# Improved Hierarchical Clustering with Non-locally Enhanced Features for Unsupervised Person Re-identification

Wanyu Zhao, Bairong Li, Qinghua Gu, Yuesheng Zhu
*Communication and Information Security Laboratory*
*Shenzhen Graduate School of Peking University*
Shenzhen, China
{1801213425, lbairong, guqh, zhuys}@pku.edu.cn

***Abstract*—Due to the high cost of data annotation in supervised person re-identification (re-ID) methods, unsupervised person re-ID methods have attracted more and more attention. The unsupervised person re-ID methods based on deep clustering have achieved good performance. However, the distance metrics used in existing unsupervised clustering methods ignore intra-cluster distance and are likely to cause some wrong merging situations and uneven distribution within clusters. Besides, these models based on deep clustering usually ignore the importance of global features for person re-ID. In this paper, we address the above problems by proposing an improved hierarchical clustering approach with non-locally enhanced features. To improve the clustering performance, we design a new metric which consists of intermediate distance as inter-cluster distance and compactness degree as intra-cluster distance. The former one can prevent some wrong merging situations and the latter one can promote the uniform distribution within clusters. In addition, we develop a non-locally enhanced feature network to take advantage of global features of images. Extensive experiments on Market-1501, DukeMTMC-reID, MARS and DukeMTMC-VideoReID demonstrate that our method obtains significant improvement over the state-of-the-art unsupervised methods.**

***Index Terms*—hierarchical clustering, intermediate distance, compactness degree, non-local features, person re-identification**

## I. Introduction

Person re-identification (re-ID) aims at matching the same person from images taken by different cameras. It has drawn lots of attention from science and industry due to its important application in safety and surveillance. Most existing person re-ID methods [1-5] are supervised methods that depend on annotation data. They are labor-consuming and expensive. The limited generalization ability of supervised methods in real scenarios motivates the research of unsupervised methods [12-19].

Traditional unsupervised methods are often based on manual features [6-8], saliency indicators [9,10] and sparsity constraints [11]. However, the performance of these traditional methods is much inferior to that of supervised methods. Recently, unsupervised methods based on deep learning have been applied to person re-ID. These methods are usually divided into two categories. One is transfer learning and the other is clustering. Cross-domain person re-ID [12-18] focuses on transferring the knowledge learned from a labeled source domain to an unlabeled target domain. It cannot be viewed as pure unsupervised learning due to its need for additional source data.

Clustering is a classical method for unsupervised learning. Fan et al. [18] present a progressive unsupervised learning (PUL) method that utilizes *k*-means algorithm to cluster different features in each iteration. However, the correct number of clusters is unknown and PUL [18] method depends on an annotated source domain. To overcome these shortcomings, Lin et al. [19] present a bottom-up clustering (BUC) approach to unsupervised person re-ID. The framework of BUC [19] applies network training and hierarchical clustering iteratively without any dependence on auxiliary data samples. Nonetheless, the existing unsupervised clustering methods still have a few problems. Firstly, inappropriate inter-cluster distance metrics are adopted, which will lead to poor clustering performance. For example, BUC [19] method adopts the minimum distance between images in two clusters as the merging criterion. The minimum distance criterion can result in some wrong merging situations and is prone to forming elongated clusters. Secondly, the intra-cluster distance is ignored, while a good clustering should have large inter-cluster distance and small intra-cluster distance. So we should consider both inter-cluster distance and intra-cluster distance when calculating the distance between clusters. Thirdly, the importance of global features is ignored. As we known, convolutional neural network and recurrent neural network are operations that only work on one local neighborhood at a time. After multi-level convolutional operations, some non-local information will lose. However, the person image takes up most of the image itself. Images of the same person may not be similar in local parts, but similar on the whole. Therefore, it is necessary to take global features into account.

To address the above problems, we present an improved hierarchical clustering approach with non-locally enhanced features for pure unsupervised person re-ID. Firstly, we present an intermediate distance (IMD) as inter-cluster distance by considering both the minimum distance and the maximum distance between clusters. IMD can avoid some wrong merg-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

ing situations to a certain extent. Secondly, we propose a compactness degree (CPD) as intra-cluster distance to relieve uneven distribution within clusters. IMD combined with CPD can promote the merging of clusters with a single sample and prevent the formation of clusters with large looseness. Thirdly, we take global features into account by designing a non-locally enhanced feature network. Specifically, the non-local operations [29] and a mixed pooling strategy are applied to unsupervised person re-ID for enhancing global features.

The experimental results demonstrate that our method is superior to the state-of-the-art pure unsupervised methods on both image-based and video-based re-ID datasets, and even better than some transfer learning and one-shot learning methods.

## II. Ralated Work

### A. Traditional Unsupervised Person re-ID

In recent years, a few unsupervised methods based on manual features [6-8], saliency indicators [9,10] and sparsity constraints [11] have been proposed. Farenzena et al. [8] present a feature extraction method based on human body symmetry to reduce the view variances. Wang et al. [9] introduce a novel unsupervised model for localized human appearance saliency selection by exploring generative probabilistic topic modelling. However, the performance of these methods is much inferior to that of supervised methods due to lack of identity labels.

