Article:
Miao, Y., Deng, J., Ding, G. et al. (1 more author) (2024) Confidence-guided centroids for unsupervised person re-identification. IEEE Transactions on Information Forensics and Security, 19. pp. 6471-6483. ISSN 1556-6013

https://doi.org/10.1109/TIFS.2024.3414310

Confidence-guided Centroids for Unsupervised Person Re-Identification

Yunqi Miao, Jiankang Deng, *Member, IEEE*, Guiguang Ding, *Senior Member, IEEE*, and Jungong Han, *Senior Member, IEEE*

***Abstract*—Unsupervised person re-identification (ReID) aims to train a feature extractor for identity retrieval without exploiting identity labels. Due to the blind trust in imperfect clustering results, the learning is inevitably misled by unreliable pseudo labels. Albeit the pseudo label refinement has been investigated by previous works, they generally leverage auxiliary information such as camera IDs and body part predictions. This work explores the internal characteristics of clusters to refine pseudo labels. To this end, Confidence-Guided Centroids (CGC) are proposed to provide reliable cluster-wise prototypes for feature learning. Since samples with high confidence are exclusively involved in the formation of centroids, the identity information of low-confidence samples, *i.e.,* boundary samples, are NOT likely to contribute to the corresponding centroid. Given the new centroids, the current learning scheme, where samples are forced to learn from their assigned centroids solely, is unwise. To remedy the situation, we propose to use Confidence-Guided pseudo Label (CGL), which enables samples to approach not only the originally assigned centroid but also other centroids that are potentially embedded with their identity information. Empowered by confidence-guided centroids and labels, our method yields comparable performance with, or even outperforms, state-of-the-art pseudo label refinement works that largely leverage auxiliary information.**

***Index Terms*—Person Re-identification, Unsupervised Learning, Centroid, Visual Surveillance.**

I. Introduction

PERSON re-identification (ReID), one of the fundamental tasks in intelligent visual surveillance, aims to retrieve a person of interest across multiple cameras [1]–[3]. Due to the label-free training manner, unsupervised person ReID methods have attracted increasing attention. Unsupervised ReID methods can be broadly categorized into two types: unsupervised domain adaptation (UDA) methods [4]–[8] and purely unsupervised learning (USL) methods [9]–[13]. The former pre-trains a model on person-related datasets, *i.e.,* source domain, and fine-tunes it on ReID-related datasets, *i.e.,* target domain. Apart from requiring additional annotated labels, UDA methods are vulnerable to the large gap between the source domain and the target domain. In contrast, USL methods do not require any labeled data for training, which are more challenging but well fit real-world scenarios. In the paper, we focus on USL methods.

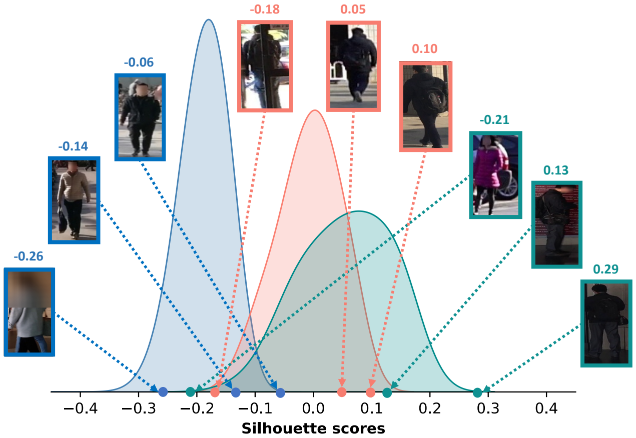


Fig. 1: Training samples (cluster ID = 1) and their silhouette scores at epoch 0 (blue), epoch 25 (orange), and epoch 50 (green) on MSMT17 [14]. Higher silhouette scores denote samples are clustered at higher confidence. **Best viewed in color**.

Existing USL methods generally follow a two-stage training scheme: 1) clustering, *i.e.,* obtaining the pseudo labels via a clustering algorithm such as DBSCAN [15], and 2) network training, *i.e.,* optimizing the network in a “supervised” manner with assigned cluster IDs. Contrastive loss such as InfoNCE [4] or ClusterNCE [9] usually serves as training objectives. Due to the blind trust in imperfect clustering results, the learning is inevitably misled by unreliable pseudo labels, where multiple identities are merged into one cluster or samples of one person are assigned to multiple clusters. Despite that some pseudo label refinement [10]–[13], [16] have been proposed, they generally leverage auxiliary information, such as camera IDs [10], [12], body part predictions [11], or are facilitated by generated samples [13]. In the paper, we aim to refine pseudo labels by merely exploiting internal characteristics within samples, *i.e.,* the sample-wise clustering confidence, which appears to be more valuable.

To measure the sample-wise clustering confidence, *i.e.,* how

Manuscript received xxx; revised xxx, xxx and xxx, xxx; accepted xxx, xxx. This work was supported by the National Natural Science Foundation of China No.61925107, by the State Key Laboratory of Reliability and Intelligence of Electrical Equipment No.EERIKF2022005, Hebei University of Technology. (Corresponding author: Jungong Han)

Yunqi Miao is with the Warwick Manufacturing Group (WMG), University of Warwick, Coventry, CV4 7AL, United Kingdom (e-mail: Yunqi.Miao.1@warwick.ac.uk).

Jiankang Deng is with the Department of Computing, Imperial College London, London, SW7 2AZ, United Kingdom (e-mail: jiankangdeng@gmail.com).

Guiguang Ding is with the School of Software, Tsinghua University, Beijing, 100084, China (e-mail: dinggg@tsinghua.edu.cn).

Jundong Han is with the Department of Computer Science, University of Sheffield, Sheffield, S10 2TN, United Kingdom (e-mail: jungonghan77@gmail.com).

well a sample fits its cluster, we employ a metric: silhouette score [17]. The score presents the ratio between intra-cluster distance and inter-cluster distance, which ranges from -1 to +1 (*higher is better*). To demonstrate the relationship between the clustering confidence and the silhouette score, we visualize silhouette scores of training samples of MSMT17 [14] in Fig. 1. Samples are from the same cluster (cluster ID=1) but at different training epochs, *i.e.,* 0, 25, and 50, respectively. As training goes on, the clustering is gradually enhanced by involving more effective features and a more discriminative network. At first, images are grouped by coarse visual features, yet by identity-related information in the end. Meanwhile, sample-wise silhouette scores continuously shift towards higher values during training. Given this consistency, a conclusion can be drawn that, *a higher silhouette score implies the sample better fits its cluster, i.e., being clustered at higher confidence.* Previous learning schemes [9], [13] adopt all-sample based centroids, which are obtained by averaging features of all samples within the cluster, and enforce instances to approach such centroids. However, our observation suggests that low-confidence samples either are poor in quality or belong to other identities. Features of such images will inevitably contaminate centroids regardless of the training stage. In light of this, we propose Confidence-Guided Centroids (CGC) to provide more reliable cluster-wise prototypes for feature learning.

Although the reliability of cluster centroids has been improved, the conventional one-hot labeling strategy aggravates a problem. Since high-confidence samples exclusively contribute to the formation of cluster centroids, the identity-related information of low-confidence samples can hardly be presented in the assigned centroid. To illustrate the problem, an analysis is conducted on MSMT17 [14], where we intend to investigate how much identity information of low-confidence samples can be presented in their assigned centroids. We found that, with the vanilla all-sample based cluster centroids, only 5.83% low-confidence samples have their identity information embedded in the assigned centroid at the beginning. Although the ratio gradually climbs to 17.19%, a large proportion of low-confidence samples (over 80%) still are pushed to "wrong" centroids. Unfortunately, the ratio achieves 14.17% at most with confidence-guided centroids. Given the situation, the one-hot labeling strategy, which enforces samples to learn from the assigned centroid solely, is unwise. To address the problem, we propose to use confidence-guided pseudo labels (CGL), which encourages instances to approach not only the assigned confidence-guided centroid but also others where their identity information is potentially embedded.

In summary, our contributions are as follows:

1) We propose Confidence-Guided Centroids (CGC) to provide cluster-wise prototypes for feature learning. The reliability of centroids is improved via filtering out low-confidence samples during formation.
2) To overcome the problem that the identity information of low-confidence samples is rarely presented in their assigned centroids, we propose to use confidence-guided pseudo labels (CGL) during training. Apart from the originally assigned centroid, instances are also encouraged to approach other centroids where their identity information are potentially embedded.
3) The proposed method only exploits internal characteristics for unsupervised person re-identification. Extensive experiments on benchmark datasets demonstrate that our method yields better or comparable performances with state-of-the-art ones that largely leverage auxiliary information.

