A Bottom-up Clustering Approach to Unsupervised Person Re-identification

Yutian Lin[1], Xuanyi Dong[1], Liang Zheng[2],Yan Yan[3], Yi Yang[1]*
[1]CAI, University of Technology Sydney, [2]Australian National University
[3]Department of Computer Science, Texas State University
yutian.lin@student.uts.edu.au, xuanyi.dxy@gmail.com
liangzheng06@gmail.com, y_y34@txstate.edu, yi.yang@uts.edu.au

Abstract

Most person re-identification (re-ID) approaches are based on supervised learning, which requires intensive manual annotation for training data. However, it is not only resource-intensive to acquire identity annotation but also impractical to label the large-scale real-world data. To relieve this problem, we propose a bottom-up clustering (BUC) approach to jointly optimize a convolutional neural network (CNN) and the relationship among the individual samples. Our algorithm considers two fundamental facts in the re-ID task, *i.e.*, ***diversity*** across different identities and ***similarity*** within the same identity. Specifically, our algorithm starts with regarding individual sample as a different identity, which maximizes the diversity over each identity. Then it gradually groups similar samples into one identity, which increases the similarity within each identity. We utilizes a diversity regularization term in the bottom-up clustering procedure to balance the data volume of each cluster. Finally, the model achieves an effective trade-off between the *diversity* and *similarity*. We conduct extensive experiments on the large-scale image and video re-ID datasets, including Market-1501, DukeMTMC-reID, MARS and DukeMTMC-VideoReID. The experimental results demonstrate that our algorithm is not only superior to state-of-the-art unsupervised re-ID approaches, but also performs favorably than competing transfer learning and semi-supervised learning methods.

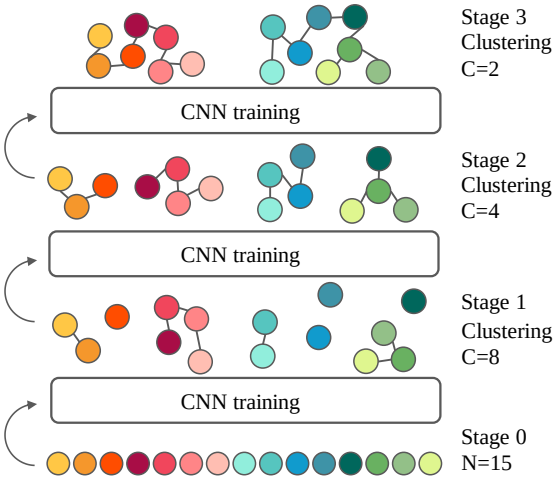

Figure 1: The bottom-up clustering method. Each circle denotes an individual image for training. N denotes the number of training samples, while C denotes the number of clusters after clustering. After the current stage of network training, we apply clustering based on the previous clustering result and the feature similarity of the current stage. By clustering from bottom to up, individual pedestrian samples are gathered to represent an identity.

1 Introduction

Person re-identification (re-ID) aims at matching a target person in a set of query pedestrian images. In recent years, the widespread adoption of deep convolutional neural networks (CNN) has led to impressive progress in the field of re-ID (Yi et al. 2014; Li et al. 2014; Varior, Haloi, and Wang 2016). However, supervised re-ID methods require intensive manual labeling. It is expensive and not applicable to the real-world applications. The limited generalization ability motivates the research into unsupervised approaches for person re-ID.

Traditional unsupervised methods focus on hand-crafted features (Farenzena et al. 2010; Liao et al. 2015; Lisanti et al. 2015), salience analysis (Zhao, Ouyang, and Wang 2013; Wang, Gong, and Xiang 2014) and dictionary learning (Kodirov, Xiang, and Gong 2015). These methods produce much lower performance than supervised methods and are not applicable to large-scale real-world data. In recent years, some transfer learning methods (Hehe et al. 2018; Peng et al. 2016; Deng et al. 2018) are proposed upon the success of deep learning (Dong et al. 2018a; Li, Tang, and Mei 2018). These methods usually learn an identity-discriminative feature embedding on the source dataset, and transfer the learned features to the unseen target domain. However, these methods require a large amount of annotated source data, which cannot be regarded as pure unsupervised approaches.

Previous deep learning based "unsupervised" person re-ID approaches leverage the prior knowledge learned from other re-ID datasets. However, we aim to solve the problem

*Corresponding author.

in a more challenge and practical setting, *i.e.*, without any re-ID annotation. To learn discriminate features in this difficult condition, we propose a novel Bottom-Up Clustering method (**BUC**) for unsupervised re-ID that maximizes the diversity over the identities while maintaining the similarity within each identity. As illustrated in Fig. 1, during the training process, the individual samples are gathered into clusters, and the clusters will merge gradually. Specifically, our framework BUC applies network training and the bottom-up clustering in an iterative way. We propose the repelled loss that can optimize the network without actually having any label and obtain decent initial accuracy. At the beginning of network training, we view individual images as exemplars, *i.e.*, each image belongs to a distinct cluster. We then gradually incorporate similarity within identities by a bottom-up clustering, which is to merge similar images (clusters) into one cluster. Moreover, in practice, different identities should have a similar probability to be captured by cameras, and thus the image number for different clusters should be balanced. To enforce this property, we incorporate a diversity regularization term in the merging procedure. Finally, during the bottom to up clustering procedure, our framework exploits the similarity and the diversity to learn discriminative features.

