# Cluster Contrast for Unsupervised Person Re-Identification

Zuozhuo Dai[1], Guangyuan Wang[1], Weihao Yuan[1], Siyu Zhu[1], and Ping Tan[2]

[1] Alibaba Cloud
[2] Simon Fraser University

**Abstract.** Thanks to the recent research development in contrastive learning, the gap of visual representation learning between supervised and unsupervised approaches has been gradually closed in the tasks of computer vision. In this paper, we focus on the downstream task of unsupervised person re-identification (re-ID). State-of-the-art unsupervised re-ID methods train the neural networks using a dictionary-based non-parametric softmax loss. They store the pre-computed instance feature vectors inside the dictionary, assign pseudo labels to them using clustering algorithm, and compare the query instances to the cluster using a form of contrastive loss. To enforce a consistent dictionary, that is the features in the dictionary are computed by a similar or the same encoder network, we present Cluster Contrast which stores feature vectors and computes contrastive loss at the cluster level. Moreover, the momentum update is introduced to reinforce the cluster-level feature consistency in the sequential space. Despite the straightforward design, experiments on four representative re-ID benchmarks demonstrate the effective performance of our method.

**Keywords:** Person re-ID · Unsupervised learning · Contrastive learning.

## 1 Introduction

Deep unsupervised person re-ID aims to train a neural network capable of retrieving a person of interest across cameras without any labeled data. This task has attracted increasing attention recently due to the growing demands in practical video surveillance and the expensive labeling cost. There are mainly two approaches to address this problem. One is the purely unsupervised learning person re-ID, which generally exploits pseudo labels from the unlabeled data [14, 15, 17, 30, 47]. The other is the unsupervised domain adaptation person re-ID, which first pre-trains a model on the source labeled dataset, and then fine-tunes the model on the target unlabeled dataset [12, 29, 48, 51, 59, 67, 68]. Generally, the performance of domain adaptation is superior to that of unsupervised learning because of the introduction of the external source domain. However, domain adaptation still suffers from the complex training procedure and requires that the difference between the source and target domain is not significant. In this paper, we focus on learning the person re-ID task without any labeled data, namely the purely unsupervised learning.
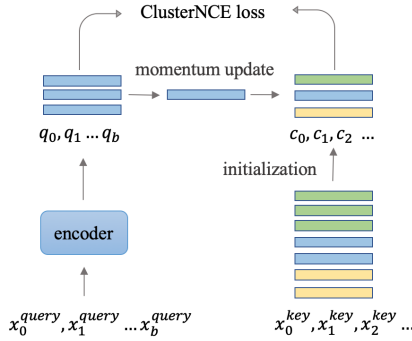
Fig. 1: The Cluster Contrast computes the contrastive loss in cluster level with momentum update. In the cluster level memory dictionary, the cluster feature is initialized as the averaged feature of the corresponding cluster and updated by the batch query instance. $x \in X$ is the training dataset. $q$ is the query instance feature vector. $c_k$ stands for the $k$-th cluster feature vector. Feature vectors with the same color belong to the same cluster.

Recently, the unsupervised representation learning methods [1, 2, 5, 6, 19, 20, 22, 24, 35, 36, 45, 53, 58] with contrastive loss [19] have gradually closed the performance gap with supervised pretraining in computer vision. Here, the contrastive loss [19] aims to compare pairs of image features so that the positive sample pairs are pulled together and the negative sample pairs are pulled away. Specifically, InstDisc [53] proposes an instance level memory bank for instance discrimination. It compares the query image features to all the instance features in the memory bank. Subsequently, the MoCo series [6, 7, 20] highlight the consistent memory dictionary in contrastive learning of visual representations. MoCo approximates the contrastive loss by sampling a subset of instances in the memory dictionary and uses the momentum-based moving average of the query encoder. Meanwhile, SimCLR [5] uses a large enough batch size to compute contrastive loss, which requires hundreds of TPU cores. Later on, SwAV [2] computes the contrastive loss in cluster level. It enforces the cluster assignment results rather than comparing sampling instance features. Since the cluster number is fixed in online clustering, SwAV does not require the large instance feature memory bank or large batch size to enforce the feature consistency.

Inspired by the great success of contrastive learning, recent works [4, 17, 47, 49, 14, 15, 30, 57, 65, 62] try to apply such ideology to the downstream re-ID tasks. In more details, such approaches exploit the memory dictionary and pseudo labels from clustering to train the neural network. At the beginning of each epoch, all the image features of the training data are extracted by the current neural network. Then, such image features are stored in a memory dictionary and a clustering algorithm, like DBSCAN [13] or K-means [34] is employed to cluster image features and produce pseudo labels. Meanwhile, the cluster ID is assigned to each image as the person identity. Finally, the neural network is trained with

a contrastive loss such as triplet loss [23, 41], InfoNCE loss [37], or other non-parametric classification loss [47] between the feature vectors of every instance inside the memory dictionary and the query instance. Since the instance features updated in one iteration are limited by the batch size, the instance features from the newly updated encoder network are not consistent with the previous ones. This problem of feature inconsistency in memory dictionary is especially serious in large-scale re-ID datasets like MSMT17 [51].