### B. Cross-domain Unsupervised Person re-ID

Cross-domain unsupervised person re-ID aims at handling an unlabeled target domain with the help of a labeled source domain. The cameras that collect data from two domains are usually different, so the data distribution of the two domains will be different eventually. To address the domain gap, many researchers have tried diverse methods, such as learning an invariant mapping from the source domain to the target domain [13] or generating new images belonging to the target domain for getting better generalization performance by utilizing GANs [15, 16, 21, 22]. In [13], Peng et al. present a multi-task dictionary learning method that can learn a dataset-shared but target-data-biased representation. To handle the discrepancy in camera styles, viewpoints and environments, Zheng et al. [22] provide an end-to-end joint learning framework that depends on data generation. These methods all depend on an auxiliary labeled dataset and the assumption that two domains share the same identity space. Different from them, our method does not use any auxiliary datasets or annotation data.

### C. Deep Clustering Unsupervised person re-ID

Recently, deep clustering has been adopted in unsupervised person re-ID [18, 19]. In [18], Fan et al. use an annotated dataset to train the model and transfer it to an unlabeled target dataset. Then, they use $k$-means algorithm to select reliable samples from the unlabeled dataset to update the model. However, this method requires a labeled source domain and depends on an assumption about the number of identities. Lin et al. [19] apply the hierarchical clustering on the CNN features to merge images from bottom to up. In the beginning, each image is regarded as a cluster, and some clusters are merged by measuring the similarity between clusters. Then the newly formed cluster result is used to update the model. In [19], the minimum distance between images in two clusters and a diversity regularization term are adopted as the merging criterion. Different from BUC [19], we introduce a new distance metric that includes both inter-cluster distance and intra-cluster distance, which can inhibit wrong merging situations and promote uniform distribution within clusters to a certain extent.

### D. Non-local Neural Network

Non-local technology [20] is a classical image denoising algorithm that calculates a weighted average of all pixels in an image. Wang et al. [29] introduce a non-local architecture that links self-attention in machine translation to the more general non-local filtering operations. In each 2D non-local operation, the response at a position is computed as a weighted sum of the features at all positions, which can enlarge the receptive field from neighbor positions to the entire image. The non-local operations can establish the connection between two pixels with a certain distance on the image. Inspired by these work, we embed non-local blocks into the CNN model to enhance global features. As far as we know, our proposed non-locally enhanced feature network is the first piece of work that applies non-local operations in unsupervised person re-ID.

## III. Proposed Method

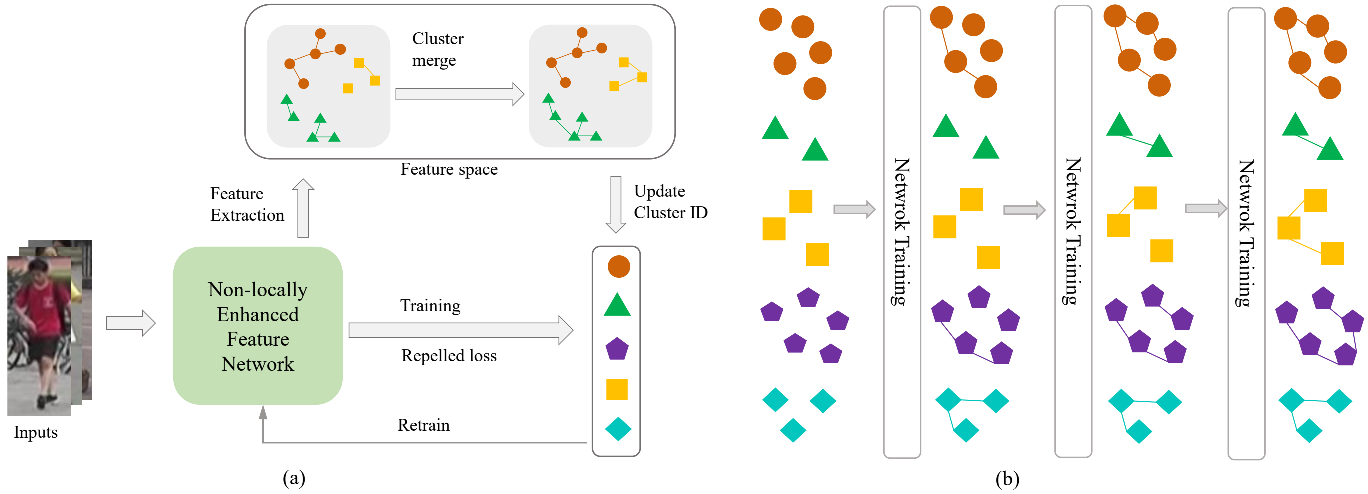
### A. Preliminary

Given an unlabeled dataset of $N$ images $X = \{x_1, x_2, \ldots, x_N\}$, we need to learn a feature embedding function $\phi(x_i;\theta)$ from $X$ without any available annotation, where $\theta$ is the weight of the network. The feature extractor can be applied to the query set $X^q = \left\{x_1^q, x_2^q, \ldots, x_{N_q}^q\right\}$ and the gallery set $X^g = \left\{x_1^g, x_2^g, \ldots, x_{N_g}^g\right\}$. The distance between each pair of images is defined as the Euclidean distance between the feature embeddings of the two images, *i.e.*, $d\left(x_i^q, x_i^g\right) = \left\|\phi\left(x_i^q;\theta\right) - \phi\left(x_i^g;\theta\right)\right\|_2$.