II. Related work

A. Unsupervised Person ReID.

The existing unsupervised person ReID methods are divided into two categories: a) Unsupervised Domain Adaptation (UDA) methods, and b) purely UnSupervised Learning (USL) methods.

UDA methods boost the performance by transferring the knowledge learned from the labeled source domain to the unlabeled target one [4]–[7], [18]. To bridge the gap between source and target domain, IDM [5] generates intermediate domains' representations based on the "shortest geodesic path". SECRET [6] maximizes the consistency between pseudo labels generated by different domains. MET [7] eliminates the noise via the aggregated information from multiple viewpoints. Meanwhile, MET improves the temporal consistency between clustering at different iterations. UST [18] presents a loss function with only one hyper-parameter for UDA person ReID. Instead of aggregating the information, UST aggregates weights from different training iterations to create the final model to better exploit the knowledge throughout the adaptation.

Since UDA methods are highly prone to the large gap between the source domain and the target domain, they are hardly applicable to real-world scenarios [1], [19].

USL methods do not require any identity labels during training [9]–[11], [13], [16], [20], [21]. Instead, they exploit pseudo labels as guidance. Pseudo labels can be generated either by the image similarity [20], [22] or clustering algorithms [9], [13], [23], [24]. Specifically, SSL [20] and MMCT [22] formulate unsupervised person ReID as a classification task and predict pseudo labels based on the image similarity. In terms of clustering-based methods, BUC [23] and HCT [24] employ the bottom-up clustering scheme to gradually merge similar individual samples into clusters. Recently, Cluster-Contrast [9] adopts a contrastive learning scheme, which initializes, updates, and performs contrastive loss computation at the cluster level. However, clustering-based methods are generally sensitive to the pseudo label noise brought by imperfect clustering results.

B. Noise Reduction of Pseudo Label.

Recently, how to handle noise pseudo labels in clustering-based methods has become a research hotspot. Generally, previous unsupervised person ReID methods refine the pseudo labels by auxiliary information or additional generated samples. Specifically, SpCL [4] employs a self-paced learning scheme to gradually obtain more reliable clusters for the pseudo label refinement. JVTC [25], RLCC [16] and OPLG-HCD [26] leverage temporal information to refine visual similarity based pseudo labels by encouraging the consistency between clustering results of two consecutive iterations. Additionally, CAP [12], IICS [27] and O2CAP [28] split each cluster into multiple proxies according to camera IDs. By applying feature learning constraints on such camera-aware proxies, the pseudo label noise brought by varying viewing points can be eliminated. Apart from generating pseudo labels via the global clustering, PPSL [29] constructs patch surrogate classes, which are then assigned to a pair of person images of different camera views via gradient-guided similarity separation. Moreover, the identity prediction consistency of different body parts is also employed to refine pseudo labels. For example, PPLR [11] employs the complementary relationship between reliable features of human global and body parts for the pseudo label refinement. Instead of building upon a single backbone network, ESSL [30] ensembles feature outputs by multiple backbones to improve the reliability of clustering, thereby the robustness of learned features.

Apart from the auxiliary information, some works focus on how to make full use of the training data [31], [32] or refine the original cluster centroids [33]. GSAM [31] suggests gathering samples of the same class into groups during the training, which alleviates the negative impact brought by individual samples. HDCLR [32] makes full use of outlier instances by enhancing self-supervised signals from both instances' self-contrastive level and probability distillation respectively. RTMem [33] proposes to use a randomly sampled instance within the mini-batch to update cluster centroids.

Moreover, some works suggest generating additional samples to reduce the pseudo label noise. Specifically, GCL [34] generates extra images towards different views for a person and GCL+ [35] further augments id-related features. Then view-invariant identity features can be disentangled by enforcing the original view and the generated ones to share the same identity representations.

The contrastive feature learning scheme is also adopted by later works [10], [36], [37]. ICE [10] alleviates the label noise by enhancing the consistency between augmented and original instances. GRACL [36] sets up two proxies for each cluster to capture inter- and intra-ID relations respectively, where samples are enforced to approach the positive proxy via relation-aware contrastive learning modules. Instead of a single label, AdaMG [37] assigns each sample with a group labels to capture complementary and diverse features through clustering. Recently, ISE [13] generates boundary samples from a given sample and their neighboring clusters. The discriminability of the network is improved by enforcing generated samples to be correctly classified.

Unlike the above methods, this work explores whether internal characteristics can facilitate pseudo label refinement. Although a previous work, CACL [38], improves the effectiveness of features by suppressing an internal characteristic - color, it is applied at the image level. In the paper, we investigate the sample-wise clustering confidence, which describes how well a sample fits its cluster at the feature level. With such a criterion, better cluster centroids and pseudo labels can be obtained for feature learning.

III. Methodology

A. Problem Statement

Let $\mathcal{T} = \{x_i\}_{i=1}^{N}$ denote an unlabeled training dataset, where x_i represents i-th image and N is the number of images. The USL ReID task aims to train a feature extractor E_θ in an unsupervised manner, where ReID features $\mathcal{F} = \{f_i\}_{i=1}^{N}$ are derived. The identity retrieval during inference is based on such ReID features. The training scheme of clustering-based USL methods [4], [9], [12], [13] alternates between two stages:
Stage I: Clustering. At the beginning of each epoch, training samples are clustered by DBSCAN [15]. Cluster IDs $y_i \in \{1, ..., C\}$ serve as one-hot pseudo labels for the network optimization. Meanwhile, based on clustering results, a cluster-based memory bank $\mathcal{M} = \{m_i\}_{i=1}^{C}$ is initialized by cluster centroids that are formulated as,

$$m_i = \frac{1}{|\mathcal{C}|} \sum_{f_i \in \mathcal{C}} f_i, \tag{1}$$

where f_i represents the feature of i-th sample in the cluster $\mathcal{C}$, and $|\mathcal{C}|$ denotes the cluster size.
Stage II: Network Training. With the obtained pseudo labels, the network is then optimized in a "supervised" manner with the training objective, *i.e.,* ClusterNCE [9], which is formulated as,

$$\mathcal{L} = -\log \frac{\exp(\Phi(f \cdot m_+)/\tau)}{\sum_{j=1}^{C} \exp(\Phi(f \cdot m_j)/\tau)}, \tag{2}$$

where m_+ refers to the centroid of the cluster that f belongs to, m_j represents j-th centroid in the memory bank, $\Phi(u \cdot v)$ represents the cosine similarity between vector u and vector v, and τ is the temperature parameter. The memory bank is updated in a momentum manner [9] as,

$$m_i \leftarrow \mu \cdot m_i + (1 - \mu) \cdot f, \tag{3}$$

where μ is the updating factor and f refers to the feature of instance belonging to i-th cluster in the current mini-batch.

In this paper, we follow the framework of iterative clustering and network training. However, our method, as illustrated in Fig. 2, differs from previous works mainly in two aspects: 1) cluster centroids. Instead of using all samples to calculate the centroids, we adopt confidence-guided centroids (CGC) to provide reliable cluster-wise prototypes for feature learning (Sec. III-C), and 2) pseudo labels. Apart from the assigned centroid, our confidence-guided pseudo labels (CGL) encourages instances to approach other centroids where their identity information is potentially embedded (Sec. III-D). Note that clusters consisting of a single point, *i.e.,* outliers, are not involved during training.

B. Silhouette Score

To describe the sample-wise clustering confidence, *i.e.,* how well a sample fits its cluster, we employ a metric named silhouette score [17]. The score simultaneously considers two key factors of clustering, *i.e.,* tightness and separation.

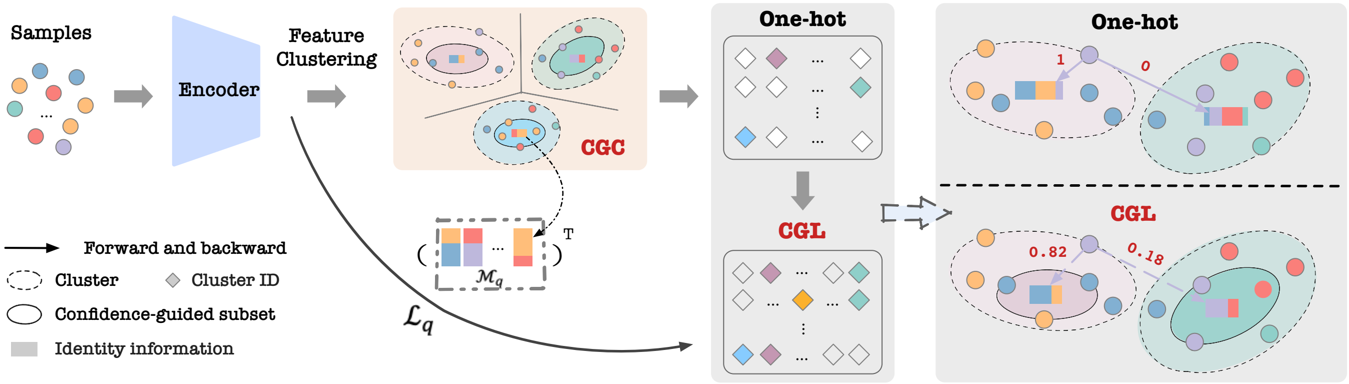


Fig. 2: **Framework of the proposed method**. At the beginning of each epoch, training samples are clustered by DBSCAN [15]. Based on the original clustering result, we select a confidence-guided subset to build our confidence-guided centroids (CGC). During optimization, samples are encouraged to approach not only the assigned centroid but others where their identity information is potentially embedded via our confidence-guided pseudo labels (CGL). **Best viewed in color**.

