Unsupervised Person Re-identification via Softened Similarity Learning

Yutian Lin[1]*, Lingxi Xie[2], Yu Wu[3,4], Chenggang Yan[1], Qi Tian[2]†
[1]Hangzhou Dianzi University, [2]Huawei Inc., [3]Baidu Research,
[4]ReLER, University of Technology Sydney
yutianlin477@gmail.com, 198808xc@gmail.com, yu.wu-3@student.uts.edu.au,
cgyan@hdu.edu.cn, tian.qi1@huawei.com

Abstract

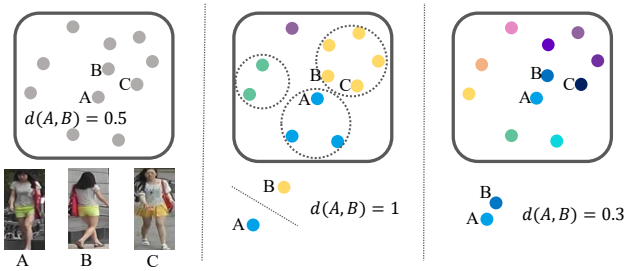
Person re-identification (re-ID) is an important topic in computer vision. This paper studies the unsupervised setting of re-ID, which does not require any labeled information and thus is freely deployed to new scenarios. There are very few studies under this setting, and one of the best approach till now used iterative clustering and classification, so that unlabeled images are clustered into pseudo classes for a classifier to get trained, and the updated features are used for clustering and so on. This approach suffers two problems, namely, the difficulty of determining the number of clusters, and the hard quantization loss in clustering. In this paper, we follow the iterative training mechanism but discard clustering, since it incurs loss from hard quantization, yet its only product, image-level similarity, can be easily replaced by pairwise computation and a softened classification task. With these improvements, our approach becomes more elegant and is more robust to hyper-parameter changes. Experiments on two image-based and video-based datasets demonstrate state-of-the-art performance under the unsupervised re-ID setting.

1. Introduction

Given a query image, person re-identification (re-ID) aims to match the person across multiple non-overlapped cameras. In the last few years, person re-ID has drawn increasing research attention [12, 45, 46, 25, 24, 23], due to its wide range of applications such as finding people of interest (*e.g.*, lost kids or criminals) and person tracking. However, most of the proposed methods are of supervised manner, which requires intensive manual labeling and is not applicable to real-world applications. To relieve the scalability problem, in this paper, we focus on the unsupervised re-ID task.

*This work was done when the first author was an intern at Huawei Noah's Ark Lab.

†Qi Tian is the corresponding author.



(a) Unlabeled feature space (b) Clustering based method (c) Our method

Figure 1. (a) Unlabeled images are represented as gray circles in the feature space. The image A and B are of the same person, with an initialized distance of 0.5. The image C is from another person. (b) The clustering based unsupervised re-ID method roughly divides images into classes for network training. Although images A and B are of the same identity, they are assigned with different pseudo labels and learn to be separated. (c) Our method push circles in similar colors (similar images) closer with a soft constraint.

Different from the unsupervised domain adaptation (UDA) methods [28, 37, 41] that leverage the prior knowledge learned from other re-ID datasets, in this paper, we aim to solve the problem without any re-ID annotation. A branch of methods [4, 14, 15] were verified effective, which adopted an iterative clustering and deep learning mechanism, where the network was trained based on the pseudo labels generated by unsupervised clustering. However, the clustering based methods roughly divided images into clusters for training, which made the model highly depends on the clustering result. As shown in Figure 1 (b), images of the same person could be divided into different clusters, which are further trained to be separated with the wrong assigned pseudo label. Since mistakes of unsupervised clustering are inevitable, learning with a hard quantization loss can be prone to fit the noisy labels produced by clustering.

In this paper, we propose a new framework of unsupervised learning in which clustering is no longer required, and thus the error of the hard quantization loss is relieved. As illustrated in Figure 1 (c), instead of using explicit labels

generated by clustering, we mine the relationship between unlabeled images as a gentle constraint to make similar images have closer representations. Specifically, our framework adopts a classification network with softened labels, where the softened labels reflect the image similarity. Unlike the original one-hot labels that force images belonging to an exact class, we treat the labels as a distribution, that an image is encouraged to be associated with several related classes. For each training data, the network is trained not only to predict the ground-truth class, but motivated to predict the similar classes. The learned embedding is then close to similar ones and has a long distance from irrelevant images. On the one hand, without learning with the hard labels, the hard quantization error is eliminated. On the other hand, the supervision of the softened label is relatively weak, which also provides more room for the algorithm. In order to fully exploit the potential of the model, we introduce some auxiliary information to help find similar images. Specifically, when measuring the similarity between images, camera ID and partial details of each pedestrian image are studied. To relieve the issue of camera variance, we propose the cross-camera encouragement term (CCE) that promotes the softened similarity learning from images under different camera views. In this way, the model will learn from more diverse data. Note that the camera ID is automatically obtained at the moment of capturing and is no need for human labeling. Moreover, we extract part features and consider the partial details along with the global appearance as an additional clue.