Our contributions can be summarized in four-fold:

- We propose a bottom-up clustering framework to solve the unsupervised re-ID problem. By exploiting the intrinsic diversity among identities and similarity within each identity, our framework can learn robust and discriminative features.
- We adopt the repelled loss to optimize the model without labels. The repelled loss directly optimizes the cosine distance among each individual sample / cluster. It can facilitate the model to exploit the similarity within each cluster and maximize the diversity among each identity.
- We propose a diversity regularization term to enable the balanced image number in each cluster. It makes the clustering results align with the real world distribution.
- The experimental results demonstrate that our approach is superior to the state-of-the-art methods on both image-based and video-based re-ID datasets. We achieves top-1 accuracy of 61.9% on Market-1501 (Zheng et al. 2015) and 55.1% on MARS (Zheng et al. 2016). Moreover, the one-shot re-ID methods utilize more annotation than ours, whereas our approach also obtains a higher performance than them.

2 Related Work

Most re-ID methods are in a supervised manner, in which sufficient labeled person pairs across cameras are given. They mainly focus on designing feature representations (Zhao, Ouyang, and Wang 2014) or learning robust distance metrics (Zheng, Gong, and Xiang 2011; Liao et al. 2015). Recently, deep learning methods achieve great success (Li et al. 2014; Varior, Haloi, and Wang 2016; Zheng et al. 2015; Zheng, Zheng, and Yang 2017) by simultaneously learning the image representations and similarities. In this paper, we focus on the unsupervised setting and do not discuss more supervised methods here.

2.1 Unsupervised Person Re-identification

The existing fully unsupervised methods usually fall into three categories, designing hand-craft features (Farenzena et al. 2010; Liao et al. 2015; Lisanti et al. 2015), exploiting localized salience statistics (Zhao, Ouyang, and Wang 2013; Wang, Gong, and Xiang 2014) or dictionary learning based methods (Kodirov, Xiang, and Gong 2015; Yan et al. 2018). However, it is a challenging task to design suitable features for images captured by different cameras, under different illumination and view condition. These methods are unable to explicitly exploit the cross-view discriminative information without pairwise identity labels. Thus the performance of these methods is much weaker than supervised methods. Recently, Xiao *et al.* (Xiao et al. 2017) propose the OIM loss for semi-supervised person search. It can also be used for unsupervised re-ID. Compared to OIM, our BUC has three advantages. (1) We constrain the feature to distribute on a unit sphere to improve its robustness. (2) We design the cluster merging to exploit the similarity among identities. (3) We propose a diversity regularization term to avoid the model collapse.

Another category of the unsupervised method makes use of additional information (Dong et al. 2018b; Peng et al. 2016; Deng et al. 2018; Li and Tang 2015). Recently, cross-domain transfer learning is used in the unsupervised re-ID task (Hehe et al. 2018; Wang et al. 2018), where information from an external source dataset is utilized. Fan *et al.* (Hehe et al. 2018) propose a progressive method, where the K-means clustering and the IDE (Zheng et al. 2015) network pre-trained on the source dataset are updated iteratively. Wang *et al.* (Wang et al. 2018) propose to learn an attribute-semantic and identity discriminative representation from the source dataset, which is transferable to the target domain. There are also some recent works (Liu, Wang, and Lu 2017; Ye et al. 2017; Ye, Lan, and Yuen 2018) focusing on the unsupervised video-based re-ID. However, these methods require some very useful annotations of the dataset, *i.e.*, the total number of identities and their appearance. To conduct experiments, they annotate each identity with a labeled video tracklet, which only reduces part of the annotation workload. As discussed in (Wu et al. 2018a), these approaches are actually the one-example methods. Different from these methods, our work focuses on the fully unsupervised setting in which there is ***no annotation*** on the dataset.

2.2 Unsupervised Feature Learning

Unsupervised feature learning is widely studied in many tasks, such as image recognition, image classification, and image retrieval (Tang and Liu 2016). Some works use hand-crafted features combined with conventional clustering methods (Han and Kim 2015; Hariharan, Malik, and Ramanan 2012; Singh, Gupta, and Efros 2012). However, the hand-designed features are not as effective as deeply learned features. A number of works (Dosovitskiy et al. 2014; Bautista et al. 2016) sample patches from images and generate labels for the patches as supervision. In (Dosovitskiy et