Formally, for i-th data point in cluster $\mathcal{C}_I$, its average distance to other data points within the cluster can be calculated as,

$$a_i = \frac{1}{|\mathcal{C}_I|} \sum_{i,j \in \mathcal{C}_I, i \neq j} d(i,j), \tag{4}$$

where $d(i,j)$ refers to the distance between i-th and j-th data points and $|\mathcal{C}_I|$ represents the cluster size. Similarly, the distance between i-th data point and samples belonging to its nearest neighboring cluster $\mathcal{C}_J$ can be denoted as,

$$b_i = \min_{J \neq I} \frac{1}{|\mathcal{C}_J|} \sum_{j \in \mathcal{C}_J} d(i,j). \tag{5}$$

Given the intra-class distance a_i and the minimal inter-class distance b_i, the silhouette score s_i is formulated as,

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}. \tag{6}$$

The silhouette score ranges from $[-1, 1]$. If an instance has a higher silhouette score, it has a smaller intra-class distance and a large inter-class distance. In other words, it is clustered at a higher confidence [17].

C. Confidence-guided Centroids

Based on the observation that images with lower silhouette scores (confidence) generally contain high uncertainty regarding person identity, previous all-sample based cluster centroids are undoubtedly unwise. To remedy the problem, we build confidence-guided centroids (CGC) with high-confidence images only.

Specifically, the confidence-guided centroid of i-th cluster m_i can be formulated as,

$$m_i = \frac{1}{|\mathcal{C}_q|} \sum_{f_i \in \mathcal{C}_q} f_i, \quad \mathcal{C}_q = \{f_i \in \mathcal{C} | s_i > \delta\}, \tag{7}$$

where a confidence-guided subset $\mathcal{C}_q$ is selected from the original cluster $\mathcal{C}$ by a silhouette score threshold δ. All confidence-guided centroids are then stored in a confidence-guided memory bank $\mathcal{M}_q = \{m_i\}_{i=1}^{C}$ for network optimization.

According to Fig. 1, our confidence-guided centroids can filter out images that are poor in quality or with cluttered backgrounds at early stages. While, at later stages, such centroids effectively exclude some low-confidence samples that possibly belong to other identities. In summary, the proposed confidence-guided centroids can provide more reliable cluster-wise prototypes for feature learning.

D. Confidence-guided Pseudo Labels

Another problem of the clustering-based USL methods is that samples, especially low-confidence ones, very likely carry different identity information with their assigned centroids. Our confidence-guided centroids also confronts the problem since only high-confidence samples are included in the formation of centroids, as illustrated in Fig. 2. Given the situation, the previous learning scheme, which enforces samples to approach their assigned centroids solely regardless of the identity consistency in-between, is unwise. To alleviate the problem, we propose to use confidence-guided pseudo labels (CGL). Such labeling encourages samples to approach not only the assigned centroid but other centroids where their identity information is potentially embedded.

Specifically, we build a distance matrix $\mathcal{D} \in \mathbf{R}^{N \times C}$, where N and C denote the number of samples and clusters at the current epoch, respectively. In the paper, clusters consisting of one sample are ignored [9]. As normalized identity features and centroids are adopted, $\mathcal{D}(i,j)$ represents the cosine distance between i-th sample and j-th confidence-guided centroid. Since similar samples are more likely to be scattered in neighboring clusters [13], the identity information of boundary samples is probably embedded in neighboring centroids. Therefore, when setting the learning target for samples, neighboring centroids should be assigned with higher confidence while distanced ones should be given lower confidence. To this end, a confidence matrix $\mathcal{P} \in \mathbf{R}^{N \times C}$ is obtained by,

$$\mathcal{P}(i,j) = \frac{p_{i,j}}{\sum_{j=1}^{C} p_{i,j}}, \quad p_{i,j} = \sigma(-\mathcal{D}(i,j)), \tag{8}$$

Algorithm 1: Pipeline of our method

```
Require: Unlabeled data with pseudo labels
    $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \ldots, C\}$
Require: Initialize the backbone encoder $E_\theta$
Require: Threshold $\delta$ for Eq. (7)
Require: Coefficient $\beta$ for Eq. (9)
for $n$ in $[1, epoch\_num]$ do
    Extracting features $\mathcal{F}$ by $E_\theta$
    Clustering $\mathcal{F}$ into $C$ clusters with DBSCAN
    Building CGC dictionary $\mathcal{M}_q$ by Eq. (7)
    for $m$ in $[1, iteration\_num]$ do
        Sampling a mini-batch from $\mathcal{T}$
        Computing CGL with Eq. (9)
        Computing loss with Eq. (10)
        Updating encoder $E_\theta$
        Updating centroids with Eq. (3)
    end
end
```

where $\mathcal{P}(i, j)$ represents the confidence of j-th centroid given by i-th sample, $\sum_{j=1}^{C} \mathcal{P}(i, j) = 1$, and $\sigma(\cdot)$ is the Sigmoid function. By integrating the confidence matrix with the originally assigned one-hot pseudo label y_i, the confidence-guided pseudo label of i-th sample $\tilde{y}_i$ can be formulated as,

$$\tilde{y}_i = \beta \cdot y_i + (1 - \beta) \cdot \mathcal{P}(i, \cdot), \tag{9}$$

where $\mathcal{P}(i, \cdot)$ indicates i-th row of confidence matrix $\mathcal{P}$, and $\beta \in [0, 1]$ is the coefficient for the pseudo label refinement.

According to a previous work [39], the training objective, *i.e.,* ClusterNCE, can be considered as a non-parametric classifier, where centroids stored in the memory bank serve as the weight matrix of the classification layer. Therefore, the training objective of our method can be rewritten as,

$$\mathcal{L}_q = \frac{1}{N} \sum_{i=1}^{N} \Big[\ell_{ce}\big(\mathcal{M}_q^T f_i, \tilde{y}_i\big) \Big], \tag{10}$$

where ℓ_{ce} refers to the cross-entropy loss. Compared to Eq. (2), the training objective of our method can be obtained by simply applying two modifications: 1) replacing the original $\mathcal{M}$ with our confidence-guided memory bank $\mathcal{M}_q$, and 2) replacing the one-hot pseudo label y_i with our confidence-guided one $\tilde{y}_i$. The training details are presented in Algorithm 1.

IV. Experiment

A. Datasets and Evaluation Protocol

Datasets. We evaluate our proposed method on the Market-1501 [40] and MSMT17 [14] datasets.

Market-1501 includes 32,668 images of 1,501 identities captured by 6 cameras. Among them, 12,936 images of 751 identities are used for training while the resting 19,732 images of 750 identities form the test set.

MSMT17 contains 126,441 images from 4,101 identities captured by 15 cameras. The training set is composed of 32,621 images of 1,041 identities and the test set consists of 93,820 images of 3,060 identities. MSMT17 is more challenging due to the diversity in backgrounds, illuminations, poses, and occlusions.

Evaluation Protocol. Following previous methods [4], [9], [10], [13], the mean average precision (mAP) [41] and the cumulative matching characteristic (CMC) [40] top-1, top-5, top-10 accuracies are adopted as evaluation metrics. Note that, there are no post-processing operations, such as reranking [42], during inference.

B. Implementation Details

Following previous works [9], [13], we adopt ResNet-50 [43] pre-trained on ImageNet [44] as our backbone feature encoder. All layers after layer-4 are replaced by a generalized mean pooling (GeM) [45] layer followed by the batch normalization layer [46]. The output 2048-dimensional ReID features are firstly normalized and then used for identity retrieval during inference. Our framework is built upon a state-of-the-art USL method [9]. For a fair comparison, we follow all experimental settings except for the formation of cluster centroids and the training objectives, as described in Sec. III. The coefficient β in Eq. (9) is empirically set as 0.8 to achieve optimal performances.