We evaluate the proposed method on two image-based and two video-based re-ID datasets. The experimental results reveal that our method is robust and stable during iterations via softened similarity learning. Our method outperforms state-of-the-art unsupervised methods on all the four datasets. With the high accuracy and the advantage that does not require any annotations, our approach is easy to be deployed in real-world applications.

Our contributions can be summarized in two-fold. **First**, we propose an unsupervised re-ID framework via softened similarity learning. A classification network is adopted with re-assigned soft label distribution to learn from similar images with smooth constraint. By pushing each person images to get closer to similar images and pushing all other person images away from each other, our framework learns a robust and discriminative model with high potential. **Second**, to make use of the high potential model, we introduce auxiliary information to guide similarity estimation. A cross-camera encouragement (CCE) term is proposed to encourage the similarity exploration between images of different camera views. The fine-grained details are also considered when measuring similarity. These strategies are also proven effective when plugging in other unsupervised re-ID methods.

2. Related Works

2.1. Supervised Person Re-identification

Most re-ID methods are in a supervised manner, in which sufficient labeled images are given. Recently, with the developing of deep learning approaches [36, 35, 34], methods with convolutional neural networks have dominated the re-ID community [12, 26, 45, 46, 25, 16]. Specifically, methods proposed to learn discriminative features from parts of pedestrian images achieve impressive performance [24, 8, 23]. For example, in [24], the feature maps are cut into uniform pieces for classification, and the part-informed features are assembled as the descriptor. A refined part pooling is further proposed to reinforce the within-part consistency in each part.

In our paper, we focus on unsupervised re-ID without annotated labels. We take advantage of the strategy of learning from the part. To exploit the fine-grained information, we directly divide the global feature into horizontal pieces to measure the similarity between each pair of the corresponding parts.

2.2. Unsupervised Domain Adaptation

To relieve the scalability problem of supervised re-ID, some unsupervised domain adaptation methods (UDA) [21, 4, 28, 3, 29, 1, 18] were proposed to learn a re-ID model from a labeled source domain and an unlabeled target domain. Wang *et al.* [28] proposed to learn an attribute-semantic and identity discriminative representation from the source dataset, which is transferable to the target domain. In [37], a PatchNet pre-trained on the source dataset is used to generate pedestrian patches. A network is then designed to pull similar patches together and push the dissimilar patches. In [41], a soft multilabel is learned for each unlabeled person by comparing the unlabeled person with a set of known reference persons from the source domain. Zhong *et al.* [49] proposed a framework that consists of a classification module and an exemplar memory module, which calculates the cross-entropy loss for labeled source data and saves the up-to-date features for target data and computes the invariance learning loss for unlabeled target data, respectively.

Unsupervised domain adaptation methods usually obtain impressive performance. However, these methods take advantage of the external source domain, which is annotated with cross-camera identity labels. In contrast, we focus on the fully unsupervised re-ID task without any external dataset or identity annotation.

2.3. Unsupervised Person Re-identification

The traditional unsupervised methods usually fall into three categories, designing hand-craft features [6, 5, 13, 17, 20], exploiting localized salience statistics [43, 42, 27]

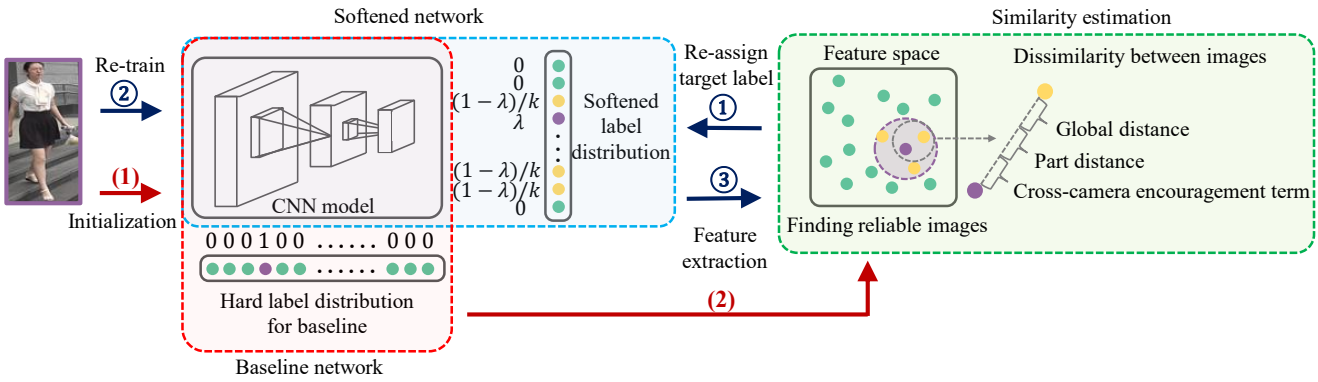


Figure 2. Overview of our method. First, a baseline network with hard label distribution is adopted for initialization, which is shown following the red arrows. Subsequently, with the initialized network, three procedures are conducted iteratively: 1. Feature embeddings of training images are extracted; 2. Similarity among images is estimated to re-assign the target label; 3. The network is re-trained with the softened labels. These procedures are shown following the blue arrows. Notably, the procedures with red arrows are conducted once, while the procedures with blue arrows are conducted iteratively.

or dictionary learning based methods [10, 9]. The performances of these methods are usually low, because it is challenging to design features for images captured by different cameras, under different illumination and view condition. In [40], camera information is used to learn view-specific projection for each camera view by jointly learning the asymmetric metric and seeking optimal cluster separations. However, this method is not suitable for dataset captured by multiple cameras, because the view-specific projection is learned from a pair of cameras.