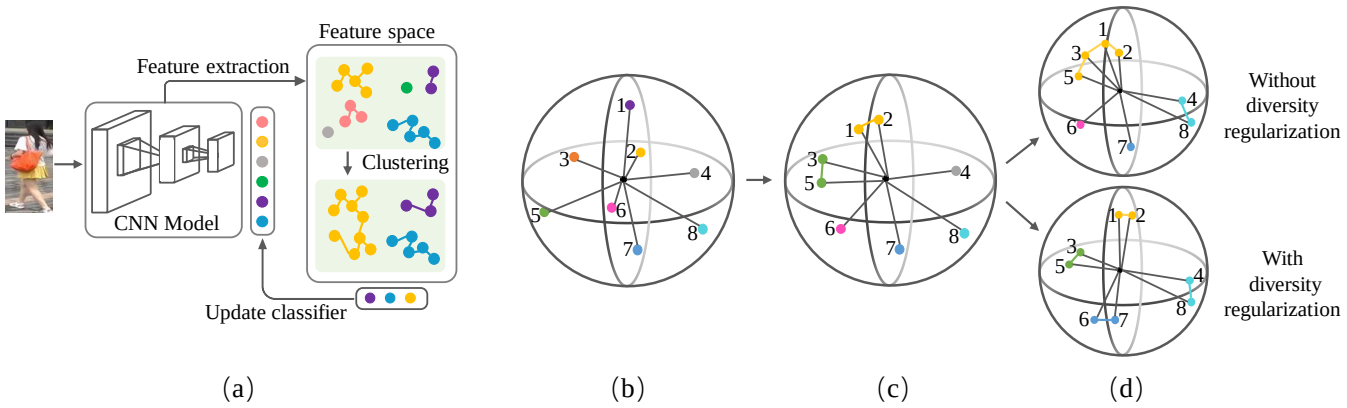

Figure 2: (a). The proposed framework takes unlabeled images as input to train the network and extract the image features for clustering. The framework do three steps alternatively, *i.e.*, extracts the CNN feature for each image, merges clusters over the whole training set, and re-train the CNN model. Fig. (b)-(d) depict the cluster merging procedure. In this example, each step we merge two cluster pairs with minimum dissimilarity. The solid points with the same color represent images in the same cluster. The colored line indicates the connected two clusters have been merged into one. In (b), the learned features discriminatively span a unit sphere, which the diversity is maximized. In (c), after merging the clusters, feature embeddings of the same cluster get closer in the sphere. In (d), the upper sphere shows the cluster merging result without diversity regularization: (Point 1, Point 3) and (Point 4, Point 8) have the shortest distances, and are then merged into one cluster. The lower sphere shows the cluster merging result with diversity regularization: though the distance between the yellow and green clusters is the shortest, these two clusters are too large and should not be merged. The Point 6 and Point 7 are merged instead.

al. 2014), exemplar-CNN is proposed to discriminate among a set of surrogate classes, where the surrogate classes are formed by applying a variety of transformations to randomly sampled image patches.

Wu *et al.* (Wu et al. 2018b) propose a non-parametric softmax classifier and use noise-contrastive estimation to tackle the computational challenges. Different from these works, our BUC not only considers the diversity over each sample but also exploits the similarity within each class. Comparing with these unsupervised feature learning methods on the classification task, our BUC obtains superior performance.

3 Methodology

3.1 Preliminary

Given a training set $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ of N images, our goal is to learn a feature embedding function $\phi(\theta; x_i)$ from $\mathbf{X}$ without any manual annotation, where parameters of ϕ are collectively denoted as θ. This feature embedding function can be applied to the testing set, $\mathbf{X}^t = \{x_1^t, x_2^t, ...x_{N_t}^t\}$ of N_t images, and the query set $\mathbf{X}^q = \{x_1^q, x_2^q, ...x_{N_q}^q\}$ of N_q images. During the evaluation, we use the feature of a query image $\phi(\boldsymbol{\theta}; x_i^q)$ to search the similar image features from the testing set. The query result is a ranking list of all testing images according to the Euclidean distance between the feature embedding of the query and testing data, *i.e.*, $d(x_i^q, x_i^t) = \|\phi(\boldsymbol{\theta}; x_i^q) - \phi(\boldsymbol{\theta}; x_i^t)\|$. The feature embeddings are supposed to assign a higher rank to similar images and keep the images of a different person a low rank.

To learn the feature embedding, traditional methods usually learn the parameters with manual annotations. That is, each image x_i is associated with a label y_i, where $1 \leq y_i \leq k$ and k is the number of identities. A classifier $f(\boldsymbol{w}; \phi(\theta; x_i)) \in \mathbb{R}^k$ parameterized by $\boldsymbol{w}$ is used to predict the identity of the image x_i. The classifier parameter $\boldsymbol{w}$ and the embedding parameter θ are jointly optimized by the following objective function:

$$\min_{\boldsymbol{\theta}, \boldsymbol{w}} \sum_{i=1}^{N} \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{1}$$

where ℓ is the softmax cross entropy loss. However, y_i is not available in the unsupervised setting, and it is challenging to find another objective function that can learn a robust embedding function ϕ.

3.2 The Bottom-up Clustering Framework

Without the manual annotation, it is important to design a supervision signal that can be used to train CNN models. To achieve this goal, we aim to exploit the similarity and diversity properties from the training data as the supervision information. As shown in Fig. 2 (a), the framework mainly contains two components: (i) A network trained with a repelled loss to let the cluster centers repelled by each other. (ii) A clustering procedure in the feature embeddings space to merge existing clusters. The clustering and network updating is done iteratively.