During training, input images are resized to 256×128. We adopt random flipping, cropping, and erasing [47] as data augmentation. Each mini-batch is formed by 16 identities, each with 16 images. Both identity and images are randomly selected from the training set. For the optimization, we adopt Adam [48] optimizer with a weight decay of 0.0005. The learning rate is set to 3.5×10^{-4} initially, and is divided by 10 every 30 epochs. We train for a total of 70 epochs on Market-1501 [40], and 50 on MSMT17 [14].

C. Comparison with State-of-the-art Methods

We compare our method with state-of-the-art (SOTA) unsupervised person ReID methods in Table I. Since our method can be an add-on to boost the performance of previous clustering-based USL methods, we adapted Cluster-Contrast [9] as our baseline following ISE [13]. As can be seen, the proposed strategies, *i.e.,* CGC and CGL, improve the mAP/top-1 accuracy by +2.9% / +1.7% on Market-1501 and +3.2% / +2.2% on MSMT17. Additionally, ISE [13] proposes to reduce the pseudo label noise of Cluster-Contrast [9] by leveraging generated samples in latent space, which is orthogonal to our method. Therefore, we apply our method to ISE, which further improves the performance by achieving 85.6% mAP and 94.3% top-1 accuracy on Market-1501 and 35.7% mAP and 66.1% top-1 accuracy on MSMT17.

Moreover, the performance can be further boosted by leveraging auxiliary information on top of the proposed strategies. Taking the camera information as an example, we compare our method with USL methods trained with camera labels: IICS [27], CAP [12], ICE [10], and PPLR [11]. Following [10], [12], we embed camera labels into confidence-guided centroids, *i.e.,* computing the centroid with features that belong to the same cluster as well as the same camera ID. As shown in Table I, with camera-aware cluster centroids, mAP improves by +2.2% on Market-1501 and +10.1% on MSMT17.

Method	Reference	Market-1501				MSMT17			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
USL methods without any labels									
BUC [23]	AAAI'19	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4
SSL [20]	CVPR'20	37.8	71.7	83.8	87.4	-	-	-	-
MMCL [22]	CVPR'20	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8
HCT [24]	CVPR'20	56.4	80.0	91.6	95.2	-	-	-	-
SpCL [4]	NeurIPS'20	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61.2
JVTC [25]	ECCV'20	41.8	72.9	84.2	88.7	15.1	39.0	50.9	56.8
GCL [34]	CVPR'21	63.4	83.7	91.6	94.3	18.0	41.6	53.2	58.4
RLCC [16]	CVPR'21	77.7	90.8	96.3	97.5	27.9	56.5	68.4	73.1
OPLG-HCD [26]	ICCV'21	78.1	91.1	96.4	97.7	26.9	53.7	65.3	70.2
ICE [10]	ICCV'21	79.5	92.0	97.0	98.1	29.8	59.0	71.7	77.0
PPLR [11]	CVPR'22	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
ISE [13]	CVPR'22	84.7	94.0	97.8	98.8	35.0	64.7	75.5	79.4
GSAM [31]	TIP'22	79.2	92.3	96.6	97.8	24.6	56.2	67.3	71.5
CACL [38]	TIP'22	80.9	92.7	97.4	98.5	23.0	48.9	61.2	66.4
GRACL [36]	TCSVT'22	83.7	93.2	97.6	98.6	34.4	64.0	75.0	79.3
AdaMG [37]	TCSVT'23	84.6	93.9	97.9	98.9	38.0	66.3	76.9	80.6
RTMem [33]	TIP'23	83.0	92.8	97.4	98.5	32.8	57.1	70.0	74.9
ESSL [30]	TIFS'23	83.4	92.9	97.1	97.8	**42.6**	**68.2**	**77.9**	**81.4**
Cluster-Contrast [9] (Baseline)	ACCV'22	82.4	92.5	96.9	98.0	31.4	61.2	72.5	76.9
Baseline+RM	-	83.3	93.0	97.1	98.0	32.8	62.4	73.6	78.1
Baseline+WS	-	83.4	93.0	97.2	98.1	32.5	62.3	73.3	77.5
Baseline+CGC	-	84.1	93.1	97.2	98.2	34.1	63.1	75.0	79.0
Baseline+CGL	-	83.4	93.2	97.1	98.2	33.7	62.5	73.9	78.4
Baseline+CGL+CGC (Ours)	-	85.3	94.2	97.6	98.5	34.6	63.4	74.6	79.3
ISE+CGL+CGC (Ours)	-	**85.6**	**94.3**	**97.9**	**98.9**	35.7	66.1	76.3	80.0
USL methods with camera labels									
IICS [27]	CVPR'21	72.1	88.8	95.3	96.9	18.6	45.7	57.7	62.8
CAP [12]	AAAI'21	79.2	91.4	96.3	97.7	36.9	67.4	78.0	81.4
ICE [10]	ICCV'21	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4
PPLR [11]	CVPR'22	84.4	94.3	97.8	98.6	42.2	73.3	83.5	86.5
O2CAP [28]	TIP'22	82.7	92.5	96.9	98.0	42.4	72.0	81.9	85.4
PPSL [29]	TIP'22	82.3	94.1	97.4	98.8	43.1	73.2	89.4	90.8
Baseline+Ours†	-	**87.5**	**95.6**	**98.2**	**98.9**	**44.7**	**75.8**	**85.4**	**87.9**

TABLE I: Comparison of ReID methods on Market-1501 and MSMT17. The best USL results WITHOUT and WITH camera information are marked in **red** and **blue**, respectively. † indicates using the additional camera knowledge.

As stated in Sec. II, existing SOTA methods generally leverage auxiliary information or extra samples to refine pseudo labels. For example, JVTC [25], RLCC [16] and OPLG-HCD [26] leverage the temporal information, CAP [12] and IICS [27] leverage the camera IDs, PPLR [11] employs the predictions of body parts, and GCL [34] as well as ISE [13] leverages generated samples. As a departure from the above methods, our method yields SOTA performances by exclusively involving internal characteristics, *i.e.,* the sample-wise clustering confidence.

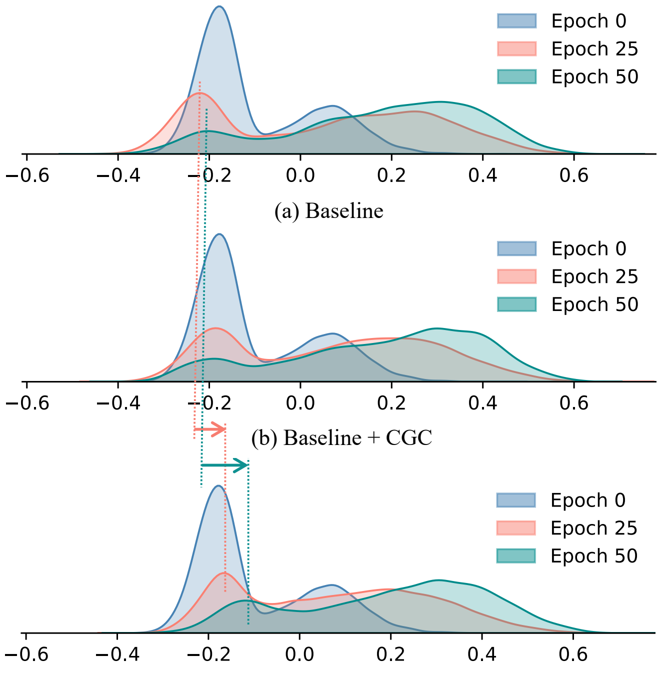
Note that AdaMG [37] and ESSL [30] outperform our method on MSMT. It is important to highlight that AdaMG utilizes three memory modules, and ESSL combines features from multiple backbones, whereas our method uses a single backbone network only. From the perspective of resource consumption, our method appears to be significantly more practical. Additionally, both AdaMG and ESSL take advantage of multiple clustering results with different hype-parameter settings, whereas ours is built upon a one-shot clustering result only. In our future work, we will integrate our method into these two methods for evaluation once their source codes are released. This will provide a comprehensive understanding of how our approach performs in conjunction with these advanced methods.

D. Ablation Study

In this section, we thoroughly analyze the effectiveness of the proposed strategies, *i.e.,*confidence-guided centroids (CGC) and confidence-guided pseudo labels (CGL).