Recently, Lin *et al.* [14] proposed a bottom-up clustering framework that iteratively trains a network based on the pseudo label generated by unsupervised clustering. However, due to the clustering error, images could be assigned wrong pseudo labels, and the network will then affected by the hard quantization error. Moreover, the clustering is applied based on the clustering result in previous iterations, which accumulates the clustering error during iterations. On the contrary, we propose a framework that mines the similarity as a soft constraint. By regarding each training image as a different class and training with the softened label distribution, we avoid quantization loss and provide more room for the algorithm.

3. Proposed Method

In this paper, we focus on the *unsupervised* re-ID problem. Given a training set of pedestrian images, we aim to learn a feature embedding function for the person images by exploring the image relationship instead of using human annotations. Then, in the evaluation stage, for both query data and gallery data, we use the learned feature embedding function to embed each image into the feature space. 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.

Under the unsupervised setting, the image labels are unknown, so that we regard each image as a different class to initialize a network and gradually mine similarity among unlabeled images as gentle supervision. As illustrated in Figure 2, our framework combines three sub-components (shown in three colored rectangles) : (1) A baseline classification network is adopted to classify each image into different classes. The baseline is used as initialization to generate feature representations; (2) The similarity between unlabeled images is explored based on the feature embedding and the auxiliary information to select reliable images for each training data; (3) The target label distribution is softened according to the reliable images, and the network is fine-tuned with the softened labels to pull the selected reliable images together and repel the other images.

3.1. Baseline: Initialization with Hard Labels

Under an unsupervised person re-ID setting, suppose we have a training set $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, where each x_i is an unlabeled person image. Our goal is to learn a feature embedding function $\phi(\theta; x_i)$ from $\mathcal{X}$ without any manual annotation, where parameters of ϕ are collectively denoted as θ. Since we do not have ground truth identity label for each image x_i, initially we assign each training data x_i by its index, *i.e.*, $\{y_i = i \mid 1 \leq i \leq N\}$. y_i is the initial pseudo label for data x_i. In this way, each training image is assumed to fall into an individual class by itself.

Following [33, 32, 14], we adopt the classification model with a non-parametric classifier, where a lookup table is used to store the features of all training images. The stored feature of each image is then used as the weight vector of each class. We formulate the classification objective using the softmax criterion. For each image x, we normalize its

feature $||\boldsymbol{v}|| = 1$ via $\boldsymbol{v} = \frac{\phi(\boldsymbol{\theta};x)}{||\phi(\boldsymbol{\theta};x)||}$. Then the probability of an image belongs to the i-th class is defined as:

$$p(y_i|x, \boldsymbol{V}) = \frac{\exp(\boldsymbol{V}_i^\top \boldsymbol{v}/\tau)}{\sum_{j=1}^{N} exp(\boldsymbol{V}_j^{\mathrm{T}} \boldsymbol{v}/\tau)}, \tag{1}$$

where $\boldsymbol{V} \in \mathbb{R}^{N \times n_\phi}$ is the lookup table that stores the feature of each class, $\boldsymbol{V}_j$ is the j-th column of $\boldsymbol{V}$ which indicates the feature of j-th class. N is the number of classes, which is the same as the number of training images. τ is a temperature parameter [7] that controls the softness of probability distribution over classes. We set $\tau = 0.1$ following [33].

The loss function is formulated as:

$$\mathcal{L} = -\sum_{j=1}^{N} \log(p(y_j|x_i, \boldsymbol{V})t(y_j), \tag{2}$$

where $t(y_j)$ is the conditional empirical distribution over class labels. We set the probability of the distribution to 1 for the ground truth class, and 0 for all other classes.

The objective Eq. 2 maximizes the cosine distance between each image feature $\boldsymbol{v}_i$ and each features in the lookup table $\boldsymbol{V}_{j \neq y_i}$, while minimizes the cosine distance between each image feature $\boldsymbol{v}_i$ and the corresponding centroid feature $\boldsymbol{V}_{j=y_i}$.

3.2. Model Learning with Softened Similarity

The initialized baseline network learns to recognize each unlabeled image and obtains an initial discriminative ability. By Eq. 1, each training sample is learned to push other training images away. However, there are images of the same identity, which are supposed to be close in the feature space. Forcing the images of the same person to have obviously different representations will have a negative effect on the network. Inspired by ECN [49, 50], we propose to learn a similar representation for images that estimated to be the same identity.

