Dual Knowledge Distillation on Multiview Pseudo Labels for Unsupervised Person Re-Identification

Wenjie Zhu, *Member, IEEE,* Bo Peng, and Wei Qi Yan, *Senior Member, IEEE*

***Abstract*—Unsupervised person re-identification (Re-ID) has made significant progress by leveraging valuable pseudo labels from completely unlabeled data. However, the predominant use of pseudo labels heavily relies on clustering results, which may lead to the accumulation of supervision deviation due to inevitable noise. In this paper, we propose a novel framework, namely Dual Knowledge Distillation on Multiview Pseudo Labels (DKD-MPL), to address this challenge. Specifically, the proposed DKD-MPL framework consists of two modules: Global Knowledge Distillation (GKD) and Self-Knowledge Distillation (SKD). In the GKD module, the pseudo labels obtained from the epoch-wise clustering procedure serve as the logits for the teacher model, while the mini-batch query images' pseudo labels act as the logits for the student model. Within the SKD module, we facilitate self-knowledge distillation by considering the pseudo labels generated by positive anchors and query images as two augmentations of the mini-batch data. As a result, DKD-MPL facilitates the exploitation of both global and local complementary knowledge across different views of pseudo labels, thereby mitigating supervision deviation. To demonstrate the effectiveness of DKD-MPL, we provide a theoretical analysis of the proposed loss and conduct extensive experiments on four popular datasets, e.g., Market-1501, DukeMTMC-reID, MSMT17, and VeRi-776. The results indicate that our method surpasses unsupervised approaches and achieves comparable performance to supervised person Re-ID methods.**

***Index Terms*—Unsupervised person re-identification, knowledge distillation, multiview pseudo labels, self-knowledge distillation.**

I. Introduction

PERSON re-identification (Re-ID) is a computer vision task that focuses on recognizing and tracking individuals across non-overlapping cameras or over time. Its primary objective is to identify a person in one camera view and match them with the same person in another camera view. The domain of person Re-ID has demonstrated wide-ranging applications in practical scenarios such as video surveillance, crowd analysis, and human-computer interaction [1]. Accurately identifying individuals across different camera views poses a significant challenge in person Re-ID, especially when there are noticeable changes in appearance caused by variations in pose, clothing, lighting conditions, occlusions, and background clutter. These factors greatly affect the performance of person Re-ID approaches. Therefore, researchers have increasingly focused on addressing various challenges in person Re-ID recently, including occlusion [2]–[4], cross-modality [5], [6] settings, clothing changes [7], [8], and video-based approaches [9], [10].

According to the learning techniques, research work on person Re-ID can be classified into three main categories: supervised learning, Unsupervised Domain Adaptation (UDA), and Purely Unsupervised Learning (PUL) methods. Supervised learning approaches [11]–[16] aim to learn discriminative features from labeled training data and have shown exceptional performance in Re-ID tasks. Shi et al. [14] propose an attribute mining and reasoning framework which makes better use of appearance attributes with localization ensemble and reasoning. Zhu et al. [15] employ self-attention mechanism to learn subtle feature embeddings for fine-grained person Re-ID. However, the labor-intensive process of labeling restricts the practicality of supervised learning methods in real-world scenarios. To overcome this limitation, UDA methods [17]–[25] have emerged, which tackle the Re-ID problem using a labeled dataset not specifically designed for Re-ID tasks. While leveraging labeled source data can improve Re-ID performance, cross-modal scenarios such as infrared, aerial, or hyperspectral domains pose significant challenges due to inherent differences between the source and target domains.

Given the scarcity or unavailability of labeled training data, PUL approaches [26]–[34] for person Re-ID have gained importance in various scenarios. PUL methods focus on learning discriminative representations and performing person matching solely based on visual information extracted from unlabeled data. Clustering is widely employed in PUL approaches to generate pseudo labels[1], thereby guiding the learning of discriminative representations. In this paper, our main focus is on addressing the person Re-ID problem without relying on labeled data in the training set, which falls within the scope of PUL methods.

Within the realm of unsupervised person Re-ID methods, a promising approach is the utilization of self-training strategies, which alternates between clustering and fine-tuning the network to align with the clustering results. Previous studies have shown the effectiveness of this approach [17], [26], [30]. Despite the remarkable performance, these methods still

This research was partially supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.LY24F030005, the National Key Research and Development Program of China under Grant No.2021YFC3340402, and the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant No.2022YW40.

Wenjie Zhu is with College of Information Engineering, China Jiliang University, Hangzhou 310018, China. He is now a visiting scholar at Auckland University of Technology, Auckland 1010, New Zealand (e-mail: zhu.wenjie@aut.ac.nz).

Bo Peng was with The University of Queensland, St. Lucia 4072, Australia.