Effectiveness of CGC. To better understand how our confidence-guided centroids benefit feature learning, we analyze how the sample-wise confidence varies throughout the training process on MSMT17. Specifically, we visualize the distribution of silhouette scores at different epochs in Fig. 3. Note that scores of outliers are excluded. Several conclusions can be drawn from the comparison between Fig. 3(a) and Fig. 3(b). 1) As training goes on, the number of valid samples gradually increases, representing as larger areas under the curve. 2) Starting from the same point (epoch 0), with our confidence-guided centroids, a noticeable shift towards higher scores can be found at epoch 25. The shift implies that CGC can effectively reduce the overall number of low-confidence samples while enhancing high-confidence ones. 3) The advantage remains until the end of training. At epoch 50, the number of high-confidence samples increases, representing by a higher peak closer to 0.4.

To better demonstrate the effectiveness of CGC, we compare models trained with the vanilla all-sample based cluster centroids ("Baseline"), with cluster centroids being built by excluding 10% samples at the cluster bound-



(a) Baseline

(b) Baseline + CGC

(c) Baseline + CGC + CGL

Fig. 3: Silhouette scores of valid samples (MSMT17 [14]) at different epochs. Comparisons are conducted between (a) baseline model, (b) baseline model with confidence-guided centroids (CGC), and (c) baseline model with CGC and confidence-guided pseudo labels (CGL). Score shifts are indicated by arrows. **Best viewed in color**.

ary ("Baseline+RM"), with centroids being built by score-weighted within-cluster samples ("Baseline+WS"), and with centroids being built by our confidence-guided strategy ("Baseline+CGC"). The performances are reported in Table I. As can be seen, centroids built with both less boundary samples and score-weighted samples are beneficial to performance boosting. For example, mAP is improved by 0.9% and 1.0% on Market-1501, respectively. Such improvements can also be found in MSMT17. However, the improvement cannot compare with ours (+1.7% / +2.7% on Market-1501/ MSMT17) as the proposed CGC considers the distance between samples and not only their assigned centroid but also those potential ones. The ablation study on CGC reveals the potential of the clustering confidence in the pseudo label refinement.

Effectiveness of CGL. We also compare the baseline model ("Baseline") and the model trained with confidence-guided pseudo labels ("Baseline+CGL"). The performances are shown in Table I. As can be seen, CGL improves mAP / top-1 accuracy by +1.0% / +0.7% on Market-1501 and +2.3% / +1.3% on MSMT17. When both CGC and CGL are employed during training, the improvements achieve +2.9% / +1.7% on Market-1501, and +3.2% / +2.2% on MSMT17.

In terms of the sample-wise clustering confidence, we visualize the distribution of silhouette scores in Fig. 3(c), when CGL is applied during training. Compared to the model trained without CGL (Fig. 3(b)), CGL further pushes the score towards a higher value at both epoch 25 and epoch 50. Fewer low-confidence samples during training imply that our CGL contributes to better clustering. In summary, the above qualitative and quantitative results prove the proposed scheme can boost performance by enhancing the sample-wise clustering confidence.

Method	Strategy	δ	Market-1501		MSMT17	
			mAP	top-1	mAP	top-1
Baseline	-	-	82.4	92.5	31.4	61.2
Ours	Linear	-	**85.3**	**94.2**	33.6	63.0
	Dynamic	-	84.9	93.9	33.0	62.8
	Constant	-0.1	83.5	93.4	32.7	62.8
		0	84.9	94.0	**34.6**	**63.4**
		0.1	84.0	93.3	34.0	63.2

TABLE II: Comparison of threshold selection strategies of confidence-guided centroids (CGC) on benchmark datasets.

E. *Parameter Analysis*

Threshold δ in CGC. To obtain the optimal threshold δ in Eq. (7) for the proposed confidence-guided centroids (CGC), three types of threshold selection strategies are explored, *i.e.,* linear, dynamic and constant, respectively. For the former two strategies, the threshold gradually increases as training goes on. The constant strategy employs a fixed threshold throughout the training process.

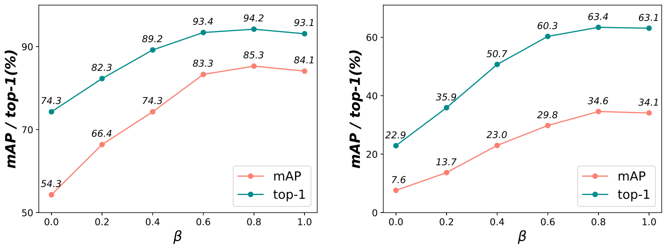
Specifically, the linear strategy updates the threshold by $\delta_t = \delta_0 * t/T + \epsilon$, where δ_0 limits the range of threshold and ϵ is the offset. In the paper, we set $\delta_0 = 0.2$ and $\epsilon = -0.1$. t and T denote the current epoch and the overall number of epochs, respectively. In terms of the dynamic strategy, the threshold is updated by $\delta = \delta_0 * tanh(0.1 * (t - T/2))$. We set $\delta_0 = 0.1$ to achieve $\delta \in [-0.1, 0.1]$, which is the same as the linear strategy. The range is set empirically in consideration of the image quality and the distribution of silhouette scores (see Fig. 3). Apart from the varying threshold, we conduct the constant strategy by fixing the threshold as $\{-0.1, 0, 0.1\}$ respectively. Comparisons between model performances with different strategies are reported in Table II. As can be observed, our method boosts the performance consistently with different scheduling schemes, which proves its robustness to different types of scheduling schemes. The best performance is achieved when adopting the linear strategy for Market-1501 and applying a fixed threshold $\delta = 0$ on MSMT17. The optimal settings are employed throughout all experiments.

Coefficient β in CGL. To analyze the impact of the coefficient β in the proposed confidence-guided pseudo labels (CGL), we tune the value of parameter β from 0 to 1 while keeping others fixed. According to Eq. (9), when β is set to 0 or 1, our method decomposes down to using the confidence matrix or the one-hot pseudo label exclusively during training. The results on these two benchmarks are illustrated in Fig. 4. As shown, as β increases from 0 to 0.8, both mAP and top-1 accuracy increase. A slight performance drop can be found when increasing β from 0.8 to 1. To achieve the best performance, we set $\beta = 0.8$ for all experiments.

To search for optimal hyper-parameters in real-world unsupervised scenarios, one approach is to randomly select and annotate a small portion (approximately $5\% \sim 10\%$) of the training data, serving as the validation set. Subsequently, appropriate scheduling schemes and hyper-parameters can be set empirically determined by monitoring performance on this labeled validation set. For example, in our validation experiment on MSMT17, we randomly select three subsets (val-1, val-2, and val-3) to mine optimal hyper-parameters, where each contains 3000 images with identity labels provided by the dataset. For each validation set, we train the model with the rest training set with different hyper-parameters and evaluate the model on the validation set. The performance in terms of mAP is illustrated in Fig. 5. As can be observed, despite the fluctuation for different validation sets, the model achieves the best performance when setting $\delta = 0$ and $\beta = 0.8$.

F. Visualization Results

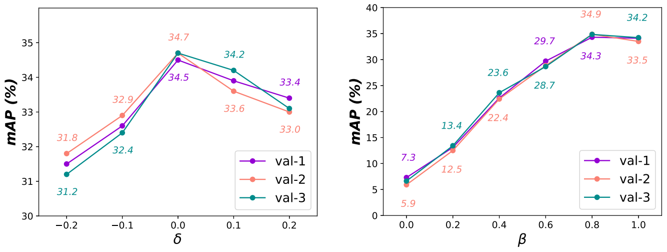
Identity Feature Distribution. To better understand the advantages of the proposed strategies, we visualize the distribution of identity features via t-SNE [49]. Specifically, 20 identities are randomly selected from Market-1501 [40] and MSMT17 [14], respectively. Features of selected identities are extracted by the baseline model and our model is trained with confidence-guided centroids (CGC) and confidence-guided pseudo labels (CGL). The distribution of identity features is illustrated in Fig. 6. As can be seen, due to the vast variety in camera views, backgrounds, and poses, the feature distribution of MSMT17 is more chaotic than that of Market-1501. Despite such challenges, with the aid of the proposed strategies, features of the same identity are distributed more compactly while features of different identities are further separated.