To find the images of the same identity, we select images with the smallest dissimilarity for each training sample. For two images x_a, and x_b, we define the dissimilarity between the two images as the distance between two images, *i.e.*, $D(x_a, x_b) = d(x_a, x_b)$, where the distance is calculated as the Euclidean distance between the two image features, *i.e.*, $d(x_a, x_b) = \|\phi(\boldsymbol{\theta}; x_a) - \phi(\boldsymbol{\theta}; x_b)\|$. Then for each training image x_i, k images with the smallest dissimilarity are selected as reliable images. We define a reliable image set $\mathcal{X}_i^{\text{reliable}} = \{x_i^1, x_i^2, ...x_i^k\}$ with label $\mathcal{Y}_i^{\text{reliable}} = \{y_i^1, y_i^2, ...y_i^k\}$. Each element x_i^j is estimated to be the same identity as x_i, and each class y_i^j is regarded as reliable class.

Instead of taking reliable images as the same class for training, we propose a softened classification network that learns the similarity among identities in a more smooth way. During training, we want the network could not only predict each image into the ground truth class, but make it acceptable to predict the training image into reliable classes. Therefore, we re-assign a nonzero value to the reliable classes in the target label. The target label distribution for data x_i is then written as:

$$t(y_j) = \begin{cases} \lambda, & y_j = y_i \\ (1-\lambda)/k, & y_j \in \mathcal{Y}_i^{\text{reliable}} \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where λ is a hyper-parameter that balances the effect of the ground truth class and the reliable classes. When λ is 1, Eq. 3 reduces to the function with only 0,1 options in the baseline network, that the model learns to recognize each image but fails to learn the similarity and consistency among images of the same person. On the other hand, when λ is too small, the model may fail to predict the ground truth label.

Comparing with the baseline network, the images are labeled with soft label distribution (which denotes probabilities) rather than hard 0,1 labels. The labels are no longer the ground truth classes, but probabilities over k possible reliable classes. By taking reliable classes into account, the confidence of the ground-truth class is reduced, and the confidence of the reliable classes is increased, which guides the network to learn the similarity among images of the identity smoothly. With Eq. 2 and Eq. 3, we define the softened cross-entropy loss as:

$$\begin{aligned} \mathcal{L} = &-\lambda \log(p(y_i|x_i, \boldsymbol{V}) \\ &- \frac{1-\lambda}{k} \sum_{j=1}^{k} \log(p(y_i^j|x_i, \boldsymbol{V}), \end{aligned} \tag{4}$$

The proposed objective not only minimizes the cosine distance between each image feature and the ground truth feature in the lookup table, but minimizes the distance between the features of each image and its reliable images. Meanwhile, the cosine distance between each image feature and the features from other classes is maximized.

With the softened classification network, we gradually learn a feature that is close to reliable images. The learning of the reliable classes is soft and gentle that tries to avoid the negative impact when we involve wrong images in the reliable set. On the other hand, the relatively weak supervision signal makes the model freer and has a higher potential. In this way, we could leverage auxiliary information to help learn a better model. In experiments, we validate that with auxiliary information, the softly learned model performs better than the model learned with hard labels, and will discuss later in Section 4.4.

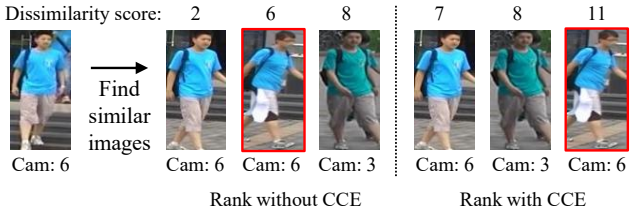


Figure 3. Illustration of the cross-camera encouragement term. When calculating the dissimilarity with and without CCE, the chosen reliable images are different. CCE promotes to find the cross-camera ground truth, instead of the hard negative sample. The negative images are shown in red.

3.3. Similarity Estimation with Auxiliary Information

As illustrated in Section 3.2, for each training sample, k images with the smallest dissimilarity are selected to be reliable. To introduce additional priors for constraints, we also think of other resources to help estimate the similarity.

Part similarity exploration. To assist the similarity measurement between the global feature, we propose to consider the similarity between part features (details) as well. Following [24], we extract the CNN feature map and divide it into p horizontal stripes. Each partition feature is then average pooled to be a part-level feature embedding. We take the average distance of the corresponding parts as the part distance between two images. The part distance between two images x_a and x_b is then formulated as:

$$d_{\text{part}}(x_a, x_b) = \frac{\sum_{i=1}^{p} \|\phi^i(\boldsymbol{\theta}, x_a) - \phi^i(\boldsymbol{\theta}, x_b)\|}{p}, \quad (5)$$

where ϕ^i is the i-th part feature embedding function.