Wei Qi Yan is with School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand (e-mail: weiqi.yan@aut.ac.nz).

[1]In unsupervised person Re-ID, pseudo labels are artificial labels assigned to unlabeled person samples through the implementation of clustering technologies.

exhibit a performance gap compared to experiments that utilize ground-truth labels [19], [30]. The use of pseudo labels in the training process does provide a form of supervised knowledge. However, it is essential to acknowledge that supervision deviation, which refers to the inconsistency between input and supervision, can lead to overconfident predictions and misclassification of samples, particularly in the context of deep learning. This issue becomes more prominent while relying on *hard* pseudo-labels, which may deviate from the true labels. More importantly, such deviations can introduce noise into the training process, leading to larger performance gaps and degraded model performance. Research outcome has shown that deep neural networks have the capability to fit any ratio of noisy labels [35]. Therefore, persistently relying on *hard* pseudo-labels may not be the optimal choice due to the inevitable noise-induced deviation in supervision. Instead, a more effective approach would be to incorporate accurate pseudo-labels, where the confidence of the predictions can precisely guide the training process.

To improve the pseudo labels in the self-training strategy and address the challenge of supervision deviation in unsupervised Re-ID methods, we propose a novel framework called Dual Knowledge Distillation on Multiview Pseudo Labels (DKD-MPL). The goal of our DKD-MPL framework is to enhance the quality of pseudo labels in unsupervised person Re-ID by incorporating the principles of knowledge distillation. Specifically, we obtain three distinct views of pseudo labels in the DKD-MPL framework, as illustrated in Fig. 1:

- **Global-level Pseudo Label (GPL).** We utilize a clustering algorithm with the embeddings of the training data to obtain pseudo labels that capture partition information at a global level. These labels are updated continuously in each epoch.
- **Query Pseudo Label (QPL).** To generate pseudo labels for query images in the current mini-batch data, we design a dynamic classifier based on class prototypes. This enables the production of query-level pseudo labels.
- **Positive-anchor Pseudo Label (PaPL).** We obtain the anchor[2] points associated with query images in the current mini-batch data and generate corresponding positive-anchor pseudo labels using the dynamic classifier described earlier. These labels can be considered as distinct augmentations of QPLs.

By incorporating these three types of pseudo labels within the DKD-MPL framework, our aim is to improve the accuracy and quality of the pseudo labels in unsupervised person Re-ID. Additionally, the DKD-MPL framework, as shown in Fig. 2, consists of two modules: Global Knowledge Distillation (GKD) module and Self-Knowledge Distillation (SKD) module. In the GKD module, the network that produces GPLs is referred to as the *teacher* model, as it leverages the relatively reliable pseudo labels at the global level. On the other hand, the network trained in a mini-batch learning manner is considered the *student* model. In the SKD module,

[2]An anchor [36] is a reference point that serves as a basis for learning the similarity or dissimilarity between other samples.

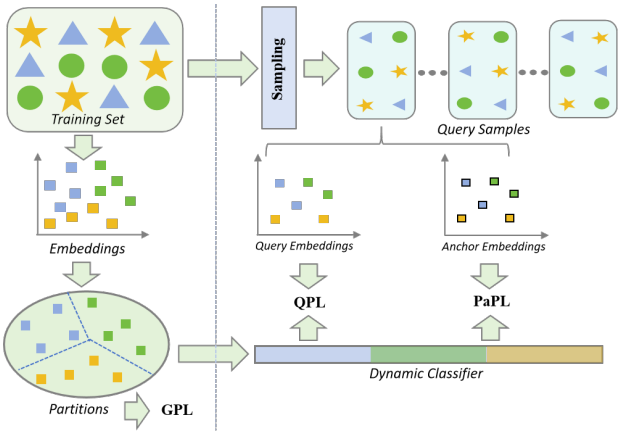


Fig. 1. Three distinct views of pseudo labels proposed in the proposed DKD-MPL framework. On the left branch, the training set is fed through deep neural networks to obtain feature embeddings, which are then clustered into multiple partitions by utilizing the clustering algorithm, achieving the **GPL**s. On the right branch, the P-K sampling method [37], i.e., P classes and K samples per class are randomly selected, is used to generate the query samples and their corresponding positive anchor samples. These samples are then processed by a dynamic classifier constructed by the cluster prototypes to obtain **QPL**s and **PaPL**s.

the positive anchors in the *student* model act as augmented versions of the current query images, enabling self-knowledge distillation and enhancing feature learning at the local level. More details about the architecture and mathematical description of the proposed DKD-MPL framework will be provided in Section III.