To enforce a consistent feature dictionary, we propose Cluster Contrast for unsupervised person re-ID. Remarkably, the ideology of Cluster Contrast is inspired by the contrasting cluster assignment technique from SwAV [2]. Different from SwAV which adopts an online clustering approach with a fixed number of clusters, we use an offline clustering method [14, 17] which demonstrates superior clustering performance in re-ID tasks and remove un-clustered outliers. Then, a cluster-level memory dictionary is built and each dictionary key corresponds to a cluster which is represented by a single feature vector. More specifically, this cluster feature is initialized as the average feature of all the images from the same cluster and updated by the batch query instance features during training. Accordingly, we propose a cluster-level InfoNCE loss, denoted as ClusterNCE loss, which computes contrastive loss between cluster feature and query instance feature as illustrated in Figure 1. Moreover, we apply the ideology of momentum update policy from MoCo [20] to the cluster level memory to further boost the feature consistency of cluster representations in the sequential space.

In summary, our proposed Cluster Contrast for unsupervised re-ID has the following contributions:

- We introduce the cluster-level memory dictionary which initializes, updates, and performs contrastive loss computation at the cluster level. The cluster feature embedding helps to alleviate the feature inconsistency problem.
- We apply the momentum updating policy to the cluster feature representation and further enforce the feature consistency in the memory dictionary.
- We demonstrate that the proposed unsupervised approach with Cluster Contrast achieves state-of-the-art performance on three purely unsupervised re-ID benchmarks.

## 2   Related Work

*Deep Unsupervised Person Re-ID.* Deep unsupervised person re-ID can be summarized into two categories. The first category is unsupervised domain adaptation re-ID, which utilizes transfer learning to improve unsupervised person re-ID [17, 47, 12, 16, 26, 61, 29, 48, 51, 59, 67, 68]. The second category is pure unsupervised learning person re-ID [14, 15, 30, 4, 49, 57, 65, 62], which trains model directory on unlabeled dataset. State-of-the-art unsupervised learning re-ID pipeline generally involves three stages: memory dictionary initialization, pseudo label generation, and neural network training. Previous works have made significant improvements either in parts or on the whole pipeline. Specifically, Lin *et al.* [30] treats each individual sample as a cluster, and then gradually groups similar

samples into one cluster to generate pseudo labels. MMCL [47] predicts quality pseudo labels comprising of similarity computation and cycle consistency. It then trains the model as a multi-classification problem. SPCL [17] proposes a novel self-paced contrastive learning framework that gradually creates more reliable cluster to refine the memory dictionary features. OPLG [65] and RLCC [62] explore the temporal label consistency for better pseudo label quality. In addition to pseudo label, another stream of camera-aware methods [49, 4] utilizes camera ID information as additional supervision signal to further improve the unsupervised re-ID performance. In this paper, we focus on purely unsupervised person re-ID, but our method can be easily generalized to unsupervised domain adaptation and camera-aware methods.

*Memory Dictionary.* Contrastive learning [1, 2, 5, 6, 19, 20, 22, 24, 35, 36, 45, 53, 58] can be thought of as training an encoder for a dictionary look-up task. Since it is too expensive in both memory and computation to compare all the image pairs within a dataset, several recent studies [1, 20, 22, 24, 37, 44, 54, 70] on unsupervised visual representation learning present promising results through building dynamic dictionaries. Moco [20] builds a memory dictionary as a queue of sampled images. The samples in memory dictionary is replaced consistently on the fly to keep the feature consistency with the newly updated model. SimCLR [5] shows that the instance memory can be replaced by a large enough batch of instances. Similar to unsupervised visual representation learning, state-of-the-art unsupervised person re-ID methods also build memory dictionaries for contrastive learning [55, 47, 16, 17]. During training, instance feature vectors in the memory dictionary are updated by the corresponding query instances features. Recently, SwAV [2] proposes an efficient online clustering method which approximates the contrastive loss of all image pairs by clustering centers, without requiring a large batch size or large memory bank. Inspired by SwAV [2] and Moco [20], we apply the ideology of cluster level contrastive learning and momentum update to the downstream unsupervised re-ID tasks and alleviate the problems of the large memory bank and memory dictionary inconsistency. Unlike SwAV in which the number of clusters is fixed, the proposed Cluster Contrast gradually selects reliable labels and dynamically refines the clustering results during training.

*Loss Functions.* In supervised person re-ID, the batch hard triplet loss has proved to be effective solutions to improve the re-ID performance [3, 9, 10, 18, 32, 42, 46, 63, 69]. In unsupervised person re-ID, since there is no ground truth person identity and the pseudo labels are changing during training, non-parametric classification loss such as InfoNCE [37] are used as identity loss. Similar to InfoNCE, Tong*et al.* [55] designs an Online Instance Matching (OIM) loss with a memory dictionary scheme which compares query image to a memorized feature set of unlabelled identities. Wang and Zhang [47] introduce the memory-based non-parametric multi-label classification loss (MMCL), which treat unsupervised re-ID as a multi-label classification problem. In order to mitigate noisy pseudo labels, MMT [16] proposes a novel soft softmax-triplet loss to support learning
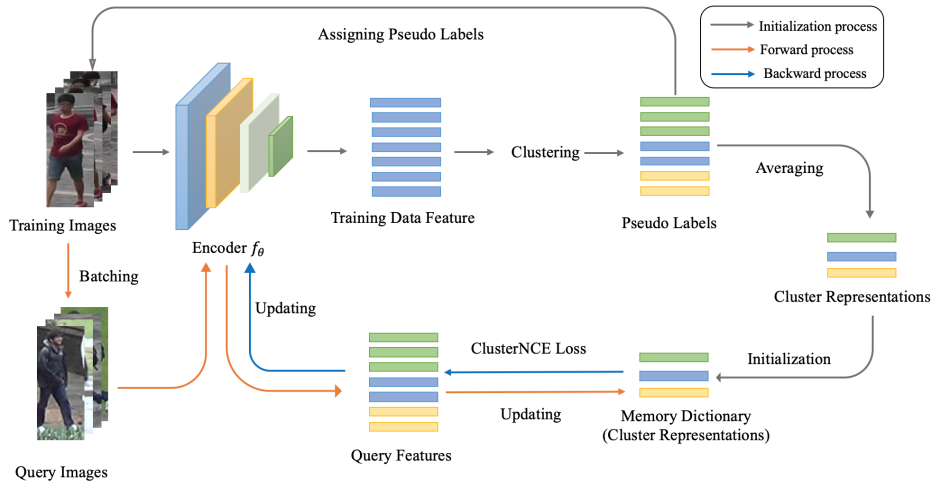
Fig. 2: The system pipeline of our unsupervised person re-ID method. The upper branch depicts the memory initialization stage. The training data features are assigned pseudo labels by clustering, where features of the same color belong to the same cluster. The lower branch represents the model training stage. Query features in iterative mini-batch are used to update the memory cluster representations with a momentum. The ClusterNCE loss computes the contrastive loss between query features and all cluster representations.