The cross-camera encouragement. We propose a cross-camera encouragement term (CCE) that added to the dissimilarity to promote images captured by different cameras be viewed as reliable images. The intuition of adding CCE is two-fold. First, comparing to inner-camera pairs, the image pairs of different camera ID would teach the network to learn cross-camera information. As a result, the model predicts similar features for a person under different camera views, which benefits the re-ID task. Second, there are many different pedestrians wearing similar clothes that appear under the same camera. CCE helps to find the cross-camera ground truth, instead of these hard negative samples. As shown in Figure 3, without CCE, although the query and the image captured by camera 3 belong to the same person, their dissimilarity is large (8) due to the camera gap. Even a negative example (the one in red) has a smaller distance to the query since they come from the same camera.

Specifically, we denote the camera ID of the training samples as $\mathcal{C} = \{c_1, c_2, ..., c_N\}$. The CCE between two images x_a and x_b is formulated as:

$$\text{CCE}(x_a, x_b) = \begin{cases} \lambda_c, & c_a = c_b \\ 0, & c_a \neq c_b \end{cases}, \quad (6)$$

where λ_c is the parameter that controls the strength of cross-camera promotion. With the CCE term, the dissimilarity between images with the same camera ID is increased. Thus CCE helps to incorporate more cross camera images in the reliable set, and reduce some inner-camera negative images.

Overall dissimilarity. Considering the part similarity exploration and the cross-camera encouragement, the overall dissimilarity $D(x_a, x_b)$ between the image x_a and x_b is then formulated as:

$$D(x_a, x_b) = (1 - \lambda_p) d(x_a, x_b) + \lambda_p d_{\text{part}}(x_a, x_b) + \text{CCE}(x_a, x_b), \quad (7)$$

where λ_p balances the contribution of the global and part similarity. As shown in the green component of Figure 2, the dissimilarity between two images consists of the global distance, the part distance, and the cross-camera encouragement term. By computing the global and the part distance, the similarity of the global appearance and local details are measured, which guarantees the accuracy of reliable image selection. By adding the CCE term, images from different cameras tend to be selected as reliable ones, which enables the network to learn from diverse images. Both of them benefit the discriminative ability of the trained model.

4. Experiments

4.1. Datasets and Implementation Details

The Market1501 dataset [45] is a large-scale dataset captured by 6 cameras for person re-ID. It contains 751 identities for training and 750 identities for testing. The training set, gallery set and query set contain 12936 images, 19732 images and 3368 query images, respectively.

The DukeMTMC-reID dataset [47] is a subset of the DukeMTMC dataset [22]. It contains 1812 identities captured by 8 cameras. Using the evaluation protocol specified in [47], we obtain 2228 query images, 16522 training images and 17661 gallery images.

The MARS dataset [44] is a large-scale video-based dataset for person re-ID. The dataset contains 17503 video tracklets of 1261 identities, where 625 identities are used for training and 636 identities are used for testing.

The DukeMTMC-VideoReID dataset [31] is a video-based re-ID dataset derived from the DukeMTMC dataset [22]. It contains 2196 tracklets of 702 identities for training, 2636 tracklets of other 702 identities for testing.

Implementation details. We adopt ResNet-50 as the CNN backbone and initialize it by the ImageNet [11] pre-trained model with the last classification layer removed.

Methods	Setting	Market-1501				DukeMTMC-reID			
		rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
OIM [33]	**Unsupervised**	38.0	58.0	66.3	14.0	24.5	38.8	46.0	11.3
EUG [31]	OneEx	49.8	66.4	72.7	22.5	45.2	59.2	63.4	24.5
ATNet [18]	UDA	55.7	73.2	74.9	25.6	45.1	59.5	64.2	24.9
ProLearn [30]	OneEx	55.8	72.3	78.4	26.2	48.8	63.4	68.4	28.5
SPGAN [3]	UDA	58.1	76.0	82.7	26.7	46.9	62.6	68.5	26.4
TJ-AIDL [28]	UDA	58.2	-	-	26.5	44.3	-	-	23.0
BUC [14]	**Unsupervised**	61.0	71.6	76.4	30.6	40.2	52.7	57.4	21.9
HHL [48]	UDA	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
Baseline	**Unsupervised**	34.4	54.1	62.3	13.2	16.5	29.9	37.3	7.9
Ours (w/o part and CCE)	**Unsupervised**	58.7	70.4	76.3	29.8	31.6	48.3	53.4	17.4
Ours (w/o part)	**Unsupervised**	68.4	80.8	84.1	35.1	49.2	61.3	65.8	26.4
Ours	**Unsupervised**	**71.7**	**83.8**	**87.4**	**37.8**	**52.5**	**63.5**	**68.9**	**28.6**

Table 1. Comparison with the state-of-the-art methods on the two image-based re-ID datasets, *i.e.*, the Market-1501 dataset and the DukeMTMC-reID dataset. In the column "Setting", "UDA" denotes the unsupervised domain adaptation methods. "OneEx" denotes the methods use the one-example annotation, in which each person in the dataset is annotated with one labeled example.