In supervised learning methods, we have handcraft annotations about $n$ identities, *i.e.*, Label $= \{y_1, y_2, \ldots, y_n\}$. A classifier $f\left(\phi\left(x_i;\theta\right);\omega\right)$ parameterized by $\omega$ is used to predict the identity of the image $x_i$. Thus, the feature extractor $\phi\left(x_i;\theta\right)$ can be learned by optimizing the following formula:

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

where $l$ is the cross-entropy loss for classification. However, there are no handcraft annotations like $y_i$ in unsupervised learning methods. To solve the problem, repelled loss [19] has been proposed to act as a classifier $f$ that can jointly consider intra-cluster and inter-cluster distance.



(a)

(b)

Fig. 1. (a) The whole framework of the unsupervised clustering model. The framework iteratively trains the network and merges clusters. The clustering results are fed back to the network for further updating. (b) The hierarchical clustering process. Each step will merge some clusters according to the distance between clusters. By the hierarchical clustering, samples of the same person are merged into a cluster to represent an identity.

The probability that image $x$ belongs to the *i-th* cluster is defined as:

$$p(i|x, V) = \frac{\exp\left(V_i^{\mathrm{T}} \boldsymbol{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(V_j^{\mathrm{T}} \boldsymbol{v}/\tau\right)} \tag{2}$$

where $\boldsymbol{v}$ is the $L_2$ normalized image feature obtained from $\phi(x;\theta)$, $V$ is a lookup table that stores the centroid feature of each cluster, $V_i$ contains the information of all images in the *i-th* cluster, $n$ is the number of clusters in current iteration and $\tau$ is a temperature parameter [27] that controls the softness of probability distribution over classes. According to [19], we set $\tau = 0.1$. The lookup table $V$ can avoid calculations that extract features from all data at each training step, and $V$ is updated by the exponential moving average [28]. Lastly, we optimize the repelled loss by the following formula:

$$\mathcal{L} = -\log\left(p\left(i|x, V\right)\right) \tag{3}$$

### *B. The Distance Metric Between Clusters*

The whole framework of unsupervised clustering model is illustrated in Fig. 1. The aim is to bring images of the same person into a cluster. At first, we set each image as a cluster to train the network. Then we use the network to extract features of images, measure the distance between clusters, merge some clusters and retrain the network with newly formed cluster iteratively. One of the critical factors of using the framework is how to measure the distance between clusters. As shown in Fig. 2, $d_1$ is smaller $d_2$. Cluster A and cluster B will be merged according to the minimum distance criterion of BUC [19], while the right merging should be cluster A and cluster C. To avoid similar wrong merging situations, we propose an intermediate distance (IMD) to evaluate inter-cluster distance. IMD between cluster A and cluster B is formulated as:

$$\mathit{IMD}(A, B) = \frac{1}{2}\left(\min_{x_a \in A, x_b \in B} d\left(x_a, x_b\right) + \max_{x_a \in A, x_b \in B} d\left(x_a, x_b\right)\right) \tag{4}$$

where $d\left(x_a, x_b\right)$ is the Euclidean distance between the feature embeddings of two images. Figuratively, $d\left(x_a, x_b\right) = \left\|\boldsymbol{v}_a - \boldsymbol{v}_b\right\|_2$. IMD considers a broader context of sample-level pairwise relationships than BUC [19], so it can avoid some wrong merging situations. In Fig. 2, IMD can inhibit the wrong merging of cluster A and cluster B and boost the correct merging of cluster A and cluster C.

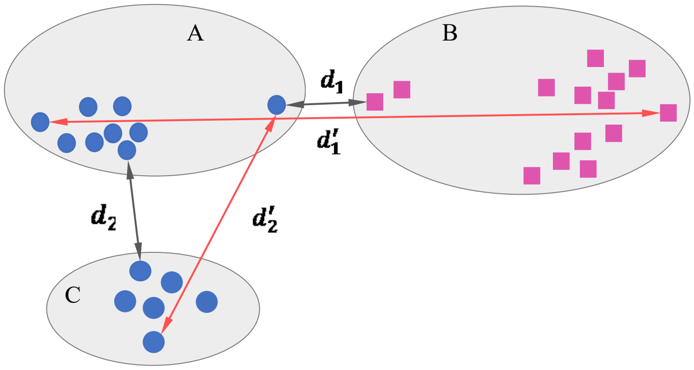


Fig. 2. Different inter-cluster distance. $d_1$ and $d_1'$ are minimum distance and maximum distance between cluster A and cluster B. $d_2$ and $d_2'$ are minimum distance and maximum distance between cluster A and cluster C.