with soft pseudo triplet labels. SPCL [17] introduces a unified contrastive loss including both source domain dataset and target domain dataset. In this paper, we apply InfoNCE loss between cluster feature and query instance feature on unsupervised re-ID.

## 3    Method

We first introduce our overall approach at a high level in Section 3.1. Then, we compare the multiple contrastive learning approaches for person re-ID with our proposed cluster contrast method in Section 3.2. Finally in Section 3.3, we explain the details of momentum update in Cluster Contrast along with its working theory.

### 3.1    Overview

State-of-the-art unsupervised learning methods [47, 16, 17, 4] solve the unsupervised learning person re-ID problem with contrastive learning. Specifically, they build a memory dictionary that contains the features of all training images. Each feature is assigned a pseudo ID generated by a clustering algorithm. During training, the contrastive loss is minimized to train the network and learn a proper feature embedding that is consistent with the pseudo ID.

(a) Multi-label classi-      (b) Instance level InfoNCE loss      (c) ClusterNCE loss (ours)
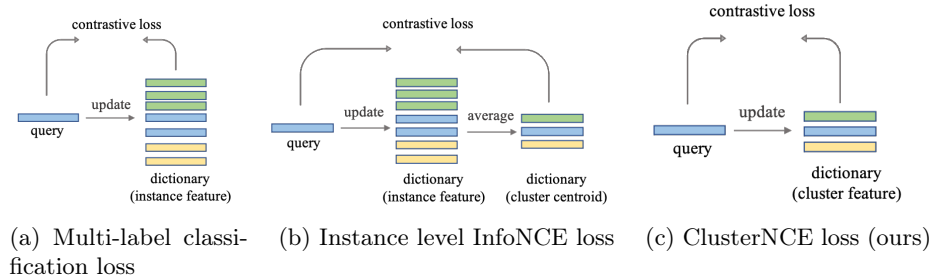fication loss

Fig. 3: Comparison of three types of memory-based non-parametric contrastive learning losses for re-ID. Different color features indicate different clusters. (a) computes the loss and updates the memory dictionary both at the instance level [47]. (b) computes the loss at the cluster level but updates the memory dictionary at the instance level [17]. (c) is our proposed approach and it computes the loss and updates the memory dictionary both at the cluster level.

We focused on designing a proper contrastive learning method to keep the whole pipeline simple while obtaining better performance. An overview of our training pipeline is shown in Figure 2. The memory dictionary initialization is illustrated in the upper branch. We use a standard ResNet50 [21] as the backbone encoder which is pretrained on ImageNet to extract feature vectors, and has basic discriminability though not optimized for re-ID tasks. We then apply the DBSCAN [13] clustering algorithms to cluster similar features together and assign pseudo labels to them. The cluster feature representation is calculated as the mean feature vectors of each cluster. The memory dictionary is initialized by these cluster feature representations and their corresponding pseudo labels. As shown in the lower branch, during the training stage, we compute the ClusterNCE loss between the query image features and all cluster representations in the dictionary to train the network. Meanwhile, the dictionary features are updated with a momentum by the query features.

To facilitate the description of methods, we first introduce the notations used in this paper. Let $X = \{x_1, x_2, \ldots, x_N\}$ denote the training set with $N$ instances. And $U = \{u_1, u_2, \ldots, u_n\}$ denotes the corresponding features obtained from the backbone encoder $f_\theta$, described as $u_i = f_\theta(x_i)$. $q$ is a query instance feature extracted by $f_\theta(\cdot)$, where the query instance belongs to $X$.

## 3.2  Cluster Contrast

In this section, we analyze different contrastive learning methods to motivate our design of Cluster Contrast. As shown in Figure 3 (a), the multi-label classification loss computes the loss in the instance level through an instance-wise contrastive loss. It stores all image feature vectors in the memory dictionary and computes

multi-class score by comparing each query feature to all of them. The memory dictionary is updated by the query features after each training iteration.

In Figure 3 (b), SPCL [17] computes the loss at cluster level through a cluster-wise InfoNCE loss. It can be defined as follows:

$$L_q = -\log \frac{exp(q \cdot c_+/\tau)}{\sum_{k=1}^{K} exp(q \cdot c_k/\tau)} \tag{1}$$

where $\tau$ is a temperature hyper-parameter, $\{c_1, c_2, \ldots, c_K\}$ are the cluster centroids and $K$ stands for the number of clusters. It uses the cluster centroid as the cluster level feature vector to compute the the distances between query instance $q$ and all the clusters. $c_+$ is the positive cluster feature which $q$ belongs to. The cluster centroids are calculated by the mean feature vectors of each cluster as:

$$c_k = \frac{1}{|\mathcal{H}_k|} \sum_{u_i \in \mathcal{H}_k} u_i \tag{2}$$

where $\mathcal{H}_k$ denotes the $k$-th cluster set and $|\cdot|$ indicates the number of instances per cluster. $\mathcal{H}_k$ contains all the feature vectors in the cluster $k$. But similar to multi-classification loss , it stores all image feature vectors in the memory dictionary. The stored image feature vectors are then updated by corresponding query image feature.