Network with Repelled Loss. Since we do not have ground truth labels, we assign each image to a different cluster initially, *i.e.*, $\{\hat{y}_i = i \mid 1 \leq i \leq N\}$.[1] In this way, the network learns to recognize each training sample instead of

[1] $\hat{y}_i$ is the cluster index for x_i and is dynamically changed.

the identities, and the diversity over each training sample is maximized. We then gradually incorporate similarity within identities by grouping similar images into clusters. The cluster ID is used as the training label, and the network is trained to minimize total intra-cluster variance and maximize the inter-cluster variance.

We define the probability that image x belongs to the c-th cluster as,

$$p(c|x, \boldsymbol{V}) = \frac{exp(\boldsymbol{V}_c^{\mathrm{T}} \boldsymbol{v}/\tau)}{\sum_{j=1}^{C} exp(\boldsymbol{V}_j^{\mathrm{T}} \boldsymbol{v}/\tau)}, \tag{2}$$

where $\boldsymbol{v} = \frac{\phi(\boldsymbol{\theta};x)}{||\phi(\boldsymbol{\theta};x)||}$, $\boldsymbol{V} \in \mathbb{R}^{C \times n_\phi}$ is a lookup table that stores the feature of each cluster, $\boldsymbol{V}_j$ is the j-th colum of $\boldsymbol{V}$, and C is the number of clusters at the current stage. At the first training stage, $C = |\boldsymbol{X}| = N$. At the following stages, our approach will merge similar images into one class, and C will gradually decrease. τ is a temperature parameter (Hinton, Vinyals, and Dean 2014) that controls the softness of probability distribution over classes. Following (Xiao et al. 2017), we set $\tau = 0.1$ in this paper. In the forward operation, we compute cosine similarities between data x_i and all the other data by $\boldsymbol{V}^T \cdot \boldsymbol{v}_i$. During backward, we update the $\hat{y}_i$-th column of the table $\boldsymbol{V}$ by $\boldsymbol{V}_{\hat{y}_i} \leftarrow \frac{1}{2}(\boldsymbol{V}_{\hat{y}_i} + \boldsymbol{v}_i)$. Finally, we minimize the repelled loss, which is formulated as,

$$\mathcal{L} = -\log(p(\hat{y}_i|x_i, \boldsymbol{V})). \tag{3}$$

During the optimization, $\boldsymbol{V}_j$ will contain the information of all images within the j-th cluster. It can be considered as a kind of "centroid" of this cluster. We do not directly calculate the centroid feature in each training stage due to the high time complexity. The lookup table V can avoid exhaustive computation of extracting features from all data at each training step. The proposed objective has two advantages. First, it can maximize the cosine distance between each image feature $\boldsymbol{v}_i$ and each centroid features $\boldsymbol{V}_{j \neq \hat{y}_i}$. Second, it can minimize the cosine distance between each image feature $\boldsymbol{v}_i$ and the corresponding centroid feature $\boldsymbol{V}_{j=\hat{y}_i}$. With these two advantages, our approach can trade off the similarity and diversity over the whole training set.

Cluster Merging. After the first training stage, the training samples are prone to be away from each other in the learned feature space. However, images of the same identity are usually visually similar and should be close, which we call *similarity*. To exploit the similarity, we apply the hierarchical clustering on the CNN features to merge the images from bottom to up. In the start, each image is treated as a cluster. Then pairs of clusters are merged into one by measuring their similarity. In order to decide which clusters should be merged, we consider the minimum distance criterion to calculate the dissimilarity value $D(A, B)$ between cluster A and cluster B.

The **minimum distance criterion** takes the shortest distance between images in two clusters as dissimilarity. This criterion only considers the shortest distance: if two images in the cluster look really alike, the clusters tend to be merged, no matter how dissimilar other images look. The advantage is that images of the same identity under the same camera are visually alike and tend to be merged into one cluster under this criterion, which guarantees the accuracy of merged images. It is formulated as:

$$D_{distance}(A, B) = \min_{x_a \in A, x_b \in B} d(x_a, x_b), \tag{4}$$

where $d(x_a, x_b)$ is defined as the Euclidean distance between the feature embeddings of two images, *i.e.*, $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$. Specifically, $d(x_a, x_b) = \|\boldsymbol{v}_a - \boldsymbol{v}_b\|$.

As shown in Fig. 2 (b)-(d), at each merging step, we aim to reduce m clusters. We define $m = N \times mp$, where $mp \in (0, 1)$ denotes the speed of cluster merging. Each time, the clusters with the shortest distance are merged. The number of clusters is initialed as $C = N$, *i.e.*, the number of training samples. After t times of cluster merging, the number of clusters is dynamically decreased to $C = N - t \times m$.