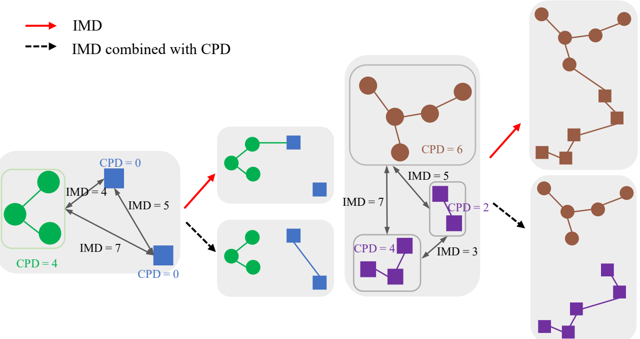
Besides, the compactness degree (CPD) is proposed to evaluate intra-cluster distance. CPD is defined as follows:

$$CPD(A) = \frac{1}{n} \sum_{i,j \in A} d\left(x_i, x_j\right) \tag{5}$$

where $n$ is the number of samples in cluster A, $d\left(x_i, x_j\right)$ is the Euclidean distance between the feature embeddings of two images in cluster A. To consider both intra-cluster and inter-cluster distance simultaneously, we define the final distance between cluster A and B as:

$$D(A, B) = IMD(A, B) + \lambda(CPD(A) + CPD(B)) \tag{6}$$

where $\lambda$ is a parameter that balances the effect of IMD and CPD. Clusters with small distance should be merged because the features of different images of the same person are close in feature space. As shown in Fig. 3(a), the value of CPD for the square clusters are 0. The final distance between the two square clusters only calculates the IMD value, leading to a relatively small distance between the two clusters. Thus, IMD combined with CPD can promote the merging of clusters with a single sample. In Fig. 3(b), the value of CPD for the circular cluster is relatively large. The final distance between the circular cluster and other clusters increases correspondingly, which will inhibit the merging of the circular cluster and other clusters to a certain extent. Therefore, IMD combined with CPD can prevent the formation of slender and loose clusters. In summary, IMD combined with CPD can promote uniform distribution within clusters.



(a)

(b)

Fig. 3. (a) and (b) show respective clustering process. The solid arrows show the clustering results with IMD and the dotted arrows show the clustering results with IMD and CPD. (a) illustrates that IMD combined with CPD can promote the merging of clusters with a singe sample. (b) shows that IMD combined with CPD can prevent the formation of slender and loose clusters.

### *C. Non-locally Enhanced Feature Network*

To make better use of global features in pictures, we design a non-locally enhanced feature network that is illustrated in Fig. 4. We adopt ResNet-50 as the CNN backbone. The last classification layer of ResNet-50 is removed and the non-local blocks [29] are inserted behind layer 2 and layer 3 of ResNet-50. Besides, we use both global average pooling operation and global max pooling operation to maintain the global relationship with the identification and preserve the discriminative part. During the training, a fully connected (FC) layer is added behind pooling layers for feature embedding. The embedding dimension is set to 2048.

The detailed architecture of the non-local blocks is shown in Fig. 4. Specifically, the input feature map is denoted as $F_x$ which has the spatial dimension of $h_x \times w_x \times c_x$. The pair-wise function that calculates the pair-wise relationship is set to dot product:

$$f\left(F_{x,i}, F_{x,j}\right) = \theta\left(F_{x,i}\right)^T \phi\left(F_{x,j}\right) \qquad (7)$$

where $F_{x,i}$, $F_{x,j}$ denote the feature activation $F_x$ at position $i$, $j$ respectively. $\theta(\cdot)$ and $\phi(\cdot)$ are two feature embedding operations with different learned parameters. The non-local blocks in the non-locally enhanced feature network are defined as:

$$y_{x,i} = \frac{1}{N} \sum_{\forall j} f\left(F_{x,i}, F_{x,j}\right) g\left(F_{x,j}\right) \qquad (8)$$

where the unary function $g$ computes a representation of the input signal at the position $j$. $N$ is the number of positions in $F_x$. We wrap the non-local blocks into the ResNet-50 by:

$$z_{x,i} = W_z y_{x,i} + F_{x,i} \qquad (9)$$

where $y_{x,i}$ is given in Eq. (8) and "$+F_{x,i}$" denotes a residual connection. $W_z$ is initialized as zero.

The whole method is described in Algorithm 1. Firstly, the number of clusters is set to the number of training images. After each merging, the labels of training images are updated. We train the network continuously and test its performance on the validation set until it reaches the highest performance.