Methods	Setting	MARS				DukeMTMC-VideoReID			
		rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
OIM [33]	**Unsupervised**	33.7	48.1	54.8	13.5	51.1	70.5	76.2	43.8
DGM+IDE [39]	OneEx	36.8	54.0	-	16.8	42.3	57.9	69.3	33.6
Stepwise [19]	OneEx	41.2	55.5	-	19.6	56.2	70.3	79.2	46.7
RACE [38]	OneEx	43.2	57.1	62.1	24.5	-	-	-	-
DAL [2]	**Unsupervised**	49.3	65.9	72.2	23.0	-	-	-	-
BUC [14]	**Unsupervised**	57.9	72.3	75.9	34.7	76.2	88.3	91.0	68.3
EUG [31]	OneEx	62.6	74.9	-	42.4	72.7	84.1	-	63.2
Ours	**Unsupervised**	**62.8**	**77.2**	**80.1**	**43.6**	**76.4**	**88.7**	**91.0**	**69.3**

Table 2. Comparison with the state-of-the-art methods on two video-based re-ID datasets, MARS and DukeMTMC-VideoReID. In the column "Setting", "OneEx" denotes the methods use the one-example annotation, in which each person in the dataset is annotated with one labeled example. "UDA" denotes the unsupervised domain adaptation methods.

The number of training epochs for the baseline network is set to be 25 for image-based datasets and 30 for video-based, the batch size is set to be 16, the dropout rate is set to be 0.5. The λ is set to 0.6. The λ_p and λ_c are set to be 0.5 and 0.02 respectively. The number of parts p is set to be 8. We use stochastic gradient descent with a momentum of 0.9 to optimize the network. The learning rate is initialized to 0.1 and changed to 0.01 after 15 epochs. For video-based datasets, we take the average feature of all frames within a tracklet to be the tracklet feature. We implement our method on both PaddlePaddle and PyTorch. 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 12 hours.

4.2. Comparison with the State-of-the-Arts

Image-based Person Re-identification. The comparisons with the state-of-the-art algorithms on Market-1501 and DukeMTMC-reID are shown in Table 1. On Market-1501, under the same setting, we obtain the best performance among the compared methods with **rank-1 = 71.7%, mAP = 37.8%**. Compared to the state-of-the-art unsupervised method BUC [14], we achieve 10.7 points and 7.2 points improvement on rank-1 accuracy and mAP, respectively. On DukeMTMC-reID, compared to BUC, our method achieves 12.3 and 6.7 points of improvement on rank-1 accuracy and mAP, respectively. The impressive performance indicates that the softened similarity learning successfully finds images of the same identity and encourages reliable images gathered in the feature space. The proposed CCE helps to learn a discriminative model cross-camera, while the part similarity estimation helps to maintain an accurate reliable image selection.

Video-based Person Re-identification. The comparisons with the state-of-the-art algorithms on the two video-based datasets are shown in Table 2. On MARS, we obtain rank-1 = 62.8%, mAP = 43.6%. Compared to BUC [14], We achieve 4.9 and 8.9 points of improvement in rank-1 accuracy and mAP, respectively. On DukeMTMC-VideoReID, we achieve rank-1 of 76.4% and mAP of 69.3%, which beats BUC by 0.2 and 1.0 points, respectively. The performance gap between ours and BUC is relatively small on