(a) Market-1501

(b) MSMT17

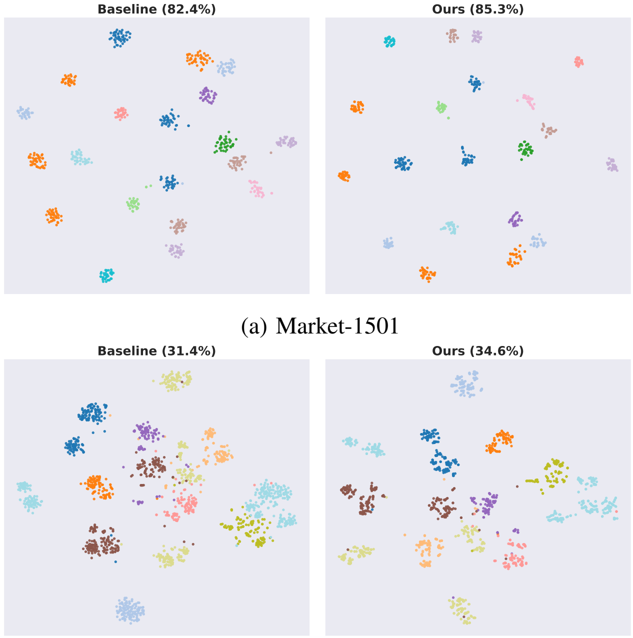
Fig. 4: Comparison of coefficient β in confidence-guided pseudo labels (CGL) on (a) Market-1501 and (b) MSMT17.



(a) δ

(b) β

Fig. 5: Performance on validation sets of MSMT17 in terms of (a) δ and (b) β.



(a) Market-1501

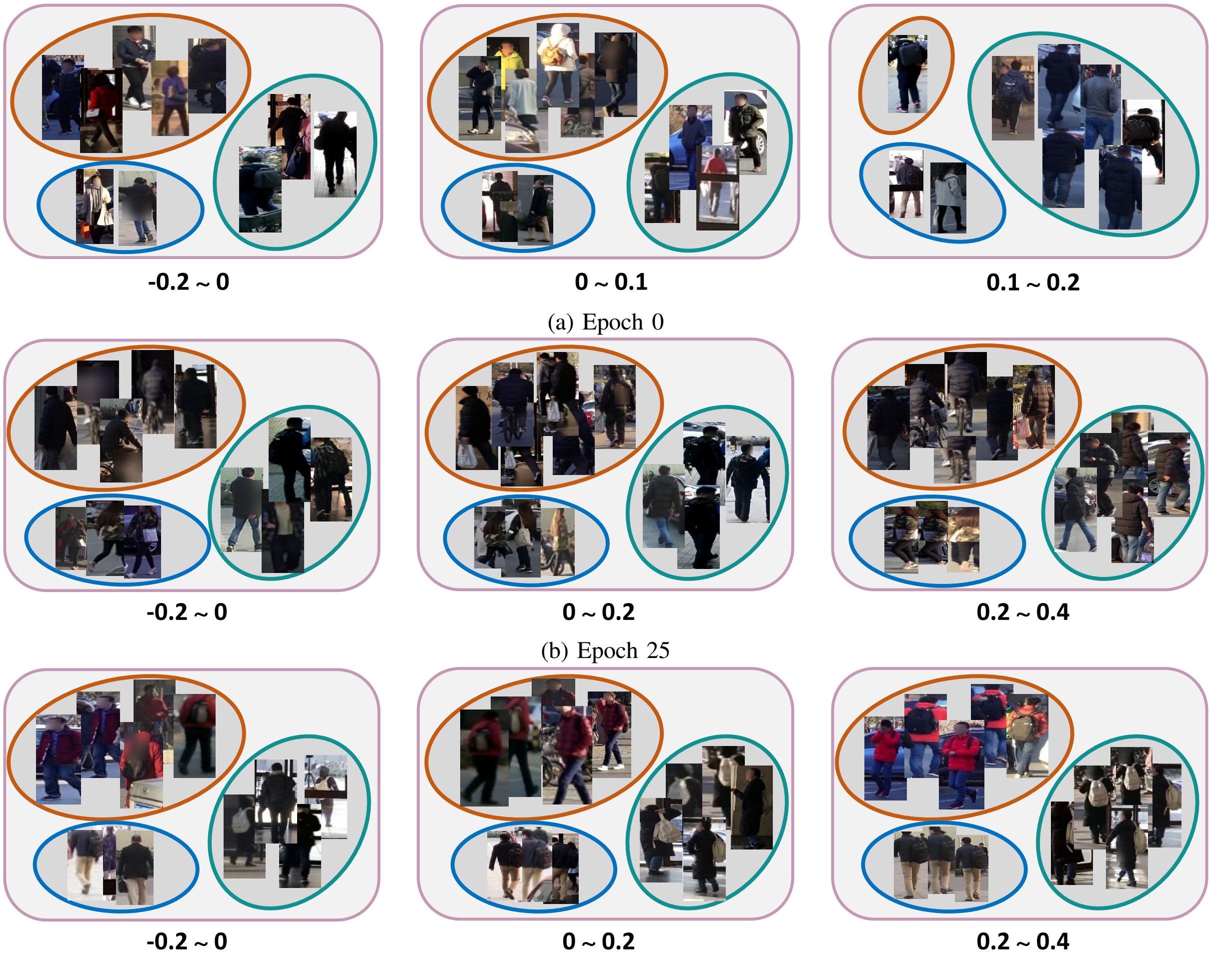
(b) MSMT17

Fig. 6: Visualization of the identity feature distribution via t-SNE [49] on (a) Market-1501 and (b) MSMT17. For each group, features are derived by the baseline model (left) and the model trained with the proposed confidence-guided centroids (CGC) and pseudo labels (CGL) (right), respectively. Model performances (mAP) are also denoted. Different identities are denoted by different colors. **Best viewed in color**.

Silhouette Scores of Samples. To better demonstrate the relationship between the sample-wise clustering confidence and silhouette scores, more clustering results of our method are provided in Fig. 7.

As can be seen, samples are coarsely clustered based on basic visual features at the beginning. As more identity-related information is learned, yet belonging to different identities, samples with similar appearances and poses are gradually grouped together. Additionally, top-ranking images have higher silhouette scores at epoch 25. Finally, at a later stage (epoch 50), better identity information is learned, and images presenting the same identity are grouped together while images belonging to different identities are scattered into different clusters.

Identity Retrieval. We present retrieval results on benchmark datasets in Fig. 8 to demonstrate the improvement of our method over the baseline. Compared to the baseline, our method achieves a better retrieval performance, presenting by more correctly matched images ranking at the top. Additionally, our method is more robust to noise caused by low image quality and cluttered backgrounds. Taking the 4th row in Fig. 8(a) as an example, the baseline fails to retrieve the correct match with cluttered backgrounds (8th image retrieved by



(a) Epoch 0

(b) Epoch 25

(c) Epoch 50

Fig. 7: Visualization of samples (MSMT17) with silhouette scores at (a) epoch 0, (b) epoch 25, and (c) epoch 50. Samples within different clusters are indicated by different colors. **Best viewed in color**.

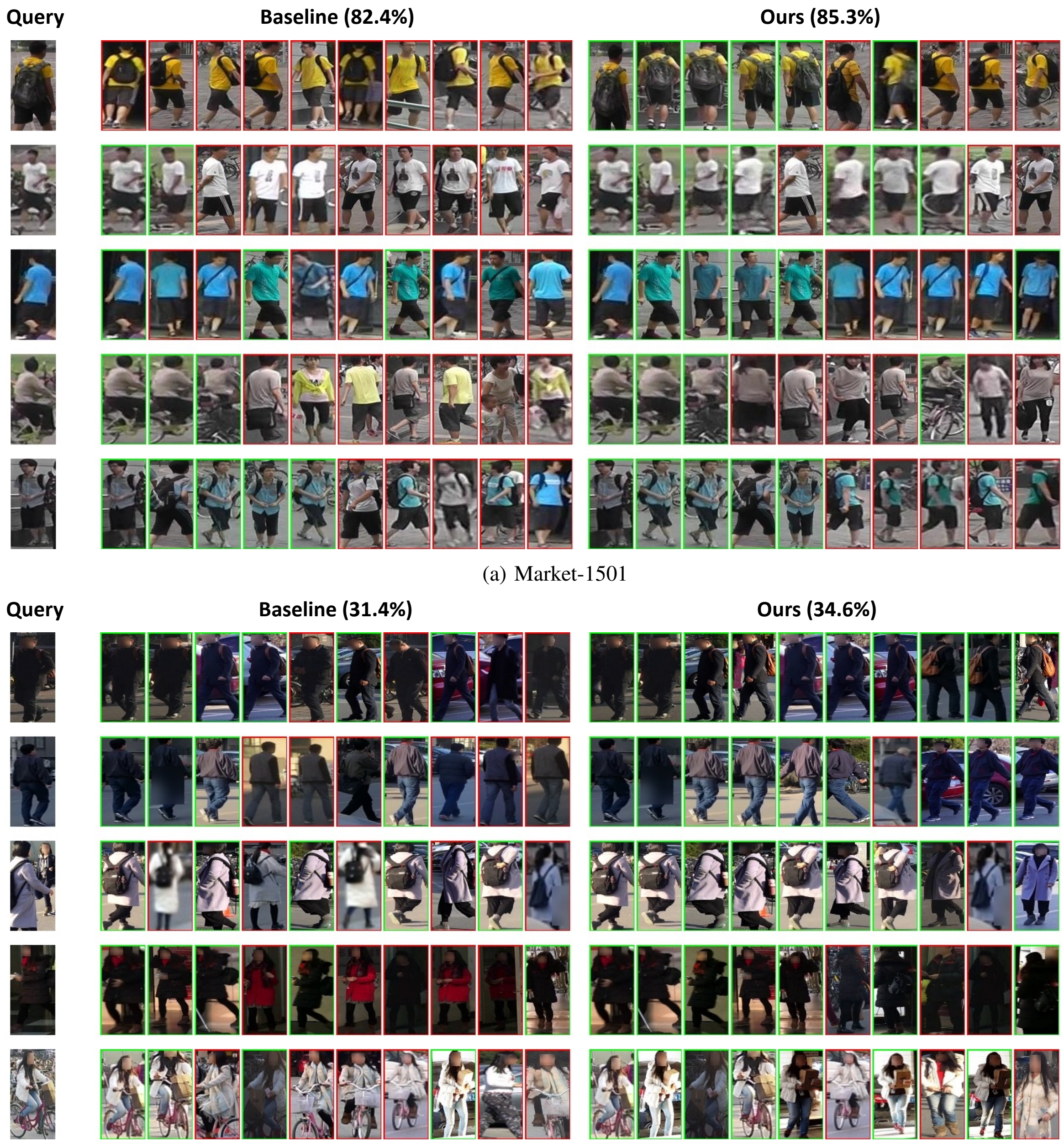
Ours). Meanwhile, due to the low quality of the query image, the features extracted by the baseline is inaccurate, which leads to visually dissimilar images (4th∼10th images) ranking higher than the similar ones (4th∼10th retrieved by Ours). This is because the proposed confidence-guided schemes encourage features to be learned from: 1) better centroids where above noisy samples are excluded due to their low confidence score, and 2) multiple potential correct identities. In this way, the extracted features are not only identity-related but noise-robust, thereby achieving better retrieval results.