**Algorithm 1** Clusering framework

```
Input: Training data $X = \{x_i\}_{i=1}^N$
       Hyperparameter $\lambda$
       Initial CNN model $\phi(\cdot;\theta_0)$
       Merge percent $p \in (0,1)$
Output:
       Optimal model $\phi(\cdot;\theta)$
Initialize: labels $Y = \{y_i = i\}_{i=1}^N$, the number of clusters $C = N$, the number of merging clusters $m = C * p$
while $C > m$ do
    Train CNN model with $X$ and $Y$
    Extract features, calculate the distance between clusters by Eq. (6) and update lookup table $V$
    Merge $m$ clusters: $C = C - m$
    Update labels in $Y$ with the newly formed cluster
    Evaluate performance $P_{new}$ on validation dataset
    if $P_{new} > P_{best}$ then
        $P_{best} = P_{new}$;
        Update parameters $\theta$ of the model
    end if
end while
```

## IV. Experiments And Analysis

### *A. Datasets*

**Market-1501** [30] includes 1,501 identities and 32,688 labeled images captured by 6 cameras, among which 12,936 images of 751 identities are used for training and 19,732 images of 750 identities are used for testing.

**DukeMTMC-reID** [31] is the subset of DukeMTMC [34] dataset. It includes 16,522 images of 702 identities for training and 17,661 images of 702 identities for testing.

**MARS** [32] is the largest video-based dataset that is extended from Market-1501. It contains 20,478 automatically generated tracklets (including 3,248 distractors) of 1,261 identities. Specifically, it is divided into 625 identities for training and 636 identities for testing.

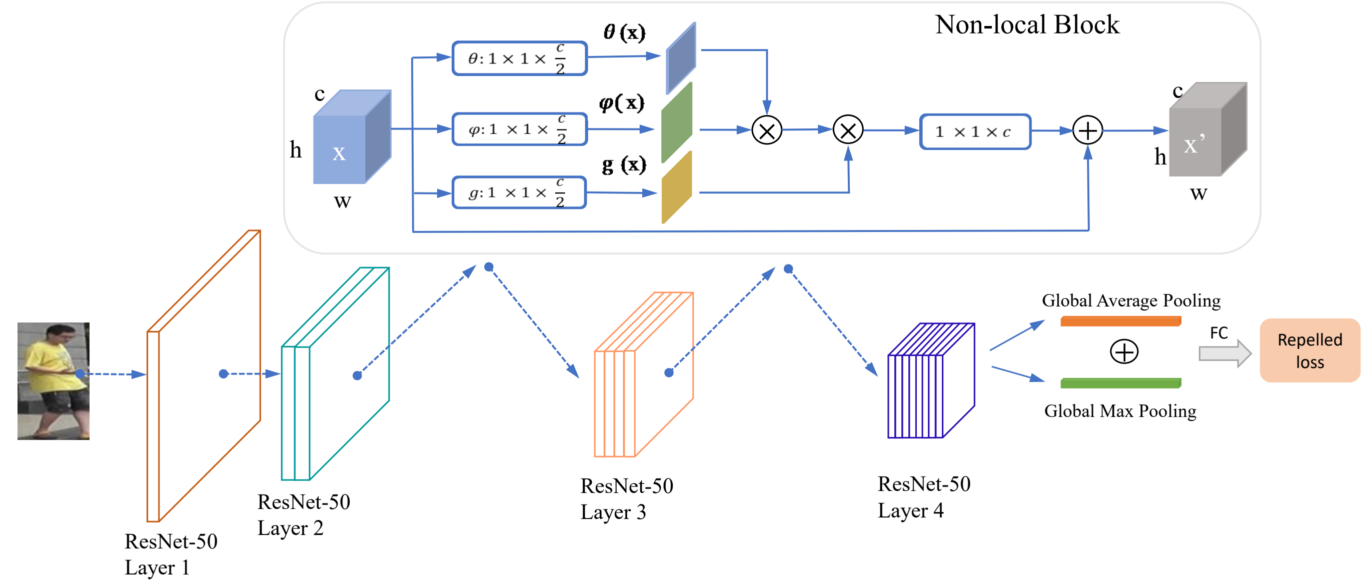


Fig. 4. Illustration of the non-locally enhanced feature network. The input image firstly goes through the backbone network and gets its feature map. Then, we use both global average pooling and global max pooling and add a fully connected (FC) layer behind pooling layers for feature embedding.

**DukeMTMC-VideoReID** [33] is a video-based dataset derived from DukeMTMC [34] dataset. It has 2,196 tracklets of 702 training identities and 2,636 tracklets of 702 testing identities.

## B. Experiment Settings

**Training.** For imaged-based datasets, *i.e.*, Market-1501 and DukeMTMC-reID, we use images removed labels to train our model. For video-based datasets, *i.e.*, MARS and DukeMTMC-VideoReID, we use tracklets to train our model and each tracklet is regarded as an individual. In video-based datasets, we use the average feature of all frames in a tracklet as the feature of the tracklet. It is worth noting that we do not use any labeled data or auxiliary datasets in our experiments.

**Evaluation Metric.** Evaluation Metrics Cumulative Matching Characteristic (CMC) is adopted in quantitative evaluation for person re-ID. The rank-$k$ records the correct matching within the top $k$ ranks to represent the CMC curve. The mean average precision (mAP) evaluates the overall performance of methods. In our study, we use mAP and rank-$k$ to evaluate our model.