In summary, the primary technical contribution of our proposed DKD-MPL framework in multimedia computing research lies in its utilization of knowledge distillation principles to improve the quality of pseudo labels in unsupervised person Re-ID. Our method addresses the challenge of supervision deviation commonly encountered in unsupervised person Re-ID approaches. By implementing dual knowledge distillation on multiview pseudo labels, our framework aims to enhance the accuracy and reliability of the Re-ID system. Importantly, the proposed framework is notably versatile and does not require any prior knowledge of a person's wearings or post-ranking strategy. Consequently, it offers comprehensive technical support for purely unsupervised person Re-ID based on pseudo labels. The main contributions of this work are concluded as follows:

- We introduce a novel insight into unsupervised person Re-ID by applying the knowledge distillation principle on multiview pseudo labels to reduce the supervision deviation caused by initial clustering results during the training process.
- A dual knowledge distillation loss is proposed, which includes a GKD loss using GPLs and QPLs obtained via a dynamic classifier, and an SKD loss utilizing positive anchors and query images from the current mini-batch data.
- We provide a thorough analysis of theoretical effectiveness of our DKD-MPL method by establishing connections with state-of-the-art unsupervised learning methods.
- Experimental results on Market-1501, DukeMTMC-reID,

MSMT17, and VeRi-776 datasets that demonstrate the superior performance of our DKD-MPL method over existing unsupervised person Re-ID approaches and its competitive performance compared to supervised person Re-ID methods.

The remainder of this paper is organized as follows: We present a brief review of related work, including unsupervised person Re-ID and knowledge distillation in Section II. In Section III, we provide an elaboration of the proposed method, outlining its framework and architecture in detail. In Section IV, we describe the experimental setup and present the results and analysis. Finally, in Section V, the conclusion summarizes the contributions, discusses potential avenues for future research, and concludes the paper.

II. Related Work

In this section, we provide a comprehensive review of recent unsupervised person Re-ID approaches, encompassing unsupervised domain adaptation and purely unsupervised methods. Additionally, we introduce the principle of knowledge distillation briefly.

A. Unsupervised Person Re-ID

Unsupervised person Re-ID has become popular due to the difficulty of manual labeling. The objective of unsupervised person Re-ID is to learn discriminative representations for person retrieval without relying on labeled data. To obtain supervision, related methods have focused on leveraging knowledge from labeled data and transforming the unsupervised person Re-ID problem into a domain adaptation task. It is crucial to bridge the distribution gap between labeled source data and unlabeled target data in UDA-based unsupervised person Re-ID. To address this, Fu et al. [38] utilized potential similarities among global body and local parts of unlabeled samples to create multiple clusters from different perspectives. These independent clusters served as pseudo labels to supervise the training process. Dai et al. [23] proposed an alternative approach to directly aligning the source and target domains against each other with their respective intermediate domains. This alignment strategy aims to facilitate a smooth knowledge transfer between the domains. In addition, augmentation-based UDA methods [39]–[41] have been explored to address the domain shift between labeled source data and unlabeled target data in re-identification tasks. The objective of these methods is to transfer the image style from the source data to the target data. They effectively bridge the gap between the different domains, allowing for improved performance when dealing with unlabeled target data.

In addition to the aforementioned UDA approaches, there is another research direction in unsupervised person Re-ID that explores the use of PUL methods, which do not rely on labeled data during training. Recent advancements in unsupervised representation learning have made it possible to learn discriminative representations solely using unlabeled target data. In line with the research work, the training process heavily relies on generated pseudo labels. One approach, called Bottom-Up Clustering (BUC) [26], initially treats each sample as an individual cluster and then performs bottom-up clustering to generate pseudo labels. Hierarchical Clustering with Triplet loss (HCT) [28] adopts a hard-batch triplet loss within a hierarchical clustering framework to reduce the influence of challenging examples and obtain high-quality pseudo labels. Self-Paced Clustering (SPCL) [19], on the other hand, employs DBSCAN algorithm [42] to estimate pseudo labels and utilizes a novel self-paced strategy to mine reliable clusters for refining the features stored in memory. Inter-instance Contrastive Encoding (ICE) [33] leverages inter instance pairwise similarity scores as soft pseudo labels to enhance the consistency between augmented and original views, which is robust to augmentation perturbations. Recently, Cluster Contrast [30] modifies the baseline pipeline by constructing a cluster-level memory, which accelerates the updating process for consistency and achieves state-of-the-art performance in Re-ID tasks.