Both Figure 3 (a) and Figure 3 (b) update the feature vectors at an instance level, resulting in feature inconsistency problem. As shown in Figure 4, the cluster size is unbalancedly distributed. In every training iteration, in a large cluster only a small fraction of the instance features can be updated due to the batch size limitation, whereas in a small cluster all the instances can be updated. Thus, the updating process is highly varied, and the contrastive loss computed by comparing all instance features is not consistent with the newest model. In each iteration, the network is constantly updated, which causes inconsistent oscillatory distribution of mini-batches. In contrast, we design our ClusterNCE loss as shown in Figure 3 (c) using the following equation:

$$L_q = -\log \frac{exp(q \cdot \phi_+/\tau)}{\sum_{k=1}^{K} exp(q \cdot \phi_k/\tau)} \tag{3}$$

where $\phi_k$ is the unique representation vector of the $k$-th cluster. It updates the feature vectors and computes the loss both in the cluster level.

We can see that, our proposed algorithm uses unique feature vectors to represent each cluster category and remains distinct throughout the updating process, which is the most significant difference from the previous contrastive loss approaches. In the next section, we will discuss in detail how our method consistently updates the cluster representation to maintain the cluster consistency with the help of momentum update.
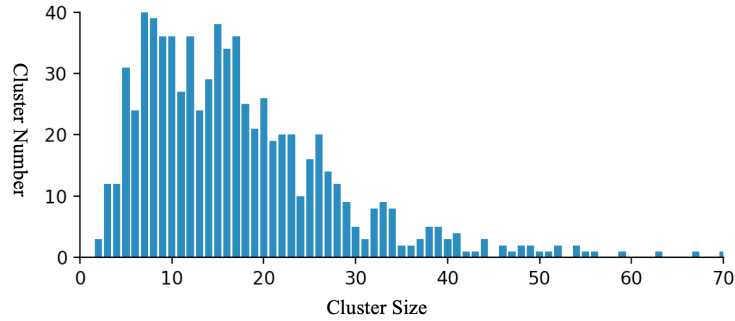
Fig. 4: The cluster size follows a normal distribution in Market1501 dataset.

### 3.3   Momentum Update

In this section, we present how to initialize and update the cluster level memory in the proposed Cluster Contrast method. The training details are presented in Algorithm 1.

*Memory Initialization.* Different from the instance level memory dictionary, we store each cluster's representation $\{\phi_1, \ldots, \phi_K\}$ in the memory-based feature dictionary. We use the mean feature vectors of each cluster to initialize the cluster representation, that is

$$\phi_k = \frac{1}{|\mathcal{H}_k|} \sum_{u_i \in \mathcal{H}_k} u_i \tag{4}$$

Eq. 4 is executed when each epoch is initialized. And the clustering algorithm runs in each epoch, so $K$ is changing as the model trains.

*Memory Updating.* During training, following [23], $P$ person identities and a fixed number $Z$ of instances for each person identity were sampled from the training set. Consequently, we obtain a total number of $P \times Z$ query images in the minibatch. We then momentum update the cluster representation iteratively by the query features in the minibatch by the Eq. 5 as illustrated in Figure 3:

$$\forall q \in Q_k, \phi_k \leftarrow m\phi_k + (1 - m)q \tag{5}$$

where $Q_k$ is the query features encoded from k-th cluster images and $m$ is the momentum updating factor. $m$ controls the consistency between the cluster feature and most updated query instance feature. As $m$ close to 0, the cluster feature $\phi_k$ is close to the newest query feature. It is worth noting that all cluster representations are stored in the memory dictionary, so we calculate loss by comparing each query instance with all cluster representations in each iteration.

---
**Algorithm 1:** Unsupervised learning pipeline with Cluster Contrast
---
**Require:** Unlabeled training data $X$
**Require:** Initialize the backbone encoder $f_\theta$ with ImageNet-pretrained
  ResNet-50
**Require:** Temperature $\tau$ for Eq. 3
**Require:** Momentum $m$ for Eq. 5
**for** $n$ *in* $[1, num\_epochs]$ **do**
>  Extract feature vectors $U$ from $X$ by $f_\theta$
>  Clustering $U$ into $K$ clusters with DBSCAN
>  Initialize memory dictionary with Eq. 4
>  **for** $i$ *in* $[1, num\_iterations]$ **do**
>  >  Sample $P \times K$ query images from $X$
>  >  Compute ClusterNCE loss with Eq. 3
>  >  Update cluster feature with Eq. 5
>  >  Update the encoder $f_\theta$ by optimizer
>
>  **end**

**end**

---

Table 1: Statistics of the datasets used in the experimental section.

| Dataset | Object | #train IDs | #train images | #test IDs | #query images | #total images |
|---|---|---|---|---|---|---|
| Market-1501 | Person | 751 | 1 2,936 | 750 | 3,368 | 32,668 |
| MSMT17 | Person | 1,041 | 32,621 | 3,060 | 11,659 | 126,441 |
| PersonX | Person | 410 | 9,840 | 856 | 5,136 | 45,792 |
| VeRi-776 | Vehicle | 575 | 37,746 | 200 | 1,678 | 51,003 |

## 4    Experiment

### 4.1    Datasets and Implementation

*Datasets.* We evaluate our proposed method on three large-scale person re-ID benchmarks: Market-1501 [64], MSMT17 [51], PersonX [43], and one vehicle ReID dataset, VeRi-776 [33]. Note that the DukeMTMC-reID [40] has been taken down for ethic issues. The Market-1501 and MSMT17 are widely used real-world person re-identification tasks. The PersonX is synthesized based on Unity [39], which contains manually designed obstacles such as random occlusion, resolution, and lighting differences. To show the robustness of our method, we also conduct vehicle re-identification experiments on the widely used real scene VeRi-776 datasets. The details of these datasets are summarized in Table 1.

