7	77.0

Table 4: Comparison of different losses. Camera-aware memory occupies up to 6, 8 and 15 times memory space than camera-agnostic memory on Market1501, DukeMTMC-reID and MSMT17 datasets.

baseline in the test set. The compactness contributes to better unsupervised ReID performance in Tab. 4.

Soft instance consistency loss. Hard instance contrastive loss reduces the intra-class variance between naturally captured views, while soft instance consistency loss mainly reduces the variance from artificially augmented perturbation. If we compare the blue (ICE full) and yellow (w/o soft) curves in Fig. 7, we can find that the model trained without $\mathcal{L}_{s\_ins}$ is less robust to augmentation perturbation. The quantitative results in Tab. 4 confirms that the $\mathcal{L}_{s\_ins}$ improves the performance of baseline. The best performance can be obtained by applying $\mathcal{L}_{h\_ins}$ and $\mathcal{L}_{s\_ins}$ on the camera-aware baseline.

Camera-agnostic scenario. Above results are obtained with a camera-aware memory, which strongly relies on ground truth camera ids. We further validate the effectiveness of the two proposed losses with a camera-agnostic memory, whose results are also reported in Tab. 4. Our proposed $\mathcal{L}_{h\_ins}$ significantly improves the performance from the camera-agnostic baseline. However, $\mathcal{L}_{s\_ins}$ **should be used under low intra-class variance, which can be achieved by the variance constraints on camera styles** $\mathcal{L}_{cross}$ **and hard samples** $\mathcal{L}_{h\_ins}$. $\mathcal{L}_{h\_ins}$ reduces intra-class variance, so that $AA \approx AP_1 \approx AP_2 \approx AP_3 \approx 1$ before augmentation in Fig. 4. $\mathcal{L}_{s\_ins}$ permits that we still have $AA \approx AP_1 \approx AP_2 \approx AP_3 \approx 1$ after augmentation. However, when strong variance exists, *e.g.*, $AA \not\approx AP_1 \not\approx AP_2 \not\approx AP_3 \not\approx 1$, maintaining this relationship equals maintaining intra-class variance, which decreases the ReID performance. On medium datasets (*e.g.*, Market1501 and DukeMTMC-reID) without strong camera variance, our proposed camera-agnostic intra-class variance constraint $\mathcal{L}_{h\_ins}$ is enough to make $\mathcal{L}_{s\_ins}$ beneficial to ReID. On large datasets (*e.g.*, 15 cameras in MSMT17) with strong camera variance, only camera-agnostic variance constraint $\mathcal{L}_{h\_ins}$ is not enough. We provide the dynamic cluster numbers of camera-agnostic ICE in Appendix.

4.5. Comparison with state-of-the-art methods

We compare ICE with state-of-the-art ReID methods in Tab. 5.

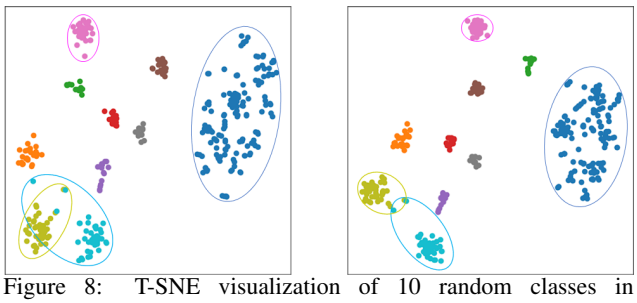

Figure 8: T-SNE visualization of 10 random classes in DukeMTMC-reID test set between **camera-aware baseline (Left)** and **ICE (Right)**.

Comparison with unsupervised method. Previous unsupervised methods can be categorized into unsupervised domain adaptation (UDA) and fully unsupervised methods. We first list state-of-the-art UDA methods, including MMCL [32], JVTC [18], DG-Net++ [51], ECN+ [50], MMT [11], DCML [1], MEB [41], SpCL [13] and ABMT [3]. UDA methods usually rely on source domain annotation to reduce the pseudo label noise. Without any identity annotation, our proposed ICE outperforms all of them on the three datasets.