There are other criteria methods to measure the dissimilarity. (1) The **maximum distance criterion** takes the maximum distance between elements of each cluster as the dissimilarity. However, images of the same identity under different cameras may have totally different visual appearance. This strategy fails to merge images from different cameras. (2) The **centroid distance criterion** takes the distance between mean features of elements in each cluster as the dissimilarity. In the re-ID task, images come from different cameras, which have different illumination, pose, and viewpoint. The mean operation omits the diversity among images within one cluster, therefore, it overlooks the useful camera information. In experiments, we demonstrate the minimum criterion is the best, and will discuss later in Section 4.4.

Dynamic Network Updating. The framework iteratively trains the network and merges the learned image features clusters. The clustering results are then fed to the network for further updating. The whole updating process is described in Algorithm 1. The number of clusters is initialized

Algorithm 1 The Bottom-Up Clustering (BUC) Framework

Require: Unlabeled data $X = \{x_1, x_2, ...x_N\}$;
Merge percent $mp \in (0, 1)$;
CNN model $\phi(\cdot; \boldsymbol{\theta}_0)$.
Ensure: Best CNN model $\phi(\cdot; \boldsymbol{\theta}^*)$.
Initialize: Cluster label $Y = \{\hat{y}_i | \ 1 \leq i \leq N\}$
  Number of cluster $C = N$
  Number of merging image $m = \lceil mp * C \rceil$
while $C > m$ **do**
  Train CNN model $\phi(x; \boldsymbol{\theta})$ with X and Y
  Clustering with m:
    $C \leftarrow C - m$
    Update Y with the new cluster labels
  Initialize the lookup table $\boldsymbol{V}$ with new dimensions
  Re-train the CNN model with parameters $\boldsymbol{\theta}$
  Evaluate on the validation set $\rightarrow$ performance P
  if $P > P^*$ **then**
    $P^* = P_t$
    Best model = $\phi(x; \boldsymbol{\theta})$
  end if
end while

Table 1: Comparison with the state-of-the-art methods on two image-based large-scale re-ID datasets, *i.e.*, the Market-1501 dataset and the DukeMTMC-reID dataset. The column "Labels" lists the labels utilized by the method. "Transfer" denotes the information from another re-ID dataset with full annotations. "OneEx" denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us.

Methods	Venue	Labels	Market-1501				DukeMTMC-reID			
			rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
BOW (Zheng et al. 2015)	ICCV15	**None**	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
OIM* (Xiao et al. 2017)	CVPR18	**None**	38.0	58.0	66.3	14.0	24.5	38.8	46.0	11.3
UMDL (Peng et al. 2016)	CVPR16	Transfer	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PUL (Hehe et al. 2018)	TOMM18	Transfer	44.7	59.1	65.6	20.1	30.4	46.4	50.7	16.4
EUG* (Wu et al. 2018a)	CVPR18	OneEx	49.8	66.4	72.7	22.5	45.2	59.2	63.4	24.5
SPGAN (Deng et al. 2018)	CVPR18	Transfer	58.1	**76.0**	**82.7**	26.7	**46.9**	**62.6**	**68.5**	**26.4**
TJ-AIDL (Wang et al. 2018)	CVPR18	Transfer	58.2	-	-	26.5	44.3	-	-	23.0
BUC without diversity regularizer	AAAI19	**None**	54.5	68.6	74.6	24.1	38.8	50.5	55.1	20.6
BUC	AAAI19	**None**	**61.9**	73.5	78.2	**29.6**	40.4	52.5	58.2	22.1

as the number of training images. After each cluster merging, the labels of the training images are re-assigned as the new cluster ID. The memory layer of the optimizer is re-initialized to zero to avoid getting stuck in local optima. We constantly train the network until we observe a performance drop on the validation set.

3.3 Diversity Regularization

With the clusters being merged, the number of classes is decreasing, and the number of images in the clusters is increasing. Although we do not know the exact number of images in each identity, we can assume that the images are evenly distributed to the identities, and different identities should be scattered in different clusters, which we call *diversity*. This implies that one cluster should not contain much more images compared to other clusters. To avoid one cluster being redundant and boost the small clusters to merge together, we incorporate a diversity regularization term into the distance criterion.

$$D_{diversity}(A, B) = |A| + |B|, \tag{5}$$

where $|A|$ denotes the number of samples belonging to the cluster A. Then, the final dissimilarity is calculated as:

$$D(A, B) = D_{distance}(A, B) + \lambda D_{diversity}(A, B), \tag{6}$$

where λ is a parameter that balances the impact of distance and regularization. The reason for adding a diversity regularization term is that, there exist some visually similar identities wearing almost the same clothes. Without the regularization term, the algorithm might merge these similar but different identities into one tremendous cluster by mistake. We tend to merge small clusters, unless the distance $d(x_a, x_b)$ is small enough. This procedure is illustrated in Fig. 2 (d).

4 Experimental Results

4.1 Datasets

Market-1501 (Zheng et al. 2015) is a large-scale dataset for person re-ID captured by 6 cameras in a university campus. It contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing.

DukeMTMC-reID (Zheng, Zheng, and Yang 2015) is a large-scale re-ID dataset derived from the DukeMTMC dataset (Ristani et al. 2016). It contains 16,522 images of 702 identities for training, 2,228 images of the other 702 identities for query, and 17,661 gallery images.