*Implementation Details.* We adopt ResNet-50 [21] as the backbone encoder of the feature extractor and initialize the model with the parameters pre-trained on ImageNet [11]. After layer-4, we remove all sub-module layers and add global average pooling (GAP) followed by batch normalization layer [25] and L2-normalization layer, which will produce 2048-dimensional features. The Gemeralized-Mean (GeM) pooling [38] can further improve the performance, which can be seen in appendix. During testing, we take the features of the global average pooling

Table 2: Comparison with state-of-the-art methods on the object re-ID benchmarks. The purely unsupervised methods, unsupervised domain adaptation methods, and camera-aware unsupervised methods are considered for the comparison. The unsupervised domain adaptation method uses additional source domain labeled dataset and the camera-aware method uses the camera ID.

(a) Experiments on Market-1501 datasets

| Methods | source | mAP | top-1 | top-5 | top-10 |
|---|---|---|---|---|---|
| **Purely Unsupervised** | | | | | |
| SSL [31] | None | 37.8 | 71.7 | 83.8 | 87.4 |
| MMCL [47] | None | 45.5 | 80.3 | 89.4 | 92.3 |
| HCT [60] | None | 56.4 | 80.0 | 91.6 | 95.2 |
| CycAs [50] | None | 64.8 | 84.8 | - | - |
| UGA [52] | None | 70.3 | 87.2 | - | - |
| SPCL [17] | None | 73.1 | 88.1 | 95.1 | 97.0 |
| IICS [57] | None | 72.1 | 88.8 | 95.3 | 96.9 |
| OPLG [65] | None | 78.1 | 91.1 | 96.4 | 97.7 |
| RLCC [62] | None | 77.7 | 90.8 | 96.3 | 97.5 |
| ICE [4] | None | 79.5 | 92.0 | 97.0 | **98.1** |
| PPLR [8] | None | 81.5 | 92.8 | 97.1 | 98.1 |
| **Ours** | None | **83.0** | **92.9** | **97.2** | 98.0 |
| **Unsupervised Domain Adaptation** | | | | | |
| MMCL [47] | Duke | 60.4 | 84.4 | 92.8 | 95.0 |
| AD-Cluster [61] | Duke | 68.3 | 86.7 | 94.4 | 96.5 |
| MMT [16] | MSMT17 | 75.6 | 89.3 | 95.8 | 97.5 |
| SPCL [17] | MSMT17 | 77.5 | 89.7 | 96.1 | 97.6 |
| TDR [26] | Duke | 83.4 | 94.2 | - | - |
| **Camera-aware Unsupervised** | | | | | |
| CAP [49] | None | 79.2 | 91.4 | 96.3 | 97.7 |
| ICE(aware) [4] | None | 82.3 | 93.8 | 97.6 | 98.4 |
| PPLR(aware) [8] | None | 84.4 | 94.3 | 97.8 | 98.6 |

(b) Experiments on MSMT17 datasets

| Methods | source | mAP | top-1 | top-5 | top-10 |
|---|---|---|---|---|---|
| **Purely Unsupervised** | | | | | |
| TAUDL [27] | None | 12.5 | 28.4 | - | - |
| MMCL [56] | None | 11.2 | 35.4 | 44.8 | 49.8 |
| UTAL [28] | None | 13.1 | 31.4 | - | - |
| CycAs [50] | None | 26.7 | 50.1 | - | - |
| UGA [52] | None | 21.7 | 49.5 | - | - |
| SPCL [17] | None | 19.1 | 42.3 | 55.6 | 61.2 |
| IICS [57] | None | 18.6 | 45.7 | 57.7 | 62.8 |
| OPLG [65] | None | 26.9 | 53.7 | 65.3 | 70.2 |
| RLCC [62] | None | 27.9 | 56.5 | 68.4 | 73.1 |
| ICE [4] | None | 29.8 | 59.0 | 71.7 | 77.0 |
| PPLR [8] | None | 31.4 | 61.1 | **73.4** | **77.8** |
| **Ours** | None | **33.0** | **62.0** | 71.8 | 76.7 |
| **Unsupervised Domain Adaptation** | | | | | |
| MMCL [47] | Duke | 16.2 | 43.6 | 54.3 | 58.9 |
| ECN [68] | Duke | 10.2 | 30.2 | 41.5 | 46.8 |
| MMT [16] | Market | 24.0 | 50.1 | 63.5 | 69.3 |
| SPCL [17] | Market | 26.8 | 53.7 | 65.0 | 69.8 |
| TDR [26] | Duke | 36.3 | 66.6 | - | - |
| **Camera-aware Unsupervised** | | | | | |
| CAP [49] | None | 36.9 | 67.4 | 78.0 | 81.4 |
| ICE(aware) [4] | None | 38.9 | 70.2 | 80.5 | 84.4 |
| PPLR(aware) [8] | None | 42.2 | 73.3 | 83.5 | 86.5 |