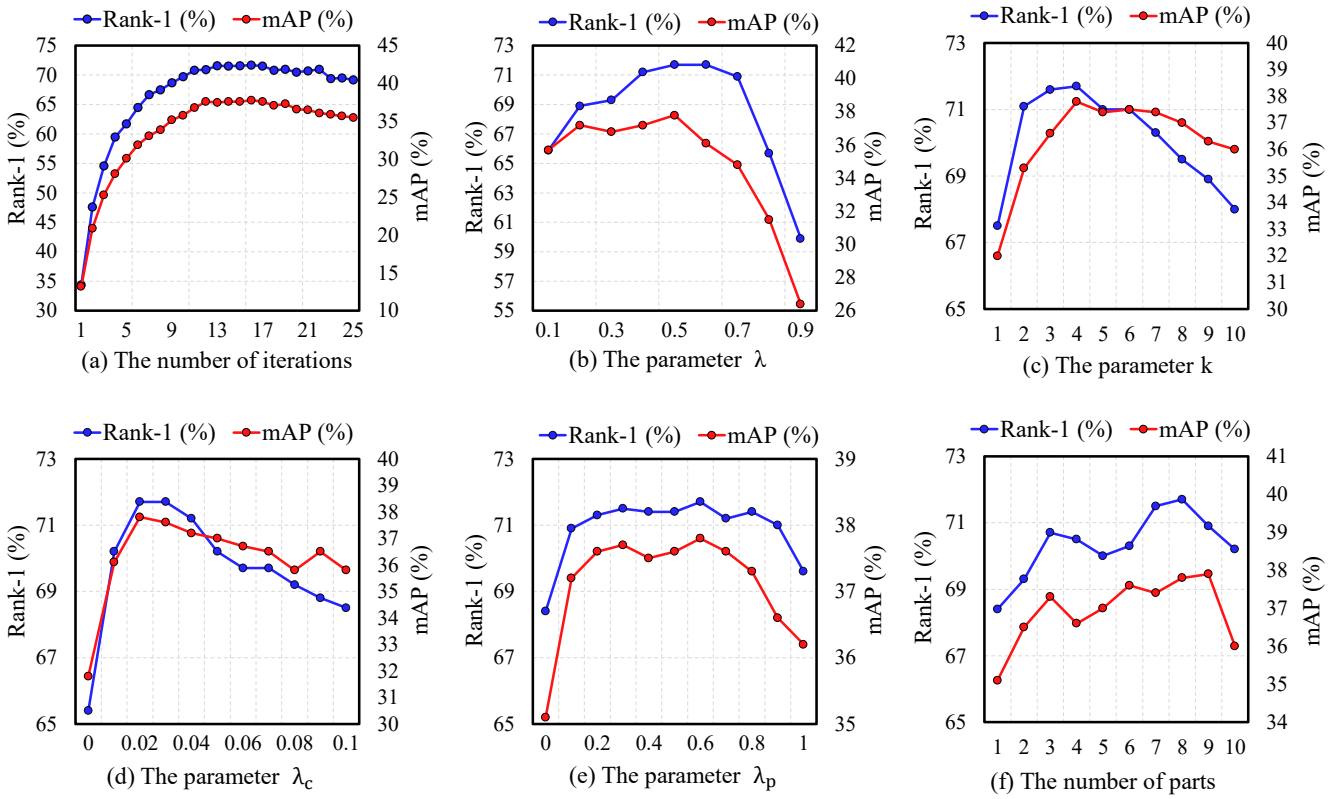


Figure 4. Parameter and method analysis on Market-1501. (a) The performance along with iterations. (b) The impact of λ for softened classification. (c) The impact of the number of reliable images k. (e) The impact of λ_c for CCE. (e) The impact of λ_p for part distance. (f) The impact of the number of parts.

DukeMTMC-VideoReID. We suspect that, the rank-1 accuracy of BUC is 76.2%, which is quite high under the unsupervised setting, and it is more difficult to make progress upon a high performance. Note that without any annotation, we still beat the EUG method in the one-example setting, where each person is annotated a tracklet as labeled data.

4.3. Diagnostic Studies

Robustness test. Figure 4 (a) illustrates the re-ID performance over iterations. Throughout iterations, the rank-1 accuracy constantly increases from 34.4% to 71.7%, which indicates that the model grows robust steadily. After the 16^{th} iteration, the re-ID performance stops to increase and shows a slight decrease. Note that, from the 10^{th} iteration to the 25^{th} iteration, our method always maintains a high re-ID performance, *i.e.* with a rank-1 accuracy higher than 69%, which demonstrates the robustness of the proposed method.

The impact of the hyper-parameter λ. In Eq. 3, the hyper-parameter λ controls the degree of softening, which balances the impact of the ground truth class and the selected reliable classes. When λ is 0, each training image is learned to be predicted to the reliable classes. When λ is 1, each training image will be predicted to its own ground truth class. We vary λ from 0.1 to 0.9 in Figure 4 (b), and observe that, when λ increases from 0.1 to 0.6, the re-ID performance continue increasing. When λ continually gets larger, we observe a obvious drop on re-ID performance.

The impact of the number of reliable images k. Figure 4 (c) shows how the re-ID performance varies with different numbers of reliable images, k. We observe that as k increases from 0 to 4, the re-ID performance continues rising, and the performance begins to drop when k gets larger. The reason is that when k is too small, the learned similarity of one identity is not adequate, which makes the model difficult to match images of the same identity. When k is too large, error cases will be involved in the reliable set, which harms the network training when forcing images of different persons to get closer.

The impact of the cross-camera encouragement term. As shown in Table 1, on Market-1501, the results of Ours (w/o part) beat Ours (w/o part and CCE) by 9.7 and 5.3 points on rank-1 and mAP respectively. On DukeMTMC-reID, the improvements on rank-1 and mAP are 17.6 and 9.0, respectively. The impressive improvements demonstrate the effectiveness of the CCE module. Without CCE, images of one identity from different cameras are hard to be

Dataset	Auxiliary	Ours		BUC	
		rank-1	mAP	rank-1	mAP
Market-1501	None	58.7	29.8	61.0	30.6
	CCE	68.4	35.1	65.9	31.8
	CCE+part	**71.7**	**37.8**	69.5	36.2
DukeMTMC	None	31.6	17.4	40.2	21.9
	CCE	49.2	26.4	48.3	24.4
	CCE+part	**52.5**	**28.6**	51.5	25.1

Table 3. Comparison with BUC [14] on Market-1501 and DukeMTMC-reID. The column "Auxiliary" lists the auxiliary information utilized by the method.

selected as reliable images because of the camera variance. CCE encourages cross-camera image selection, which enables the model learning from diverse images and getting robust to camera views. Besides, we evaluate CCE based on BUC [14], and the result is shown in Table 3. We observe that on Market-1501, the improvements of using CCE are 4.9 points for rank-1 and 1.2 points for mAP. This further indicates that CCE is effective and can be easily adopted to other unsupervised methods to achieve better performance.