MARS (Zheng et al. 2016) is a large-scale video-based dataset for person re-ID captured by six cameras in a university campus. The dataset contains 17,503 video tracklets of 1,261 identities, where 625 identities are used for training and 636 identities are used for testing.

DukeMTMC-VideoReID (Wu et al. 2018a) is a large-scale video-based re-ID dataset derived from the DukeMTMC dataset (Ristani et al. 2016). It contains 2,196 tracklets of 702 identities for training, 2,636 tracklets of other 702 identities for testing.

4.2 Experimental Settings

Evaluation Protocols. For the image-based re-ID datasets Market-1501 and DukeMTMC-reID, we take all the training image without ID labels to train the framework. For the video-based datasets MARS and DukeMTMC-VideoReID, each training tracklet is regarded as an individual sample in the model training. Note, our method does not utilize any annotation information (*e.g.* ID labels or other annotated datasets) for model initialization or training.

Evaluation Metrics. For person re-ID, we use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the performance of each method. For each query, its average precision (AP) is computed from its precision-recall curve. The mAP is calculated as the mean value of average precision across all queries. We report the Rank-1, Rank-5, Rank-10 scores to represent the CMC curve. These CMC scores reflect the retrieval precision, while mAP reflects the recall.

Implementation Details. We adopt ResNet-50 as the CNN backbone to conduct all the experiments. We initialize it by the ImageNet (Krizhevsky, Sutskever, and Hinton 2012) pre-trained model with the last classification layer removed. For all the experiments if not specified, we set the number of training epochs in the first stage to be 20, the batch size to be 16, the dropout rate to be 0.5, mp to be 0.05 and λ in Eq. (6) to be 0.003. We use stochastic gradient descent with a momentum of 0.9 to optimize the model. The learning rate is initialized to 0.1 and changed to 0.01 after 15

Table 2: Comparison with the state-of-the-art methods on two video-based re-ID datasets, MARS and DukeMTMC-VideoReID. The column "Labels" lists the labels utilized by the method. "OneEx" denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us.

Methods	Venue	Labels	MARS				DukeMTMC-VideoReID			
			rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
OIM* (Xiao et al. 2017)	CVPR18	**None**	33.7	48.1	54.8	13.5	51.1	70.5	76.2	43.8
DGM+IDE (Ye et al. 2017)	ICCV17	OneEx	36.8	54.0	-	16.8	42.3	57.9	69.3	33.6
Stepwise (Liu, Wang, and Lu 2017)	ICCV17	OneEx	41.2	55.5	-	19.6	56.2	70.3	79.2	46.7
RACE (Ye, Lan, and Yuen 2018)	ECCV18	OneEx	43.2	57.1	62.1	24.5	-	-	-	-
DAL (Chen, Zhu, and Gong 2018)	BMVC18	Camera	49.3	65.9	72.2	23.0	-	-	-	-
EUG (Wu et al. 2018a)	CVPR18	OneEx	62.6	74.9	-	42.4	72.7	84.1	-	63.2
BUC without diversity regularizer	AAAI19	**None**	53.8	67.9	72.5	28.9	72.6	84.9	88.9	64.4
BUC	AAAI19	**None**	55.1	68.3	71.8	29.4	**74.8**	**86.8**	**89.7**	**66.7**

epochs. For video-based datasets, we take the average feature of all frames within a tracklet to be the tracklet feature for cluster merging and final evaluation. On Market-1501 and DukeMTMC-reID, it takes about 4 hours to finish the training procedure with a GTX 1080TI GPU. On Mars and DukeMTMC-VideoReID, it takes about 5 hours.

4.3 Comparison with the State of the Art

Image-based Person Re-identification. The comparisons with the state-of-the-art algorithms on image-based datasets are shown in Table 1. Note that the performances in (Peng et al. 2016) are reproduced by (Hehe et al. 2018) and we borrow the numbers to our table. On Market-1501, we obtain the best performance among the compared methods with **rank-1 = 61.9%, mAP = 29.6%**. Compared to the state-of-the-art method OIM (Xiao et al. 2017) in the fully unsupervised setting, we achieve 23.9 points (absolute) and 15.6 points improvement in rank-1 accuracy and mAP, respectively. Similarly, our method achieves 15.5 points (absolute) and 10.8 points improvement in rank-1 and mAP on DukeMTMC-reID. The significant improvement is mainly due to the further cluster merging that exploits similarity from the instances for supervision.

We also compare our method to the state-of-the-art transfer learning methods in Table 1. Although these methods utilize external images and human annotations, our method with *zero* annotation still surpasses them by a large margin. On Market-1501, our method outperforms the state-of-the-art transfer learning method (Wang et al. 2018) by 3.7 points and 3.1 points in rank-1 accuracy and mAP, respectively.

Video-based Person Re-identification. Table 2 shows the comparisons with the state-of-the-art algorithms on video-based datasets. On MARS, we obtain **rank-1 = 55.1%, mAP = 29.4%**. We beat the fully-unsupervised method OIM (Xiao et al. 2017) by a large margins with 21.4 points in rank-1 accuracy and 15.9 points for mAP. On DukeMTMC-VideoReID, our results achieve 23.7 points and 22.9 points improvement on rank-1 accuracy and mAP, respectively.