(c) Experiments on PersonX datasets

| Methods | source | mAP | top-1 | top-5 | top-10 |
|---|---|---|---|---|---|
| MMT [16] | Market | 78.9 | 90.6 | 96.8 | 98.2 |
| SPCL [17] | Market | 78.5 | 91.1 | 97.8 | 99.0 |
| SPCL [17] | None | 72.3 | 88.1 | 96.6 | 98.3 |
| **Ours** | None | 84.7 | 94.4 | 98.3 | 99.3 |

(d) Experiments on VeRi-776 datasets

| Methods | source | mAP | top-1 | top-5 | top-10 |
|---|---|---|---|---|---|
| MMT [16] | VehicleID | 35.3 | 74.6 | 82.6 | 87.0 |
| SPCL [17] | VehicleID | 38.9 | 80.4 | 86.8 | 89.6 |
| SPCL [17] | None | 36.9 | 79.9 | 86.8 | 89.9 |
| PPLR [8] | None | **41.6** | 85.6 | **91.1** | **93.4** |
| **Ours** | None | 40.8 | **86.2** | 90.5 | 92.8 |

layer to calculate the consine similarity. At the beginning of each epoch, we use DBSCAN [13] for clustering to generate pseudo labels.

The input image is resized 256 x 128 for Market-1501, PersonX and MSMT17 datasets, and 224 x 224 for VeRi-776. For training images, we perform random horizontal flipping, padding with 10 pixels, random cropping, and random erasing [66]. Each mini-batch contains 256 images of 16 pseudo person identities and each person identity contains 16 images. In the case that a person identity has less than 16 images, images are sampled with replacement to compose

Table 3: Our method is more robust to batch size changing.

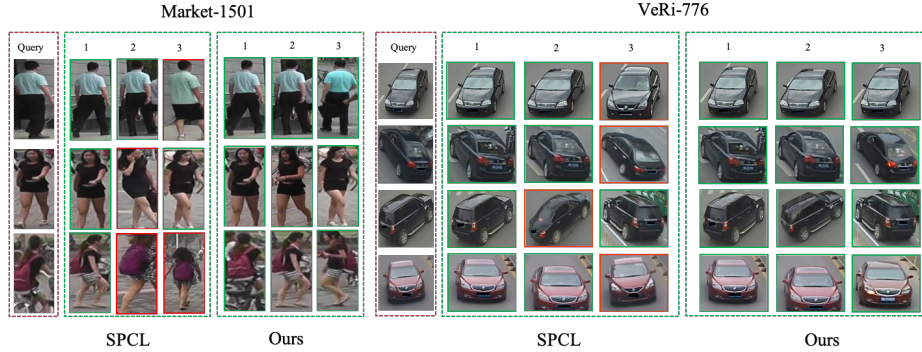| Batch size | 32 | | 64 | | 128 | | 256 | |
|---|---|---|---|---|---|---|---|---|
| | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 |
| Baseline | 62.5 | 80.3 | 65.5 | 82.7 | 71.5 | 86.7 | 73.1 | 87.3 |
| Ours | 79.0 | 90.4 | 80.2 | 91.2 | 81.8 | 92.0 | 83.0 | 92.9 |



Fig. 5: The comparison of top-3 ranking list between SPCL [17] and our method on Market1501 and VeRi776 datasets. The correct results are highlighted by green borders and the incorrect results by red borders.

16 images.The momentum value $m$ is set to 0.1 and the loss temperature $\tau$ is set to 0.05 for all datasets. We adopt Adam optimizer to train the re-ID model with weight decay 5e-4. The initial learning rate is set to 3.5e-4 with a warm-up scheme in the first 10 epochs, and then reduced to 1/10 of its previous value every 20 epoch in a total of 50 epochs.

## 4.2   Comparison with State-of-the-arts

We first compare our method to State-of-the-arts unsupervised learning methods which is the main focus of our method. From Table 2, we can see that our method is significantly better than all existing purely unsupervised methods, which proves the effectiveness of our method. Based on the same pipeline and DBSCAN clustering method, the mAP of our method surpasses the state-of-the-art purely unsupervised learning method by 2.5%, 2.6%, and 12.4% on person re-ID datasets Market-1501 [64], MSMT17 [51], and PersonX [43] dataset. Our method also performs comparable on vehicle re-ID dataset VeRi-776 [33]. And our method performs inferior to SOTA UDA and camera-aware unsupervised re-ID methods as they use additional source labeled dataset and camera id information. The Cluster Contrast can be easily generalized on UDA and camera-aware unsupervised re-ID methods. Details can be found in appendix.

Table 4: (a) We can see that the cluster level memory remarkably improves the performance and the momentum update strategy can further bring the improvement. (b) The performance is superior when a larger fraction of instances are updated on the baseline method. The statistics of both tables are obtained from the Market1501 dataset.