In recent years, researchers have been increasingly interested in addressing the issue of noisy pseudo labels in unsupervised person Re-ID. Several approaches have been proposed to refine these labels [29], [34], [43], [44]. One notable method is Refining Pseudo Label with Clustering Consensus (RLCC) [29], which builds upon the SPCL framework. RLCC introduces a self-paced strategy as a post-processing step on the cluster results to further refine the pseudo labels. Chen et al. [44] propose a label reliability perception technique for person Re-ID that refines the noisy labels. These methods typically perform label refinement as a pre-processing step or involve determining the reliability of the labels. Unlike those PUL methods that rely on pseudo labels generated from global clustering, Wu et al. [34] take a different approach by constructing patch surrogate classes as initial supervision. They propose a method that assigns pseudo labels to samples using pairwise gradient-guided similarity separation. In contrast to these approaches, our proposed framework focuses on directly improving the pseudo labels during training using knowledge distillation techniques. By incorporating knowledge distillation, we aim to enhance the accuracy and reliability of the pseudo labels. This, in turn, leads to improved representation learning and overall performance in unsupervised person Re-ID tasks.

B. Knowledge Distillation

Knowledge distillation was initially introduced in [45] as a method for model compression. The basic idea behind this approach is to leverage the distilled dark knowledge from a pre-trained complex model to generate soft labels that can then assist in optimizing a simpler model. In the classic form of knowledge distillation, the student model is trained to replicate the output probabilities (logits) of the teacher model by minimizing the cross-entropy loss between the output probabilities of the two models. Conversely, feature-based knowledge distillation involves training the student model to replicate the intermediate feature representations of the teacher model. Recently, several works have attempted to use the student network itself as a teacher model, known as self-knowledge distillation (or self-distillation), achieved through augmentation on network architecture [46] or input

data [47], which avoids substantial computational resources on pre-training and improves the efficiency of network training.

In contrast to supervised person Re-ID approaches, our method eliminates the requirement for manual annotation efforts as it does not necessitate any labeled data for training. This renders it cost-effective and scalable, leveraging the abundance of unlabeled data that is available in real-world scenarios. Furthermore, our approach possesses an advantage over UDA methods by being adaptable to diverse domains and datasets without depending on the availability or quality of labeled data, enabling its application in dynamic and varied settings.

Despite the progress of previous PUL approaches for resolving the pseudo labels issues, we are interested in developing a dual knowledge distillation framework with multiview pseudo labels to tackle the supervision deviation challenge for purely unsupervised person Re-ID. However, defining a reliable teacher model directly in an unsupervised setting poses challenges due to the absence of labeled data. To overcome this, we leverage the pseudo labels obtained from epoch-wise clustering as the logits of a teacher model and employ Mahalanobis distance-based embeddings and prototypes to generate pseudo labels for mini-batch query images, serving as the logits of a student model. In addition, we introduce a self-knowledge distillation module in the student branch that utilizes pseudo labels from positive anchors and query images within the same cluster to enhance representation learning. By updating the teacher model with the entire data and refining the student model through self-knowledge distillation, our method effectively exploits both global and local complementary knowledge from different views of pseudo labels, mitigating supervision deviation.

III. Methodology

In this section, we will provide a detailed explanation of the proposed dual knowledge distillation framework that leverages multiview pseudo labels. In order to provide a comprehensive understanding of the proposed framework, this section will begin by introducing the mathematical symbols used. Following this, we will present the technical details, including an overview of the framework, the modules involved, and the corresponding loss functions. Finally, we will provide training details and theoretical analysis to further support and explain the proposed method.

A. Preliminary

Before delving into the details of our DKD-MPL framework, we specify the notations for concise statements. In this paper, 3-ordered tensors are represented by blackboard bold letters, such as $\mathbb{X}$, while matrices and vectors are denoted by uppercase and lowercase bold letters $\boldsymbol{X}$, $\mathbf{x}$, respectively. Table I provides an overview of all the notations utilized in this paper.

Based on the clarified notations, let $\mathbb{X} = \{\boldsymbol{X}_i\}_{i=1}^{n}$ represent the training data. Here, $\mathbb{X} \in \Re^{h\times w\times n}$ denotes the set of input samples, where $\boldsymbol{X}_i \in \Re^{h\times w}$ represents the data of an individual sample. The variables h, w, and n correspond to the height, width, and total number of image samples, respectively. In the context of person Re-ID, our objective is to learn an effective feature embedding function that assigns higher rankings to similar samples within the resulting embedding space. Since there is a lack of labeled training data in unsupervised person Re-ID, we address this limitation by leveraging clustering techniques and training the model using pseudo labels as a form of supervision.

TABLE I
Notations in this paper

Notation	Description
$\mathbb{X} \in \Re^{h\times w\times n}$	A 3-order Tensor $\mathbb{X}$
$\boldsymbol{X} \in \Re^{h\times w}$	A matrix with h rows and w columns
$\boldsymbol{X}_i \in \Re^{h\times w}$, $\mathbf{x}_i \in \Re^{h}$	The matrix or vector denoted by index i
$\mathbf{x}[i]$	The i-th element of a vector
$\mathcal{F}_\theta(\cdot)$	Mapping function parameterized by θ
$d_j(\mathbf{x})$	The distance between x and the feature distribution of the j-th cluster
$\lvert\mathcal{B}\rvert$	The number of samples in the mini-batch $\mathcal{B}$

B. Overview of Framework

As previously discussed, the motivation behind our framework is to enhance features through the application of knowledge distillation using multiview pseudo labels. Illustrated in Fig. 2, our proposed DKD-MPL framework consists of two main modules: GKD and SKD. Each module utilizes two types of pseudo labels to facilitate the knowledge distillation task. Now, we will delve into an overview of the data streams performed by the GKD and SKD modules.