In Table 2, we also compare our method to the state-of-the-art methods (Liu, Wang, and Lu 2017; Ye et al. 2017; Ye, Lan, and Yuen 2018) in the video-based one-example setting. These methods initialize their models by annotating each person with a labeled video tracklet. As discussed in (Wu et al. 2018a), these approaches are the one-example methods, hence, are not actually unsupervised. Their methods rely on some very useful annotations on the dataset, *i.e.*, how many identities exist in the dataset and what they look like (from a tracklet for each person). Without any annotation, our method still beats all of these methods with one-example annotation, which indicates that our method is more effective in exploiting the unlabeled data.

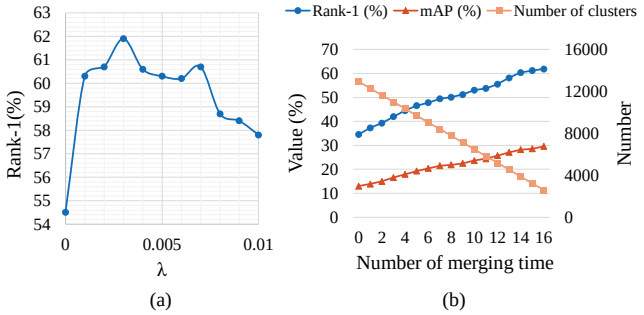

Figure 3: (a) Performance curve with different values of the diversity regularization parameter λ on Market-1501. (b) The rank-1 accuracy, mAP, and the number of clusters on Market-1501 after each cluster merging step.

4.4 Ablation Studies

The Impact of Diversity Regularization. The performance of with and without the diversity regularization item is shown in Table 1 and Table 2, respectively. The diversity regularization provides a large performance improvement on all the four datasets. Specifically, on Market-1501 and DukeMTMC-reID, the diversity regularization item improves the rank-1 accuracy by 7.4 points and 1.6 points, respectively. We suspect that without the diversity regularization, two similar identities may be easily merged into one cluster by mistake. With the diversity regularization term, we tend to merge small clusters first.

The diversity regularization parameter λ in Eq. (6) balances the cluster size and cluster distance. We evaluate different values for the parameter λ in Fig. 3 (a). As λ increases from 0 to 0.003, the rank-1 accuracy on Market-1501 increases from 54.5% to 61.9%. If we set λ to be greater than

Table 3: The comparison of different merging criteria on Market-1501.

Criterion	rank-1	rank-5	rank-10	rank-20	mAP
Maximum	58.6	69.2	72.3	77.1	25.0
Centroid	61.4	72.8	76.3	81.0	28.3
Minimum	**61.9**	**73.5**	**78.2**	**82.0**	**29.6**

0.003, the too large diversity regularization term would begin to introduce a negative effect.

The Impact of Cluster Merging Criterion. As shown in Table 3, the results of three cluster merging criteria are listed. We get the best result with the rank-1 = 61.9% when using the minimum distance criterion. When using the centroid distance criterion, we observe a slightly lower performance with the rank-1 = 61.4%. When using the maximum distance criterion, we observe a rank-1 accuracy of 58.6%. We assume that images of the same identity from different cameras suffer from large visual appearance difference. Using this criterion may fail to merge clusters including images captured from different cameras.

4.5 Algorithm Analysis

Analysis over Cluster Merging. We show the performance of re-ID and the number of remaining clusters on Market-1501 in Fig. 3 (b). As the number of the remaining clusters gradually decreases, the rank-1 accuracy and the mAP accuracy are both increasing. After 16 times of merging, the rank-1 accuracy increases from 34.6% to 61.9%, and the mAP accuracy increases from 13.0% to 29.6%. The number of clusters is decreased from 12,936 to 2,600, while the ground truth number of identities is 751. We observe that both the improvement of the performance and the reduction of the clusters are continuous and gradual. It indicates that our method gradually learns from the diversified images to generate a more discriminative feature representation.

Qualitative Analysis. To further understand the discriminative ability of our unsupervised learned feature, we utilize t-SNE (Maaten and Hinton 2008) to visualize the feature embeddings of the merged clusters by plotting them to the 2-dimension map. As illustrated in Fig. 4, the images of the same identity usually gather together, which represents the learned similarity within identities. Besides, most identities are distinguishable from each other, which represents the diversity among the identities. More qualitative results over iterations can be found in the supplementary material.

4.6 Compare to Unsupervised Feature Learning

To compare with the unsupervised feature learning methods, we also conduct image classification experiments on CIFAR-10 (Krizhevsky and Hinton 2009) to make a fair comparision with them. CIFAR-10 contains 60,000 images of 10 different classes. Following (Wu et al. 2018b), we take ResNet18 as the backbone model and extract the last pooling layer's features. The nearest neighbor classifier is adopted to assess the learned feature, which reflects the quality of the representation. As shown in Table 4, we achieve 85.2% top-1 accuracy on CIFAR-10, showing a 4.4% accuracy gain over (Wu et al. 2018b). This improvement proves the superiority of the cluster merging and network updating strategies.