**Experimental Details.** Our proposed network framework is shown in Fig. 4. The ResNet-50 parts of the non-locally enhanced network are initialized by the ImageNet pre-trained model, and the parameters of non-local blocks are initialized as 0. For the training process, the training epoch in the first stage is set to be 20, the batch size to be 16, merge percent $p$ to be 0.05, the dropout rate to be 0.5 and $\lambda$ in Eq. (6) to be 0.3. The stochastic gradient descent with a momentum of 0.9 is utilized to optimize the model. The learning rate of parameters is initialized as 0.1 and decreases to 0.01 after 15 epochs.

## C. Comparision with the State-of-the-art Methods

**Image-based Person Re-identification.** The comparison with the state-of-the-art unsupervised methods on two large image-based datasets is reported in Table I. On Market-1501, our method achieves the best performance of which rank-1 is 77.5% and mAP is 50.3% among all pure unsupervised methods. Compared with the best pure unsupervised method BUC [19], our method achieves 11.3% improvement in rank-1 and 12% improvement in mAP. Similarly, our method achieves 63.2% in rank-1 and 38.6% in mAP on DukeMTMC-reID and exceeds other pure unsupervised methods to a large extent. We also compare our method with transfer learning methods and one-shot learning methods. Although these methods use additional datasets or manual annotations, our method is better than most of them, which demonstrates the superiority of our method.

**Video-based Person Re-identification.** Table II shows the comparison with the state-of-the-art unsupervised methods on two large video-based datasets. On MARS, our method achieves 67.5% in rank-1 and 44.1% in mAP, exceeding the most competitive BUC [20] by 6.4% in rank-1 and 6.1% in mAP. On DukeMTMC-VideoReID, our method obtains 80.2% in rank-1 and 73.6% in mAP. It exceeds BUC [19] by 11% in rank-1 and 11.7% in mAP. As can be seen from Table II, our method is also superior to other transfer learning methods and one-shot learning methods.

## D. Ablation Study

**The Effectiveness of IMD and CPD.** We evaluate the effectiveness of our proposed metric method by comparing to the closest work BUC [19]. To verify the effectiveness of IMD, we compare BUC [19] without diversity regularization term with IMD without CPD. As shown in row 1 and row 3

TABLE I
Comparison with the state-of-the-art methods on two image-based datasets. The column “Labels” denotes the type of supervision used by the corresponding method.“None” represents no extra information is used.“Transfer” represents an additional dataset is used.“OneEx” represents one labeled image in per identity is used. * denotes that the results are reproduced by Lin [19].

| Methods | Venue | Labels | Market-1501 | | | | DukeMTMC-reID | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | rank-1 | rank-5 | rank-10 | mAP | rank-1 | rank-5 | rank-10 | mAP |
| UMDL[13] | CVPR’16 | Transfer | 34.5 | 52.6 | 59.6 | 12.4 | 18.5 | 31.4 | 37.6 | 7.3 |
| PUL[18] | TOMM’18 | Transfer | 44.7 | 59.1 | 65.6 | 20.1 | 30.4 | 46.4 | 50.7 | 16.4 |
| EUG[33] | TIP’19 | OneEx | 55.8 | 72.3 | 83.5 | 26.2 | 48.8 | 63.4 | 68.4 | 28.5 |
| SPGAN[16] | CVPR’18 | Transfer | 58.1 | 76.0 | 82.7 | 26.7 | 46.9 | 62.6 | 68.5 | 26.4 |
| TJ-AIDL[14] | CVPR’18 | Transfer | 58.2 | 74.8 | - | 26.5 | 44.3 | 59.6 | - | 23.0 |
| ATNet[36] | ICCV’19 | Transfer | 55.7 | 73.2 | 79.4 | 25.6 | 45.1 | 59.5 | 64.2 | 24.9 |
| MAR[12] | CVPR’19 | Transfer | 67.7 | 81.9 | 87.3 | 40.0 | **67.1** | **79.8** | **84.2** | **48.0** |
| BOW[30] | ICCV’15 | **None** | 35.8 | 52.4 | 60.3 | 14.8 | 17.1 | 28.8 | 34.9 | 8.3 |
| OIM*[35] | CVPR’17 | **None** | 38.0 | 58.0 | 66.3 | 14.0 | 24.5 | 38.8 | 46.0 | 11.3 |
| BUC[19] | AAAI’19 | **None** | 66.2 | 79.6 | 84.5 | 38.3 | 47.4 | 62.6 | 68.4 | 27.5 |
| **Ours** | - | **None** | **77.5** | **88.5** | **92.2** | **50.3** | 63.2 | 75.1 | 79.2 | 38.6 |

TABLE II
Comparison with the state-of-the-art methods on two video-based datasets. The column “Labels” denotes the type of supervision used by the corresponding method. “None” represents no extra information is used. “Transfer” represents an additional dataset is used. “OneEx” represents one labeled image in per identity is used. * denotes that the results are reproduced by Lin [19].