(a) Ablation study of effective components

| Method | mAP | top-1 |
|---|---|---|
| Baseline | 73.1 | 87.5 |
| +Cluster Memory | 80.9 | 91.7 |
| +Cluster Memory+Momentum | 83.0 | 92.9 |

(b) Impact of cluster size

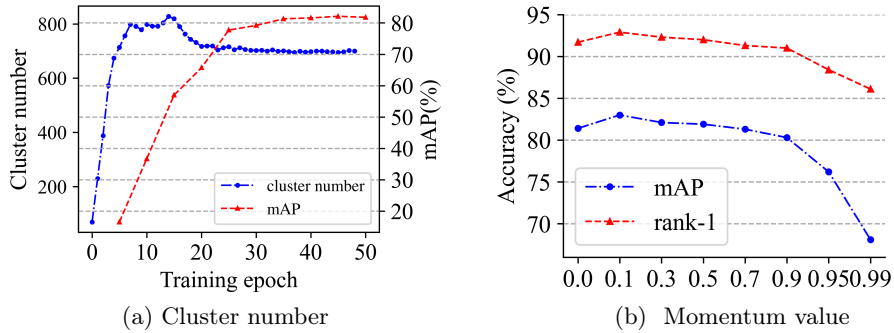| Cluster Size | #Instance | Fraction | mAP | top-1 |
|---|---|---|---|---|
| 20 | 4 | 0.2 | 67.6 | 84.4 |
| 8 | 4 | 0.5 | 71.4 | 85.9 |
| 4 | 4 | 1.0 | **76.5** | **89.0** |



(a) Cluster number



(b) Momentum value

Fig. 6: (a) We can see that the cluster number is dynamically changing as the model trains. (b) Our method performs comparably well when the momentum value is smaller than 0.9. All the statistics are obtained from the Market-1501.

## 4.3   Ablation Studies

In this section, we study the effectiveness of various components in Cluster Contrast method. We define the unsupervised learning pipeline with instance-level memory dictionary (Figure 3 (b)) as the **baseline** method.

*Cluster Memory.*  In section 3.3, we argue that compared to instance-level memory, the cluster-level memory could update cluster feature more consistently. As shown in Figure 3 (b), the instance-level memory maintains the feature of each instance of the dataset. In every training iteration, each instance feature in the mini-batch will be updated to its own memory dictionary. Since the cluster size is unbalancedly distributed, only a small fraction of the instance features could be updated in a large cluster when all instances in a small cluster are updated. Table 4a shows the effectiveness of Cluster Memory without momentum updating, where the instance feature is directly replaced by query feature.

The simplest solution is to increase the batch size similar to SimCLR [5]. As the batch size increases, more instance features could be updated inside one
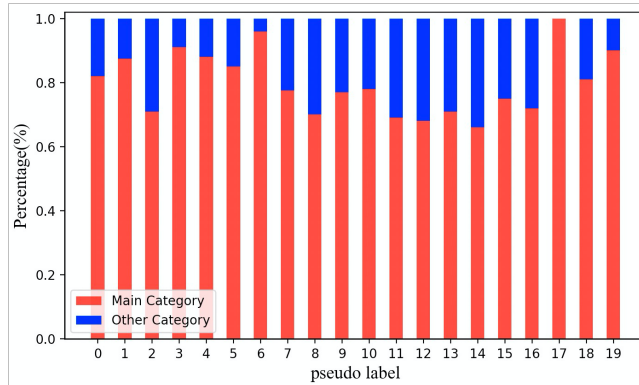
Fig. 7: We randomly select 20 categories from the Market1501 clustering results and calculate the percentage of different categories using ground truth labels.

cluster. However, the batch size reaches its upper limit of 256 due to the GPU memory. To deal with the limitation of the GPU memory, we came up another solution that we restrict the cluster size to a constant number. Therefore, in every iteration a fixed fraction of the instance features could be updated. In this way, the instance feature vectors can be updated consistently with a small batch size. The results in Table 4b demonstrate that the performance of the baseline increases with the rising of the fraction of the updated instance features, until all instance feature vectors inside one cluster could be updated in a single iteration. In sum, we propose the Cluster Contrast, which can update the cluster feature representation in single iteration. As shown in Table 3, our method is more robust to batch size changing. And the Cluster Contrast is more memory efficient since the number of cluster features is an order of magnitude smaller than the number instance features.

*Momentum Updating.* SwAV optimizes the cluster feature end to end by gradient. Our method employs offline clustering so the cluster feature cannot be directly optimized. Instead we follow Moco to use the momentum updating strategy to update cluster representations to maintain the feature consistency. Table 4a shows the effectiveness of the momentum updating strategy. As shown in Eq. 5, the momentum value $m$ controls the update speed of cluster memory. The larger the value of $m$, the slower the cluster memory update. We conducted experiments on the Market-1501 dataset to explore the influence of different $m$ values on our method. As shown in Figure 6 (b), it performs reasonably well when $m$ is less than 0.9. When m is too large (e.g., greater than 0.9), the accuracy drops considerably. These results support us to build better cluster representation.

*Cluster Feature Representation.* As shown in Figure 3 (b), the instance-level memory averages all instance feature vectors to represent the cluster feature. However, in unsupervised learning re-ID, the pseudo label generation stage would

Table 5: Comparison with unsupervised pretrained Resnet50 backbones

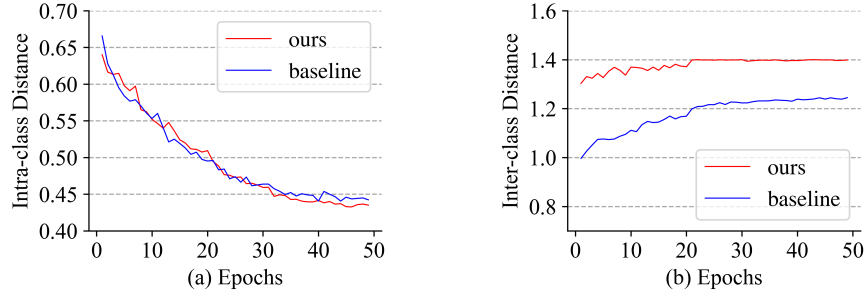| Method | Market-1501 | | MSMT17 | | PersonX | | VeRi-776 | |
|---|---|---|---|---|---|---|---|---|
| | mAP | top-1 | mAP | top-1 | mAP | top-1 | mAP | top-1 |
| supervised | 83.0 | 92.9 | 33.0 | 62.0 | 84.7 | 94.4 | 40.8 | 86.2 |
| SwAV | **84.8** | **93.5** | **38.2** | **67.5** | **86.2** | **95.1** | **42.0** | **87.4** |