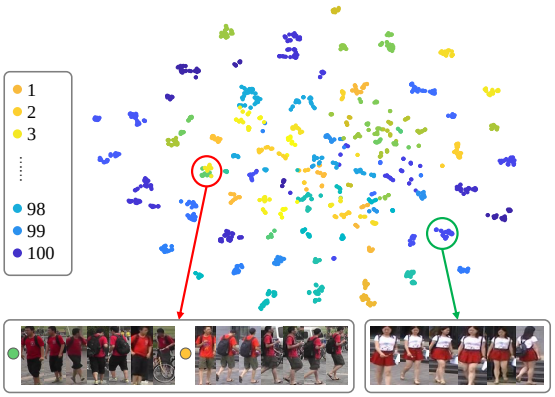

Figure 4: T-SNE visualization of the learned feature embeddings on a part of the Market-1501 training set (100 identities, 1,700 images). Points of the same color represent images of the same identity. We show the detailed images of a positive example (the green circle) and a negative example (the red circle). The points in the green circle are of the same identity. In the red circle, the green and yellow points gathered together, indicating that our algorithm merges them into one cluster by mistake. However, the samples looks very similar and are hard to be discriminated between each other.

Table 4: The top-1 accuracy on CIFAR-10.

Methods	top-1
S-CNN (Ghaderi and Athitsos 2016)	72.7
NID (Wu et al. 2018b)	80.8
Roto-Scat + SVM (Singh and Kingsbury 2017)	82.3
DCGAN (Radford, Metz, and Chintala 2016)	82.8
Ours	**85.2**

5 Conclusions

In this paper, we propose a bottom-up clustering approach (BUC) to tackle the unsupervised re-ID task. It jointly optimizes a CNN model and the relationship among the individual samples. Specifically, the network training starts by treating each individual image as an identity. Then, bottom-up clustering is applied to the feature embedding extracted from the network to reduce the number of classes. During the whole process, the network gradually exploits similarity from diverse unlabeled images. In experiments, BUC achieves higher performance than the state-of-the-art methods in both image-based and video-based re-ID datasets.

Acknowledgment. We acknowledge the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centers Programme for funding this research. We also acknowledge the gift donation from Cisco, Inc for this research.

References

Bautista, M. A.; Sanakoyeu, A.; Tikhoncheva, E.; and Ommer, B. 2016. Cliquecnn: Deep unsupervised exemplar learning. In *NIPS*.

Chen, Y.; Zhu, X.; and Gong, S. 2018. Deep association learning for unsupervised video person re-identification. In *BMVC*.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*.

Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018a. Style aggregated network for facial landmark detection. In *CVPR*.

Dong, X.; Yu, S.-I.; Weng, X.; Wei, S.-E.; Yang, Y.; and Sheikh, Y. 2018b. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Ghaderi, A., and Athitsos, V. 2016. Selective unsupervised feature learning with convolutional neural network (s-cnn). In *ICPR*.

Han, D., and Kim, J. 2015. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*.

Hariharan, B.; Malik, J.; and Ramanan, D. 2012. Discriminative decorrelation for clustering and classification. In *ECCV*.

Hehe, F.; Liang, Z.; Chenggang, Y.; and Yi, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *NIPS-W*.

Kodirov, E.; Xiang, T.; and Gong, S. 2015. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Li, Z., and Tang, J. 2015. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Transactions on Image Processing*.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*.

Li, Z.; Tang, J.; and Mei, T. 2018. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Lisanti, G.; Masi, I.; Bagdanov, A. D.; and Del Bimbo, A. 2015. Person re-identification by iterative re-weighted sparse ranking. *IEEE T-PAMI*.

Liu, Z.; Wang, D.; and Lu, H. 2017. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research*.

Peng, P.; Xiang, T.; Wang, Y.; Pontil, M.; Gong, S.; Huang, T.; and Tian, Y. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

Singh, A., and Kingsbury, N. 2017. Dual-tree wavelet scattering network with parametric log transformation for object classification. In *ICASSP*.

Singh, S.; Gupta, A.; and Efros, A. A. 2012. Unsupervised discovery of mid-level discriminative patches. In *ECCV*.

Tang, H., and Liu, H. 2016. A novel feature matching strategy for large scale image retrieval. In *IJCAI*.

Varior, R. R.; Haloi, M.; and Wang, G. 2016. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*.

Wang, H.; Gong, S.; and Xiang, T. 2014. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018a. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

Wu, Z.; Xiong, Y.; Stella, X. Y.; and Lin, D. 2018b. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *CVPR*.

Yan, C.; Luo, M.; Liu, W.; and Zheng, Q. 2018. Robust dictionary learning with graph regularization for unsupervised person re-identification. *Multimedia Tools and Applications*.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. *ICCV*.

Ye, M.; Lan, X.; and Yuen, P. C. 2018. Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *ICPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*.

Zheng, W.-S.; Gong, S.; and Xiang, T. 2011. Person re-identification by probabilistic relative distance comparison. In *CVPR*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2015. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*.