Under the fully unsupervised setting, ICE also achieves better performance than state-of-the-art methods, including BUC [20], SSL [21], MMCL [32], JVTC [18], HCT [40], CycAs [35], GCL [4], SpCL [13] and CAP [34]. CycAs leveraged temporal information to assist visual matching, while our method only considers visual similarity. SpCL and CAP are based on proxy contrastive learning, which are considered respectively as camera-agnostic and camera-aware baselines in our method. With a camera-agnostic memory, the performance of ICE(agnostic) remarkably surpasses the camera-agnostic baseline SpCL, especially on Market1501 and MSMT17 datasets. With a camera-aware memory, ICE(aware) outperforms the camera-aware baseline CAP on all the three datasets. By mining hard positives to reduce intra-class variance, ICE is more robust to hard samples. We illustrate some hard examples in Fig. 9, where ICE succeeds to notice important visual clues, *e.g.*, characters in the shirt (1st row), blonde hair (2nd row), brown shoulder bag (3rd row) and badge (4th row).

Method	Reference	Market1501				DukeMTMC-reID				MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Unsupervised Domain Adaptation													
MMCL [32]	CVPR'20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0	16.2	43.6	54.3	58.9
JVTC [18]	ECCV'20	61.1	83.8	93.0	95.2	56.2	75.0	85.1	88.2	20.3	45.4	58.4	64.3
DG-Net++ [51]	ECCV'20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.4	22.1	48.8	60.9	65.9
ECN+ [50]	TPAMI'20	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4	16.0	42.5	55.9	61.5
MMT [11]	ICLR'20	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5	23.3	50.1	63.9	69.8
DCML [1]	ECCV'20	72.6	87.9	95.0	96.7	63.3	79.1	87.2	89.4	-	-	-	-
MEB [41]	ECCV'20	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2	-	-	-	-
SpCL [13]	NeurIPS'20	76.7	90.3	96.2	97.7	68.8	82.9	90.1	92.5	26.8	53.7	65.0	69.8
ABMT [3]	WACV'21	78.3	92.5	-	-	69.1	82.0	-	-	26.5	54.3	-	-
Fully Unsupervised													
BUC [20]	AAAI'19	29.6	61.9	73.5	78.2	22.1	40.4	52.5	58.2	-	-	-	-
SSL [21]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9	-	-	-	-
JVTC [18]	ECCV'20	41.8	72.9	84.2	88.7	42.2	67.6	78.0	81.6	15.1	39.0	50.9	56.8
MMCL [32]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0	11.2	35.4	44.8	49.8
HCT [40]	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4	-	-	-	-
CycAs [35]	ECCV'20	64.8	84.8	-	-	60.1	77.9	-	-	26.7	50.1	-	-
GCL [4]	CVPR'21	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5	21.3	45.7	58.6	64.5
SpCL(agnostic) [13]	NeurIPS'20	73.1	88.1	95.1	97.0	65.3	81.2	90.3	92.2	19.1	42.3	55.6	61.2
ICE(agnostic)	This paper	79.5	92.0	97.0	98.1	67.2	81.3	90.1	93.0	29.8	59.0	71.7	77.0
CAP(aware)[34]	AAAI'21	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
ICE(aware)	This paper	82.3	93.8	97.6	98.4	69.9	83.3	91.5	94.1	38.9	70.2	80.5	84.4
Supervised													
PCB [28]	ECCV'18	81.6	93.8	97.5	98.5	69.2	83.3	90.5	92.5	40.4	68.2	-	-
DG-Net [44]	CVPR'19	86.0	94.8	-	-	74.8	86.6	-	-	52.3	77.2	-	-
ICE (w/ ground truth)	This paper	86.6	95.1	98.3	98.9	76.5	88.2	94.1	95.7	50.4	76.4	86.6	90.0

Table 5: Comparison of ReID methods on Market1501, DukeMTMC-reID and MSMT17 datasets. The best and second best unsupervised results are marked in red and blue.