| Methods | Venue | Labels | MARS | | | | DukeMTMC-VideoReID | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | rank-1 | rank-5 | rank-10 | mAP | rank-1 | rank-5 | rank-10 | mAP |
| DGM+IDE[37] | TIP’19 | OneEx | 36.8 | 54.0 | - | 16.8 | 42.3 | 57.9 | 69.3 | 33.6 |
| Stepwise[38] | ICCV’17 | OneEx | 41.2 | 55.5 | - | 19.6 | 56.2 | 70.3 | 79.2 | 46.7 |
| RACE[39] | ECCV’18 | OneEx | 43.2 | 57.1 | 62.1 | 24.5 | - | - | - | - |
| DAL[40] | BMVC’18 | OneEx | 49.3 | 65.9 | 72.2 | 23.0 | - | - | - | - |
| EUG[33] | TIP’19 | OneEx | 62.8 | 75.2 | 80.4 | 42.6 | 72.9 | 84.3 | 88.3 | 63.3 |
| OIM*[35] | CVPR’17 | **None** | 33.7 | 48.1 | 54.8 | 13.5 | 51.1 | 70.5 | 76.2 | 43.8 |
| BUC[19] | AAAI’19 | **None** | 61.1 | 75.1 | 80.0 | 38.0 | 69.2 | 81.1 | 85.8 | 61.9 |
| **Ours** | - | **None** | **67.5** | **79.2** | **82.6** | **44.1** | **80.2** | **92.2** | **95.1** | **73.6** |

TABLE III
The effectiveness of our proposed components. Regularization term of BUC [19] is cluster size. In IMD based experiments, if MP is not used, we use global average pooling by default.

| Methods | CPD | NL | MP | Market-1501 | | DukeMTMC-reID | | MARS | | DukeMTMC-VideoReID | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP |
| BUC w/o regularization[19] | | | | 62.9 | 33.8 | 41.3 | 22.5 | 55.5 | 31.9 | 60.7 | 50.8 |
| BUC with regularization[19] | | | | 66.2 | 38.3 | 47.4 | 27.5 | 61.1 | 38.0 | 69.2 | 61.9 |
| Ours (based IMD) | | | | 66.9 | 39.4 | 53.5 | 30.4 | 63.9 | 41.5 | 75.1 | 68.7 |
| Ours (based IMD) | ✓ | | | 70.1 | 41.9 | 56.3 | 32.3 | 64.6 | 41.9 | 76.5 | 70.9 |
| Ours (based IMD) | ✓ | ✓ | | 74.7 | 47.4 | 60.0 | 35.0 | 65.3 | 42.6 | 78.9 | 72.1 |
| Ours (based IMD) | ✓ | ✓ | ✓ | **77.5** | **50.3** | **63.2** | **38.6** | **67.5** | **44.1** | **80.2** | **73.6** |

of Table III , the experimental performance of our proposed IMD is superior to that of the minimum distance in BUC [19] on both image-based and video-based datasets. This benefits from the improvement of distance metric between clusters. It is worth noting that IMD without CPD performs better than BUC [19] with diversity regularization, as shown in row 2 and row 3 of Table 3. This demonstrates the superiority of IMD as a metric of inter-cluster distance. Moreover, the data in row 3 and row 4 of Table III show that CPD can bring performance improvement on all four datasets. This proves the importance of intra-cluster distance in clustering metric and the validity of CPD as intra-cluster distance. IMD and CPD balance the similarity and dissimilarity in clustering.

**The Effectiveness of Non-locally Enhanced Feature Network.** From row 4, row 5 and row 6 in Table III, we can see that both non-local blocks and the mixed pooling strategy can improve the experimental performance to a certain extent. This demonstrates that both the non-local blocks and the mixed pooling strategy can enhance global features. Our proposed non-locally enhanced network is more suitable for the unsupervised clustering person re-ID.

**The Impact of Pooling Strategies.** Table IV shows the performance of our proposed method under different pooling strategies on Market-1501. It can be observed that the performance of global max pooling is better than that of global average pooling. This is mainly because global average pooling considers all positions of a particular part and all positions contribute to the final feature embedding equally. Therefore, the discrimination ability of the feature embedding generated by global average pooling can be easily affected by the irrelevant background patterns. On the contrary, global max pooling preserves the largest response value of a local

TABLE IV
THE IMPACT OF DIFFERENT POOLING STRATEGIES ON MARKET-1501.

| Model | Market-1501 | | | |
|---|---|---|---|---|
| | rank-1 | rank-5 | rank-10 | mAP |
| IMD + CPD + NL + Avg pool | 73.9 | 87.1 | 90.7 | 46.2 |
| IMD + CPD + NL + Max pool | 76.4 | 87.7 | 91.6 | 48.9 |
| IMD + CPD + NL + Mixed pool | **77.5** | **88.5** | **92.2** | **50.3** |

part. Namely, it retains the most discriminative information. Global average pooling operations and global max pooling operations are complementary, which can obtain the feature of global part and the most discriminative part of features. Thus, we integrate the two pooling strategies into an unified model to make full use of their advantages. Experimental results in Table IV indicate that mixing the two pooling strategies gets a better result than using either of them.