Fig. 8: Comparison of the intra-class distance and inter-class distance between our method and baseline method on Market1501 datasets.

inevitably introduce the outlier instances, which are harmful to compute cluster centroid. In Figure 7, we count the proportions of different real categories being clustered into the same category on the Market-1501 dataset. It shows there still around 20% noisy instances when model training in finished. Our method can get better feature representation as shown in Figure 8. The feature quality of our method measured by the intra-class distance and the inter-class distance are much better than the baseline method. From this we can speculate that better representation of features between classes is an important factor for our method to achieve better results.

*Compared with unsupervised pretrained backbones.* The pseudo-label based unsupervised re-ID methods [14, 15, 17, 30, 47, 4, 49, 57, 65, 62] use the supervised ImageNet pretrained backbone for clustering initialization. In order to make the pipeline full unsupervised, we also evaluate our method with unsupervised pretrained backbone in Figure 5. Thanks to the rich feature representation learned by the unsupervised SwAV method, our fully unsupervised re-ID pipeline with unsupervised backbone achieves better results on all four re-ID datasets.

## 5    Conclusion

In this paper, we present the Cluster Contrast for unsupervised re-ID, which stores feature vectors and computes contrast loss in cluster level memory dictionary. It unifies the cluster feature updating progress regardless the cluster size or dataset size. Momentum updating is used to further reinforce the cluster feature consistency. Experiments show demonstrate the effectiveness of our method.

# References

1. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910 (2019)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
3. Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J.: Self-critical attention learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9637–9646 (2019)
4. Chen, H., Lagadec, B., Bremond, F.: Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14960–14969 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
7. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
8. Cho, Y., Kim, W.J., Hong, S., Yoon, S.E.: Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7308–7318 (2022)
9. Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P.: Batch dropblock network for person re-identification and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3691–3701 (2019)
10. Dai, Z., Chen, M., Zhu, S., Tan, P.: Batch feature erasing for person re-identification and beyond. arXiv preprint arXiv:1811.07130 **1**(2),  3 (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 994–1003 (2018)
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
14. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(4), 1–18 (2018)
15. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6112–6121 (2019)
16. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526 (2020)

17. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, h.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 11309–11321. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/821fa74b50ba3f7cba1e6c53e8fa6845-Paper.pdf

18. Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.G., Han, K.: Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3642–3651 (2019)

19. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)

20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

22. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning. pp. 4182–4192. PMLR (2020)

23. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)

24. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)

25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)

26. Isobe, T., Li, D., Tian, L., Chen, W., Shan, Y., Wang, S.: Towards discriminative representation learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8526–8536 (2021)

27. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European conference on computer vision (ECCV). pp. 737–753 (2018)

28. Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person re-identification. IEEE transactions on pattern analysis and machine intelligence **42**(7), 1770–1782 (2019)

29. Lin, S., Li, H., Li, C.T., Kot, A.C.: Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv preprint arXiv:1807.01440 (2018)

30. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8738–8745 (2019)

31. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3390–3399 (2020)

32. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4099–4108 (2018)

33. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2016)
34. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)
35. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020)
36. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
37. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
38. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
39. Riccitiello, J.: John riccitiello sets out to identify the engine of growth for unity technologies (interview). VentureBeat. Interview with Dean Takahashi. Retrieved January **18**(3) (2015)
40. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016)
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
42. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 719–728 (2019)
43. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 608–617 (2019)
44. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
45. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
46. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 365–381 (2018)
47. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10981–10990 (2020)
48. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2275–2284 (2018)
49. Wang, M., Lai, B., Huang, J., Gong, X., Hua, X.S.: Camera-aware proxies for unsupervised person re-identification. In: AAAI. vol. 2, p. 4 (2021)
50. Wang, Z., Zhang, J., Zheng, L., Liu, Y., Sun, Y., Li, Y., Wang, S.: Cycas: Self-supervised cycle association for learning re-identifiable descriptions. arXiv preprint arXiv:2007.07577 (2020)

51. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
52. Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., Li, S.Z.: Unsupervised graph association for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8321–8330 (2019)
53. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
54. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
55. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
56. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3415–3424 (2017)
57. Xuan, S., Zhang, S.: Intra-inter camera similarity for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11926–11935 (2021)
58. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6210–6219 (2019)
59. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2148–2157 (2019)
60. Zeng, K., Ning, M., Wang, Y., Guo, Y.: Hierarchical clustering with hard-batch triplet loss for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13657–13665 (2020)
61. Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y.: Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9021–9030 (2020)
62. Zhang, X., Ge, Y., Qiao, Y., Li, H.: Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3436–3445 (2021)
63. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 667–676 (2019)
64. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
65. Zheng, Y., Tang, S., Teng, G., Ge, Y., Liu, K., Qin, J., Qi, D., Chen, D.: Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8371–8381 (2021)
66. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)

67. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–188 (2018)
68. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 598–607 (2019)
69. Zhou, S., Wang, F., Huang, Z., Wang, J.: Discriminative feature learning with consistent attention regularization for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8040–8049 (2019)
70. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6002–6012 (2019)