- The GKD module employs an encoder $\mathcal{F}_\theta$, parameterized by using θ, to obtain deep representations for all samples in the dataset. A clustering algorithm, specifically DBSCAN, is then utilized to generate both the clustering results and GPLs for the training samples. Moreover, mini-batch data, sampled from the training set, is fed into a weight-sharing encoder to extract deep representations for these query samples. A classifier $\mathcal{C}(\cdot|\boldsymbol{W})$, parameterized by the weight matrix $\boldsymbol{W}$ derived from the clustering results, is designed to obtain the QPLs for these query samples.
- In the SKD module, a new data stream is introduced. This stream is obtained by computing the positive anchor points using the intermediate representations of the query samples. Subsequently, the classifier $\mathcal{C}(\cdot|\boldsymbol{W})$ from the GKD module is utilized to generate pseudo labels for these positive anchor points, known as PaPLs.

To enable the dual knowledge distillation process using multiview pseudo labels, we have carefully designed loss functions known as the GKD loss and SKD loss for the GKD and SKD modules, respectively. As mentioned earlier, the GKD loss incorporates GPLs and QPLs to facilitate knowledge distillation, while the SKD loss leverages QPLs and PaPLs for self-knowledge distillation.

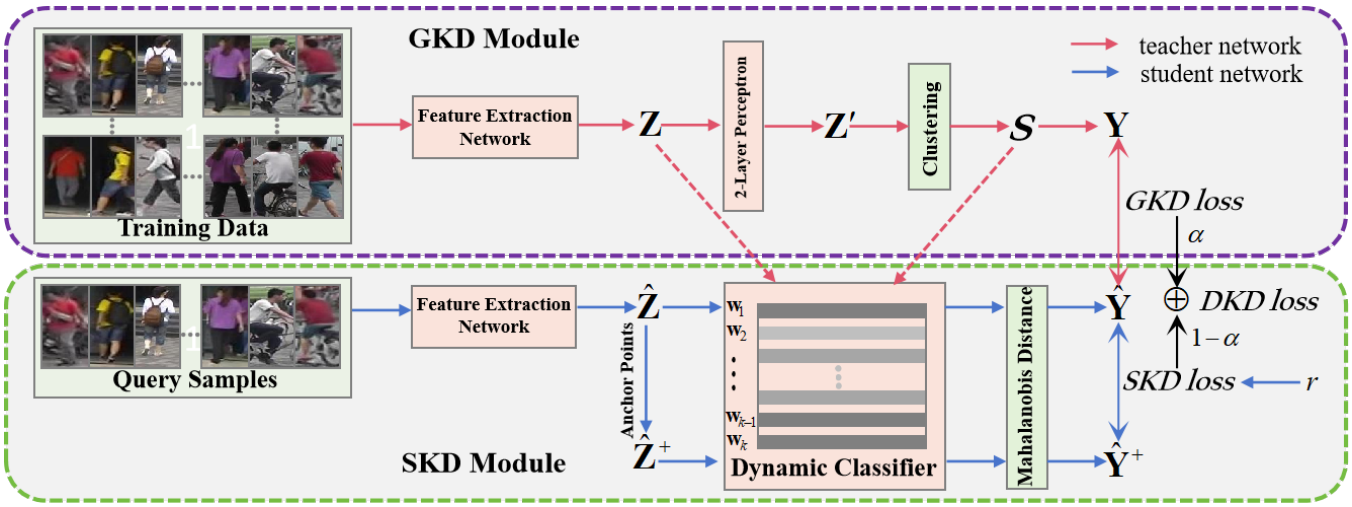


Fig. 2. Overview of the proposed DKD-MPL framework. The framework consists of two modules: the GKD module and the SKD module. In the GKD module, we apply the DBSCAN clustering algorithm [42] on the low-dimensional embeddings $\boldsymbol{Z}^{'}$, obtained from $\boldsymbol{Z}$ via a 2-layer perceptron, to partition the training set into clusters denoted as $\mathcal{S}$ which can be used to represent the the GPL vectors $\mathbf{Y}$. Simultaneously, using the query samples from the current mini-batch and the dynamic classifier based on Mahalanobis distance, we derive the QPL vectors $\hat{\mathbf{Y}}$. In the SKD module, we obtain anchor points by using the query images' features $\hat{\boldsymbol{Z}}$ and consider them as augmentations of the query samples. These anchor points are then used to derive the PaPL vectors $\hat{\mathbf{Y}}^{+}$. Both the PaPL vectors $\hat{\mathbf{Y}}^{+}$ and QPL vectors $\hat{\mathbf{Y}}$ are utilized in calculating the SKD loss, with a score hyperparameter r.