The parameter λ_c of CCE balances the effect of appearance and camera diversity. We evaluate different values of λ_c in Figure 4 (d). As λ_c increases from 0 to 0.02, the rank-1 accuracy on Market-1501 increases from 65.4% to 71.7%. If we set λ_c to be greater than 0.02, the too large encouragement term would lead to a negative effect on the performance.

The impact of part similarity. As shown in Table 1, on Market-1501, the results of Ours beat "Ours (w/o part)" by 3.3 and 2.7 points on rank-1 accuracy and mAP, respectively. On DukeMTMC-reID, the improvements on rank-1 and mAP are 3.3 and 2.2, respectively. We also evaluate the part similarity based on BUC [14], and the result is shown in Table 3. We observe that on Market-1501, the improvements of using the part similarity are 3.6 points for rank-1 and 4.4 points for mAP. This demonstrates that investigating the appearance between pedestrian parts is beneficial for similarity estimation. This idea is also effective on other unsupervised methods and can be easily adopted.

The parameter λ_p balances the effect of the global distance and part distance. We evaluate different values for the parameter λ_p in Figure 4 (e). When $\lambda_p = 0$, we only adopt the global distance. As λ_p increases, retrieval accuracy improves at first. When $\lambda_p = 0.5$, we obtain the best performance. After that, the performance begins to drop.

In Figure 4 (f), we vary the number of parts p from 1 to 10. When $p = 1$, the part distance is the same as the global distance. When $p = 8$, we obtain the best accuracy.

4.4. Delving into the Softened Similarity Learning and the Hard Label Learning

We conduct experiments to examine the effectiveness of our method of softened similarity learning and BUC that learn from hard labels. The models learning with and without the auxiliary information are compared. The experimental results are summarized in Table 1 and Table 3.

We first observe that both approaches without the auxiliary information yield improvement over the baseline. As shown in Table 1, the rank-1 of the baseline and Ours (w/o part and CCE) are 34.4% and 58.7%, respectively. The large performance gap demonstrates the effectiveness of softened similarity learning. Second, as shown in Table 3, without any auxiliary information, BUC achieves better performance than ours on both datasets. We think the reason is that our network is trained with softened labels, which avoids pushing images of the same identity away, but it is also has a relatively small strength to force images of different identities separate. Nevertheless, from Table 3, we find that when we adopt CCE or part similarity into the two approaches, our method exceeds BUC on both datasets. This indicates that given a better similarity estimation, the softened similarity learning has a higher potential to learn better embeddings. We suspect that when learning with hard one-hot labels, the model is forced to fit the noise labels, which limits its accuracy. In contrast, our method hardly affected by the inaccurate similarity estimation and thus has more room to learn and improve.

5. Conclusions

In this paper, we investigate the problem of unsupervised re-ID. Following the pipeline of iterative person recognition and feature update, we propose not to assign each sample with a hard label so as to avoid quantization loss as well as provide more room for the learning algorithm. To introduce additional priors for constraints, we introduce several auxiliary information, including a camera-based term which is easy to obtain yet useful for distance amendment. Experiments on both image-based and video-based re-ID tasks validate the effectiveness of our approach.

This work puts forward a point that classification may not be the optimal supervision, in particular, for unsupervised re-ID. This reminds us of the difference between classification-based and metric-learning-based methods in supervised re-ID. The potential connections between them remain uncovered, which we will investigate in the future research.

Acknowledgments. This work is supported by National Nature Science Foundation of China (61931008, 61671196, 61701149, 61801157, 61971268, 61901145, 61901150, 61972123), National Natural Science Major Foundation of Research Instrumentation of PR China under Grants 61427808 , Zhejiang Province Nature Science Foundation of China (LR17F030006, Q19F010030), 111 Project, No. D17019.

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018.

[2] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. In *BMVC*, 2018.

[3] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[4] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMM*, 2018.

[5] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[6] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Workshops*, 2014.

[8] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[9] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised l1 graph learning. In *ECCV*, 2016.

[10] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[13] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

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

[15] Yutian Lin, Yu Wu, Chenggang Yan, Mingliang Xu, and Yi Yang. Unsupervised person re-identification via cross-camera similarity exploration. *IEEE TIP*, 2020.

[16] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[17] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 2014.

[18] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019.

[19] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017.

[20] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[21] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016.

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

[23] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, 2019.

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

[25] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.

[26] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[27] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

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

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

[30] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. Progressive learning for person re-identification with one example. *IEEE TIP*, 2019.

[31] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

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

[33] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[34] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE TPAMI*, 2020.

[35] Chenggang Yan, Biyao Shao, Hao Zhao, Ruixin Ning, Yongdong Zhang, and Feng Xu. 3d room layout estimation from a single rgb image. *IEEE TMM*, 2020.

[36] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai.

Stat: spatial-temporal attention mechanism for video captioning. *IEEE TMM*, 2019.

[37] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, 2019.

[38] Mang Ye, Xiangyuan Lan, and Pong C. Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*, 2018.

[39] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. *ICCV*, 2017.

[40] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[41] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019.

[42] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, 2013.

[43] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[44] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

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

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

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

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

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

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