G. More Discussions

Identity Consistency Score. The current learning scheme forces samples to approach their assigned cluster centroids, where their identity information is embedded. However, the existence of noisy labels will lead samples to "wrong" centroids. It is especially problematic for low-confidence samples, *i.e.,* boundary samples because they can be closer to other centroids than the assigned ones.

To investigate the problem, we conduct an experiment on MSMT17 [14] to analyze how much the identity information of boundary samples can be presented in the assigned centroids, *i.e.,* the identity consistency in-between. Specifically, we select clusters whose size is over 100 at each epoch. For each cluster, samples whose silhouette scores rank at the bottom 5% are empirically marked as boundary samples. Formally, let $\mathcal{C} = \{(x_i, g_i)\}_{i=1}^{N_c}$ denote a cluster with N_c samples, where g_i refers to the ground-truth identity label provided by the dataset. An identity set $\mathcal{G} = \{g_k\}_{k=1}^{M}$ is then constructed by overall M identities occurring in the cluster. Taking a cluster with 3 samples as an example, *i.e.,* $\mathcal{C} = \{(x_0, 0), (x_1, 0), (x_2, 1)\}$, the corresponding identity set is constructed by overall $M = 2$ identity labels: $\mathcal{G} = \{0, 1\}$. Following the formation of vanilla all-sample based cluster centroids (Eq. (1)), the identity information embedded in the centroid can be obtained by linearly integrating all identities within the cluster via weights $\mathcal{Q} = \{q_k\}_{k=1}^{M}$, where q_k is obtained by $q_k = \frac{1}{N_c}\sum_{g_i \in \mathcal{C}} \mathbb{1}\{g_i = g_k\}$. $\mathbb{1}\{g_i = g_k\}$ equals to 1 when $g_i = g_k$, otherwise 0. Then, the identity consistency score (ICS) between boundary samples and the cluster centroid of $\mathcal{C}$ can be calculated as, $ICS = \frac{1}{N_c}\sum_{g_i \in \mathcal{C}} q_k \cdot \mathbb{1}\{g_i = g_k\}$.

Similar to the vanilla scheme, ICS of our confidence-guided centroids (CGC) scheme can be computed by simply replacing



(a) Market-1501

(b) MSMT17

Fig. 8: Retrieval results of the baseline model (Baseline) and the model trained with the proposed schemes (Ours) on (a) Market-1501 and (b) MSMT17. The performance of models (mAP) is also reported. For each group, the query image is shown at the leftmost, followed by the top-10 images of its ranking list given by different models. The green rectangles indicate correct retrieval results, while red ones denote false retrieval results. **Best viewed in color**.

$\mathcal{C}$ with the confidence-guided subset $\mathcal{C}_q$ during the computation of the weight q_k. Since low-confidence samples are filtered out in the formation of confidence-guided centroids, the identity set $\mathcal{G}$ only includes identities of samples with high confidence scores. We compare the average ICS throughout the training with vanilla all-sample based cluster centroids and the proposed confidence-guided ones, and obtain the curves in Fig. 9.

For the vanilla scheme, only 5.83$\%$ boundary samples carry the same identity information with their assigned centroid at the beginning. Although the ratio gradually climbs to 17.19%, a large proportion of boundary samples (over 80$\%$) still are pushed to centroids where their identity information is rarely presented. Unfortunately, the problem has been aggravated by confidence-guided centroids, where the ratio achieves 14.17$\%$ at most. The low identity consistency scores point out the seriousness of the problem and validate the necessity of our confidence-guided pseudo labels.

Clustering Quality. We also intend to analyze the improvements brought by the proposed strategies in terms of clustering quality. Following ISE [13], four evaluation metrics are employed, which are fowlkes_mallows_score,

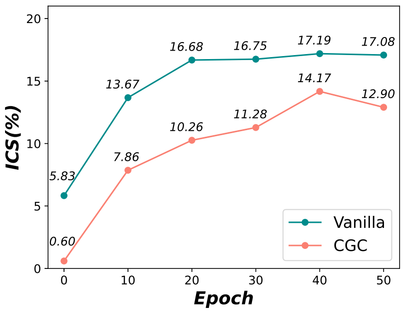


Fig. 9: Identity consistent score (ICS) of boundary samples at different epochs. Vanilla and CGC refer to the previous all-sample based cluster centroids and the proposed confidence-guided centroids, respectively.

adjusted_rand_score, adjusted_mutual_info_score and v_measure_score, respectively. All the above metrics represent the consistency between clustering results and ground-truth labels (*higher is better*). In the paper, we investigate how the four metrics vary throughout the whole training process on two benchmark datasets, *i.e.,*Market-1501 [40] and MSMT17 [14]. The results are illustrated in Fig. 10. To demonstrate the advantages of our method, we also plot the clustering quality curve of "Baseline" models and a state-of-the-art method, ISE [13], in Fig. 10. Note that the ISE curve is plotted by the estimated values from their publication.

As can be seen, for all training schemes, the clustering quality gradually improves during training due to the involvement of more effective features as well as more discriminative networks. Additionally, the model trained with the proposed schemes outperforms both baseline models and ISE on all metrics. The improvements validate the effectiveness of the proposed schemes in clustering quality boosting.

More Backbones. We evaluate the proposed method on two more backbones: Resnet101 [43] and ConvNeXt-Tiny [50]. Performances are reported in Table III. The performance of the original backbone (ResNet50) is also reported for reference.

As can be seen, when a more powerful backbone (ResNet101) is adopted, the performance of our method is improved on two benchmark datasets. In terms of ConvNeXt, we choose ConvNeXt-Tiny (28.6M) since it has a comparable number of parameters to ResNet50 (25.6M). However, the performance on both datasets drops significantly, which is possibly due to the insufficient representability of backbone features. Specifically, the dimensions of features output by ConvNeXT-Tiny and ResNet50 are 768 and 2048, respectively. The representability is limited by the feature dimension, which fails to capture some important cues for person re-identification. Therefore, our method with ConvNeXT-Tiny backbone achieves inferior performances on both datasets.

Backbone	Market1501		MSMT17	
	mAP	top-1	mAP	top-1
ResNet101	87.1	95.2	36.2	68.6
ConvNeXt-Tiny	48.7	71.2	16.3	40.1
ResNet50	85.6	94.3	35.7	66.1

TABLE III: Performance of our method on Resnet101 and ConvNeXt-Tiny backbone network.