C. GKD Module

In contrast to supervised learning methods that rely on ground-truth labels, the GKD module in our DKD-MPL framework introduces a teacher model that utilizes reliable pseudo labels generated instead. As shown in Fig. 2, following the unsupervised person Re-ID pipeline, we apply a clustering algorithm to the complete feature set of the training data, denoted as $\boldsymbol{Z} \in \Re^{d\times n}$, where d is the feature dimension. This process generates global partitions, represented as $\mathcal{S} := \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k\}$, where k is the number of classes, for the image samples. The pseudo labels obtained from the clustering results serve as initial supervision. obtained from the clustering results serve as initial supervision. Subsequently, we utilize class prototypes obtained from these features to create a classifier, denoted as $\boldsymbol{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_j, \cdots, \mathbf{w}_k] \in \Re^{d\times k}$, where $\mathbf{w}_j \in \Re^d$ represents the prototypical representation of the class indexed by j.

During the training stage, the query image samples $\boldsymbol{X}_i$s are selected from a mini-batch of data using P-K sampling. These query images are then passed through the encoder network, and the resulting encoded representations are fed into the classifier $\mathcal{C}(\cdot)$. The classifier $\mathcal{C}(\cdot)$ produces predicted probabilities in the form of a vector $\hat{\mathbf{y}}_i = \mathcal{C}\left(\mathcal{F}_\theta\left(\boldsymbol{X}_i\right)|\boldsymbol{W}\right) \in [0,1]^k$ for the given image sample $\boldsymbol{X}_i$, obtained by applying softmax over distances to the distribution of each cluster j:

$$\hat{\mathbf{y}}_i[j] = \frac{\exp(-d_j(\mathbf{z}_i))}{\sum_{j=1}^{k}\exp(-d_j(\mathbf{z}_i))} \triangleq \frac{\exp(-d_M(\mathbf{z}_i,\mathbf{w}_j))}{\sum_{j=1}^{k}\exp(-d_M(\mathbf{z}_i,\mathbf{w}_j))}. \tag{1}$$

In Eqn. (1), the encoded feature of $\boldsymbol{X}_i$ is denoted as $\mathbf{z}_i = \mathcal{F}_\theta(\boldsymbol{X}_i) \in \Re^d$, where $d_j(\mathbf{z}_i)$ represents the distance between the embedding $\mathbf{z}_i$ and the feature distribution of the j-th cluster. This distance is based on the assumption that there exists an embedding in which points are distributed around a prototypical representation for each cluster. To measure this distance, we use Mahalanobis distance d_M between $\mathbf{z}_i$ and the corresponding prototype $\mathbf{w}_j$, with a covariance matrix $\boldsymbol{S}_j$. The Mahalanobis distance is:

$$d_M(\mathbf{z}_i,\mathbf{w}_j) = (\mathbf{z}_i-\mathbf{w}_j)^T\boldsymbol{S}_j^{-1}(\mathbf{z}_i-\mathbf{w}_j). \tag{2}$$

Assuming that the distribution around each weight vector is isotropic in the embedding space, such that $\boldsymbol{S}_j = s_j^2\mathbf{I}$ for $j = 1, 2, \cdots, k$, where $\mathbf{I}$ denotes the identity matrix. By substituting Eqn. (2) into Eqn. (1), we can obtain the predicted probability w.r.t. the given image sample $\boldsymbol{X}_i$:

$$\hat{\mathbf{y}}_i[j] = \frac{\exp\left(-||\mathbf{z}_i-\mathbf{w}_j||^2/s_j^2\right)}{\sum_{j=1}^{k}\exp\left(-||\mathbf{z}_i-\mathbf{w}_j||^2/s_j^2\right)}, \tag{3}$$

where $s_j > 0$ reflects the concentration level of the distribution around a weight vector in the corresponding cluster and is dynamically estimated during training. By employing Eqn. (3), we can predict the probability w.r.t. the given image sample $\boldsymbol{X}_i$ applying softmax over Mahalanobis distances to the distribution of each cluster j.