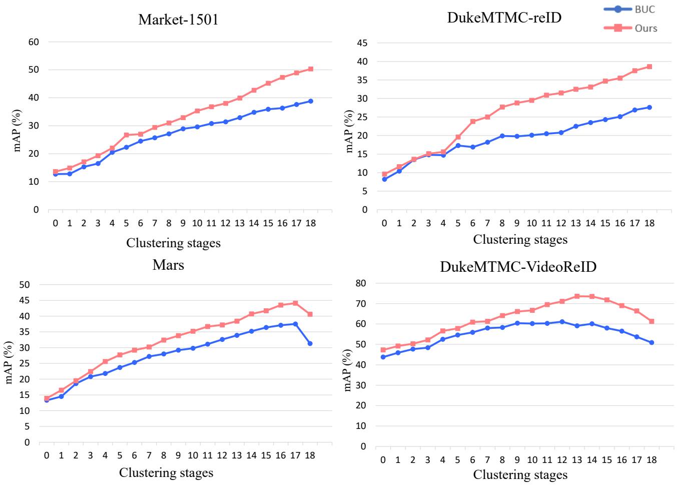


Fig. 5. The mAP performance in different clustering stages on four large datasets.

**Performance comparison of different clustering stages.** Unsupervised hierarchical clustering framework trains the model with generated pseudo labels and repeats the merging and fine-tuning until getting the best performance. The mAP performance of BUC [19] and our proposed method in different clustering stages on four datasets is shown in Fig. 5. Before achieving their respective best performance, the growth rate of mAP in our method is higher than that in BUC [19] on the whole. This is mainly because the wrong merging in early stages of BUC [19] generates false labels, which will affect the optimization direction of the model. These negative false labels have an additive effect in later stages, so they will result in poor experimental performance. Different from BUC [19], our method can effectively reduce the wrong merging of early stages.

## V. CONCLUSION

In this paper, we have presented an improved hierarchical clustering approach with non-locally enhanced features for unsupervised person re-ID. Specifically, we have proposed a new metric composing of intermediate distance (IMD) as inter-cluster distance and compactness degree (CPD) as intra-cluster distance. IMD and CPD ensure the similarity and dissimilarity of clustering, which promotes the quality of clustering and avoids negative false labels effectively. Besides, the designed non-locally enhanced feature network that aims at enhancing global features can bring performance improvement. Experimental results on four large datasets (Market-1501, DukeMTMC-reID, MARS and DukeMTMC-VideoReID) demonstrate the superiority of the proposed method, which outperforms the existing state-of-the-art unsupervised methods in terms of mAP and rank-*k*.

## ACKNOWLEDGMENT

This work was supported in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing), in part by Shenzhen International cooperative research projects GJHZ20170313150021171, and in part by NSFC-Shenzhen Robot Jointed Founding (U1613215).

## REFERENCES

[1] B. Chen, W. Deng, and J. Hu, "Mixed high-Order attention network for person re-rdentification," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 371-381, 2019.

[2] Y. Cho and K. Yoon, "Pamm: Pose-aware multi-shot matching for improving person re-identification," in *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3739C-3752, 2018.

[3] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 1179-1188, 2018.

[4] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," in *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472C-3483, 2018.

[5] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," in *IEEE Trans. Circuits Syst. Video Technol.*, 2018.

[6] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629C-1642, 2014.

[7] S. Liao, Y. Hu, X. Zhu, and S. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2197-2206, 2015.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2360-2367, 2010.

[9] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *British Machine Vision Conference*, 2014.

[10] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 3586-3593, 2013.

[11] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterativelaplacian regularization for unsupervised person re-identification," in *British Machine Vision Conference*, 2015.

[12] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, and J. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2148-2157, 2019.

[13] P. Peng, T. Xiang, Y. Wang, M. Pontil, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 1306-1315, 2016.

[14] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2275-2284, 2018.

[15] S. Bak, P. Carr, and J. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, pp.189-205, 2018.

[16] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 994-1003, 2018.

[17] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 7948-7956, 2018.

[18] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: clustering and fine-tuning," in *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 14, no. 4, p. 83, 2018.

[19] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8783-8745, 2019.

[20] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 60-65, 2005.

[21] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, pp. 650-667, 2018.

[22] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2138-2147, 2019.

[23] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis.*, pp. 172-188, 2018.

[24] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning*, pp. 478-487, 2016.

[25] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, pp. 132-149, 2018.

[26] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *International Conference on Machine Learning*, vol. 70, pp. 3861-3870. 2017.

[27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[28] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," in *Technometrics*, vol. 32, no. 1, pp. 1-C12, 1990.

[29] X. Wang, R. Girshick, A. Gupta, and K. HE, "Non-local neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 7794-7803, 2018.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1116-1124, 2015.

[31] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3754-3762, 2017.

[32] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, pp. 868-884, 2016.

[33] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," in *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872-2881, 2019.

[34] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, pp. 17-35, 2016.

[35] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 3415-3424, 2017.

[36] J. Liu, Z. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 7202-7211, 2019.

[37] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," in *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976-2990, 2019.

[38] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2429-2438, 2017.

[39] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis.*, pp. 170-186, 2018.

[40] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," in *British Machine Vision Conference*, 2018.