V. Conclusion

This paper focused on the pseudo label refinement for clustering-based unsupervised person ReID, which aims to alleviate the pseudo label noise brought by imperfect clustering results. Instead of relying on auxiliary information such as camera IDs, body parts, or generated samples, we refined pseudo labels with internal characteristics, *i.e.,* the sample-wise clustering confidence. Specifically, we proposed to use confidence-guided centroids (CGC) to provide reliable cluster-wise prototypes for feature learning, where low-confidence instances are filtered out during the formation of centroids. Additionally, targeting the problem that a large proportion of samples are pushed to "wrong" centroids, we propose to use confidence-guided pseudo labels (CGL). Such labeling enables samples to approach not only the assigned centroid but other clusters where their identities are potentially embedded. With the aid of CGC and CGL, our method yields comparable performances with, or even outperforms, state-of-the-art pseudo label refinement works that largely leverage auxiliary information.

Limitations and Future Works. Although we conducted multiple scheduling strategies in the paper, the parameters are selected empirically. We attempt to explore adaptive thresholds in the future. Additionally, since the proposed method does not leverage identity labels, it is applicable to diverse re-id related tasks, such as vehicle re-id [51], [52] and text-to-image reid [53]. In the future, we attempt to explore its potential in the multi-modality context with the help of vision-language models such as CLIP [54].

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2021.

[2] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1367–1376.

[3] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2285–2294.

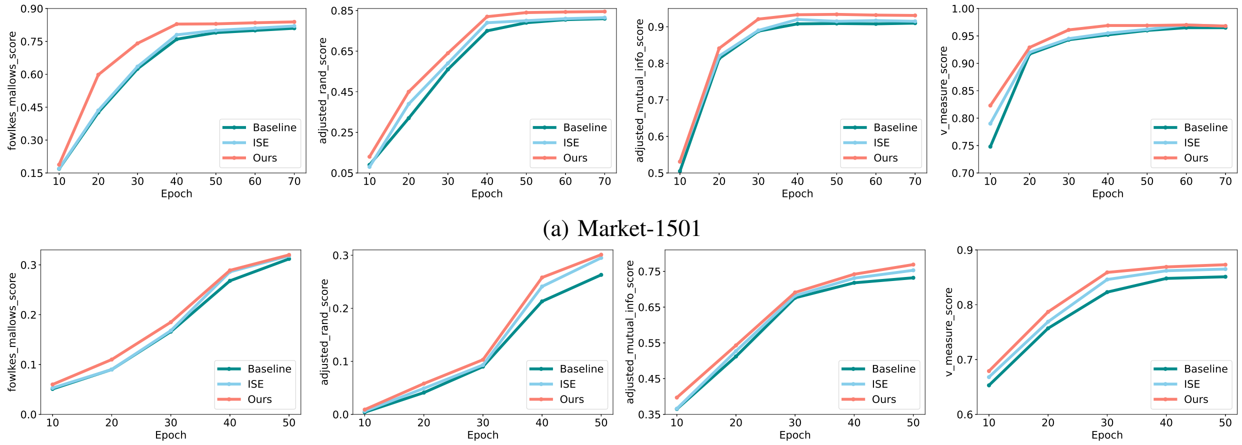
[4] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Proc. NeurIPS*, vol. 33, 2020, pp. 11 309–11 321.

[5] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan, "Idm: An intermediate domain module for domain adaptive person re-id," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 11 864–11 874.

[6] T. He, L. Shen, Y. Guo, G. Ding, and Z. Guo, "SECRET: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 1, 2022, pp. 879–887.

[7] J. Gu, W. Chen, H. Luo, F. Wang, H. Li, W. Jiang, and W. Mao, "Multi-view evolutionary training for unsupervised domain adaptive person re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 344–356, 2022.

[8] H. Wang, M. Yang, J. Liu, and W.-S. Zheng, "Pseudo-label noise prevention, suppression and softening for unsupervised person re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 3222–3237, 2023.



(a) Market-1501

(b) MSMT17

Fig. 10: Clustering quality during the training of the baseline model (Baseline), ISE [13], and the proposed method (Ours) on (a) Market-1501 and (b) MSMT17. **Best viewed in color**.

[9] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 1142–1160.

[10] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14 960–14 969.

[11] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7308–7318.

[12] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 4, 2021, pp. 2764–2772.

[13] X. Zhang, D. Li, Z. Wang, J. Wang, E. Ding, J. Q. Shi, Z. Zhang, and J. Wang, "Implicit sample extension for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7369–7378.

[14] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 79–88.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD*, vol. 96, no. 34, 1996, pp. 226–231.

[16] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3436–3445.

[17] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[18] G. Bertocco, F. Andaló, and A. Rocha, "Unsupervised and self-adaptative techniques for cross-domain person re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4419–4434, 2021.

[19] X. Lin, P. Ren, C.-H. Yeh, L. Yao, A. Song, and X. Chang, "Unsupervised person re-identification: A systematic survey of challenges and solutions," *arXiv preprint arXiv:2109.06057*, 2021.

[20] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3390–3399.

[21] C. Chen, J. Han, and K. Debattista, "Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[22] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10 981–10 990.

[23] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, no. 01, 2019, pp. 8738–8745.

[24] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13 657–13 665.

[25] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 483–499.

[26] Y. Zheng, S. Tang, G. Teng, Y. Ge, K. Liu, J. Qin, D. Qi, and D. Chen, "Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8371–8381.

[27] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11 926–11 935.

[28] M. Wang, J. Li, B. Lai, X. Gong, and X.-S. Hua, "Offline-online associated camera-aware proxies for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 6548–6561, 2022.

[29] L. Wu, D. Liu, W. Zhang, D. Chen, Z. Ge, F. Boussaid, M. Bennamoun, and J. Shen, "Pseudo-pair based self-similarity learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 4803–4816, 2022.

[30] G. Bertocco, A. Theophilo, F. Andaló, and A. Rocha, "Leveraging ensembles and self-supervised learning for fully-unsupervised person re-identification and text authorship attribution," *IEEE Trans. Inf. Forensics Secur.*, 2023.

[31] X. Han, X. Yu, G. Li, J. Zhao, G. Pan, Q. Ye, J. Jiao, and Z. Han, "Rethinking sampling strategies for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 29–42, 2022.

[32] D. Cheng, J. Zhou, N. Wang, and X. Gao, "Hybrid dynamic contrast and probability distillation for unsupervised person re-id," *IEEE Trans. Image Process.*, vol. 31, pp. 3334–3346, 2022.

[33] J. Yin, X. Zhang, Z. Ma, J. Guo, and Y. Liu, "A real-time memory updating strategy for unsupervised person re-identification," *IEEE Trans. Image Process.*, 2023.

[34] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2004–2013.

[35] ——, "Learning invariance from generated variance for unsupervised person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[36] H. Zhang, G. Zhang, Y. Chen, and Y. Zheng, "Global relation-aware contrast learning for unsupervised person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8599–8610, 2022.

[37] J. Peng, G. Jiang, and H. Wang, "Adaptive memorization with group labels for unsupervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[38] M. Li, C.-G. Li, and J. Guo, "Cluster-guided asymmetric contrastive learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 3606–3617, 2022.

[39] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3733–3742.

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

[41] S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2530–2539.

[42] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1318–1327.

[43] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[45] F. Radenović, G. Tolias, and O. Chum, “Fine-tuning cnn image retrieval with no human annotation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2018.

[46] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[47] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Int. Conf. Learn. Repre. (ICLR)*, 2015.

[49] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[50] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 976–11 986.

[51] X. Zhu, Z. Luo, P. Fu, and X. Ji, “Voc-reid: Vehicle re-identification based on vehicle-orientation-camera,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 602–603.

[52] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, “Parsing-based view-aware embedding network for vehicle re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7103–7112.

[53] S. Li, L. Sun, and Q. Li, “Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 1, 2023, pp. 1405–1413.

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

Yunqi Miao received the Ph.D. degree from the Warwick Manufacturing Group (WMG), University of Warwick in 2023. Her research interests include person re-identification, face recognition, and blind face restoration.

Jiankang Deng received the Ph.D. degree from Imperial College London (ICL) in 2020. He is one of the main contributors to the widely used open-source platform Insightface. He has more than 8K citations to his research work. His research interests include deep learning-based face analysis, including detection, alignment, reconstruction, recognition, and generation.

Guiguang Ding (Senior Member, IEEE) received the B.E. and Ph.D. degrees from Xidian University, Xi’an, China, in 1999 and 2004, respectively. He is currently a Professor with the School of Software, Tsinghua University, Beijing, China. His research interests include the areas of multimedia information retrieval, computer vision, and machine learning.

Jungong Han (Senior Member, IEEE) is the Chair Professor in Computer Vision at the Department of Computer Science, University of Sheffield, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning. He is a Fellow of the International Association of Pattern Recognition.