Given the GPL vectors $\mathbf{y}_i$s from a mini-batch data in a P-K sampling way, the output QPL vectors $\hat{\mathbf{y}}_i$s produced via Eqn. (1) are encouraged to be close to GPL vectors $\mathbf{y}_i$s achieved from the teacher model. To achieve this, the GKD loss can be defined using the cross-entropy function as:

$$\mathcal{L}_{\text{GKD}} = \sum_{i=1}^{|\mathcal{B}|} H(\mathbf{y}_i,\hat{\mathbf{y}}_i) = -\sum_{i=1}^{|\mathcal{B}|}\mathbf{y}_i\cdot\log\hat{\mathbf{y}}_i, \tag{4}$$

where $H(a,b)$ denotes the cross-entropy between a and b, and $|\mathcal{B}|$ denotes the number of samples in the mini-batch $\mathcal{B}$. By minimizing the cross-entropy loss with respect to the GPL vector $\mathbf{y}_i$ and QPL vector $\hat{\mathbf{y}}_i$ in Eqn. (4), we aim to utilize

the latent knowledge distilled from the teacher model in the query images of the current mini-batch.

D. SKD Module

Clustering is an effective approach for generating labels for unlabeled data, which can provide supervision for network training in an unsupervised setting. However, clustering techniques may group visually similar samples with different identities, leading to incorrect or misleading supervision during training. This challenge is compounded with the memorization effect observed in deep neural networks, which may struggle to discern deviations from accurate supervision. To alleviate this issue, we propose the SKD module, which leverages self-knowledge distillation to refine feature learning and mitigate the negative impact of clustering-based supervision during training.

During training, the network optimization is guided by two sources of supervision signals. The first source consists of pseudo labels defined in Eqn. (1). The second source is obtained from the anchor points $\boldsymbol{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_i, \cdots, \mathbf{q}_P\}$, which are used to evaluate the consistency of the supervision signal with the input. To this end, we directly use the model prediction of anchor points $\mathcal{C}(\boldsymbol{Q}|\boldsymbol{W})$ as soft labels, without relying on additional pre-trained models. To optimize the learning process within the SKD module, we leverage KL-divergence to define the SKD loss as:

$$\mathcal{L}_{\text{SKD}} = \sum_{i=1}^{|\mathcal{B}|} \mathbb{KL}\left(\hat{\mathbf{y}}_i^+ \| \hat{\mathbf{y}}_i\right), \tag{5}$$

where $\mathbb{KL}(\cdot||\cdot)$ denotes the KL-divergence function, $\hat{\mathbf{y}}_i^+ = \mathcal{C}(\mathbf{q}_i^+|\boldsymbol{W})$ which can be obtained from Eqn. (1), and $\mathbf{q}_i^+$ represents the corresponding positive anchor point for the sample $\boldsymbol{X}_i$. In Eqn. (5), the SKD loss measures the discrepancy of soft labels between the current query images and the corresponding positive anchors. Given a sample $\boldsymbol{X}_i$, we consider the classes with the highest r scores in $\hat{\mathbf{y}}_i^+$, constituting a set denoted by $\mathcal{P}_r(\hat{\mathbf{y}}_i^+)$. Here, r is a introduced hyper-parameter and will be empirically discussed in the experiment section. Based on the above, the KL-divergence between $\hat{\mathbf{y}}_i^+$ and $\hat{\mathbf{y}}_i$ defined in Eqn. (5) would be altered as follows:

$$\mathbb{KL}\left(\hat{\mathbf{y}}_i^+ || \hat{\mathbf{y}}_i\right) = \sum_{j \in \mathcal{P}_r(\hat{\mathbf{y}}_i^+)} \hat{\mathbf{y}}_i^+[j] \log \frac{\hat{\mathbf{y}}_i^+[j]}{\hat{\mathbf{y}}_i[j]}, \tag{6}$$

where $\hat{\mathbf{y}}_i^+[j]$ and $\hat{\mathbf{y}}_i[j]$ denote the j-th entries of $\hat{\mathbf{y}}_i^+$ and $\hat{\mathbf{y}}_i$, respectively. In the real application, $\hat{\mathbf{y}}_i[j] \neq 0$ is always the case, where $j \in \mathcal{P}_r\left(\hat{\mathbf{y}}_i^+\right)$ represents the subset of indices corresponding to the highest r scores in $\hat{\mathbf{y}}_i^+$.