Comparison with supervised method. We further provide two well-known supervised methods for reference, including the Part-based Convolutional Baseline (PCB) [28] and the joint Discriminative and Generative Network (DG-Net) [44]. Unsupervised ICE achieves competitive performance with PCB. If we replace the clustering generated pseudo labels with ground truth, our ICE can be transformed into a supervised method. The supervised ICE is competitive with state-of-the-art supervised ReID methods (*e.g.*, DG-Net), which shows that the supervised contrastive learning has a potential to be considered into future supervised ReID.

5. Conclusion

In this paper, we propose a novel inter-instance contrastive encoding method ICE to address unsupervised ReID. Deviated from previous proxy based contrastive ReID methods, we focus on inter-instance affinities to make a model more robust to data variance. We first mine the hardest positive with mini-batch instance pairwise similarity ranking to form a hard instance contrastive loss, which effectively reduces intra-class variance. Smaller intra-class variance contributes to the compactness of clusters. Then, we use mini-batch instance pairwise similarity scores as soft labels to enhance the consistency before and after data augmentation, which makes a model robust to artificial augmentation variance. By combining the proposed hard instance contrastive loss and soft instance consistency loss,

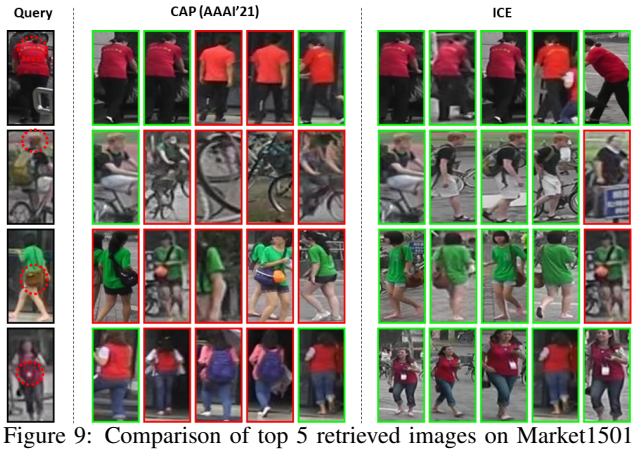

Figure 9: Comparison of top 5 retrieved images on Market1501 between CAP [34] and ICE. Green boxes denote correct results, while red boxes denote false results. Important visual clues are marked with red dashes.

ICE significantly outperforms previous unsupervised ReID methods on Market1501, DukeMTMC-reID and MSMT17 datasets.

Acknowledgements. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] Guangyi Chen, Yuhao Lu, Jiwen Lu, and Jie Zhou. Deep credible metric learning for unsupervised domain adaptation person re-identification. In *ECCV*, 2020. 2, 7, 8

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In *WACV*, 2020. 1

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification. In *WACV*, 2021. 2, 7, 8

[4] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, 2021. 2, 7, 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2

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

[8] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, 2019. 2

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 3

[10] 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 *ICCV*, 2019. 2

[11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 1, 2, 3, 7, 8

[12] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, 2020. 2

[13] 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 *NeurIPS*, 2020. 1, 2, 3, 7, 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 4

[17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2, 5

[18] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 2, 7, 8

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

[20] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 1, 2, 7, 8

[21] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020. 1, 2, 7, 8

[22] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, June 2019. 1

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 5

[24] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshops*, 2016. 5

[25] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[26] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 5

[27] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *PR*, 2020. 1, 2

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1, 8

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2, 5

[30] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 2008. 6

[32] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020. 1, 2, 7, 8

[33] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, 2018. 2

[34] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, 2021. 2, 3, 4, 7, 8

[35] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 2, 7, 8

[36] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. In *ICLR*, 2021. 2, 5

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 2, 5

[38] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *CVPR*, 2018. 1, 2

[39] Hong-Xing Yu, W. Zheng, Ancong Wu, X. Guo, S. Gong, and J. Lai. Unsupervised person re-identification by soft multilabel learning. *CVPR*, 2019. 2

[40] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, 2020. 2, 7, 8

[41] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020. 1, 7, 8

[42] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019. 2

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *ICCV*, 2015. 5

[44] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 8

[45] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 2021. 2

[46] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6

[48] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 2

[49] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 2

[50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE TPAMI*, 2020. 7, 8

[51] Yang Zou, Xiaodong Yang, Zhiding Yu, B. V. K. Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020. 2, 7, 8