E. Dual Knowledge Distillation Loss

In the previous section, we introduced the dual knowledge distillation framework, which consists of two modules: GKD and SKD. The corresponding loss functions are referred to as GKD loss and SKD loss, respectively. These losses are designed using different types of labels, such as GPLs, QPLs, and PaPLs. By combining the GKD and SKD losses, the overall loss function of our dual knowledge distillation framework using multiview pseudo labels is defined as follows:

$$\mathcal{L}_{\text{DKD}} = \alpha \mathcal{L}_{\text{GKD}} + (1-\alpha)\mathcal{L}_{\text{SKD}}, \tag{7}$$

where $0 < \alpha < 1$ is a hyperparameter used to balance between the GKD and SKD losses. The advantages of these two losses can be summarized as follows:

- The GKD loss incorporates two forms of labels for the same inputs, obtained from weight-sharing encoders but with different approaches. GPLs are generated using a global clustering algorithm that clusters the embeddings based on their similarity and assigns pseudo-labels to each cluster. QPLs, on the other hand, are generated through query sample classification in a mini-batch, where the embeddings are used as inputs to the student model. The GKD loss encourages the student model to learn both the overall clustering structure captured by the global clustering algorithm and the finer-grained classification information from query sample classification.
- The objective of SKD is to update the embeddings of query image samples within the current mini-batch data by leveraging knowledge from themselves, i.e., positive anchor points. To ensure consistency between the output and deep features, an SKD loss is designed, incorporating QPLs and PaPLs. The classifier generated in the GKD module acts as a weight-sharing mapping function for self-knowledge distillation.

F. Training Strategy

Initially, the training procedure of the proposed DKD-MPL framework for unsupervised person Re-ID is outlined in Algorithm 1. Given the unlabeled training set and pretrained model, our approach comprises two primary processes: *Clustering and Initialization* (Steps 3-5), and *Weights Updating* (Steps 7-10). The *Clustering and Initialization* process is executed in each epoch, where low-dimensional embeddings of the features are clustered using the DBSCAN algorithm [42] into k groups, generating pseudo labels. In the process of *Weights Updating*, the current mini-batch is utilized in each iteration to update the weights. To ensure a comprehensive understanding of the proposed method's flow, we will provide detailed explanations for each of these processes.

1) Clustering and Initialization: In each epoch, the pseudo labels are generated by clustering low-dimensional embeddings of the features $\{\widetilde{\mathbf{z}}_i | \widetilde{\mathbf{z}}_i = \mathcal{F}_{\widetilde{\theta}}(\boldsymbol{X}_i)\}_{i=1}^n$. To achieve optimal clustering performance, low-dimensional embeddings are obtained through a two-layer perception instead of using $\widetilde{\mathbf{z}}_i$s directly. The two-layer perceptron is particularly effective in generating high-quality embeddings. $\mathcal{F}_{\widetilde{\theta}}(\cdot)$ is given by the encoder with the parameters $\widetilde{\theta}$ learned in the last epoch, resulting in k clusters denoted by $\mathcal{S} := \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k\}$. The classifier is then initialized by setting each weight vector as the mean of the cluster it represents:

$$\mathbf{w}_i = \texttt{Norm}\left(\frac{1}{|\mathcal{S}_i|} \sum_{\boldsymbol{X} \in \mathcal{S}_i} \mathcal{F}_{\widetilde{\theta}}(\boldsymbol{X})\right), \tag{8}$$

where $\texttt{Norm}(\cdot)$ indicates the ℓ_2-normalization function.

The parameter s reflects the concentration degree of an embedding distribution in terms of the average deviation from the mean in a distribution. Smaller values of s indicate higher levels of concentration. Ideally, we expect s to be small if the corresponding weight vector's average distance of samples in the same cluster to the mean is small and the cluster size is large. Therefore, we define s for each cluster as follows:

$$s_i = \sqrt{\frac{\sum_{\boldsymbol{X}\in\mathcal{S}_i} ||\mathcal{F}_{\tilde{\theta}}(\boldsymbol{X}) - \mathbf{w}_i||}{|\mathcal{S}_i| \log\left(|\mathcal{S}_i| + \varepsilon\right)}}, \tag{9}$$

where ε is introduced to prevent s from being overly large in the case of small cluster size. By integrating Eqn. (9) with Eqn. (3), a mechanism is introduced where samples in a loose cluster (with larger s_i) are encouraged to move closer to the cluster centroid, while those in a tight cluster (with smaller s_i) are encouraged to maintain their proximity to the centroid. This approach promotes the creation of more balanced clusters with similar concentration levels. It is worth noting that previous methods typically do not explicitly consider cluster diversity, except for BUC [26], which employs a diversity regularization technique to prevent clusters from becoming redundant in a heuristic manner.

2) Weights Updating: After forwarding the input data into the network, we compute the overall loss and propagate the gradient backwards to update the encoder parameter θ using Eqn. (10), while the classifier weights $\boldsymbol{W}$ are updated using a momentum averaging formulation defined in Eqn. (11).

$$\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}, \tag{10}$$

$$\mathbf{w}_i \leftarrow \texttt{Norm}\Big(\xi \mathbf{w}_i + (1-\xi)\frac{1}{|\mathcal{B}_i|}\sum_{\boldsymbol{X}\in\mathcal{B}_i} \mathcal{F}_\theta(\boldsymbol{X})\Big), \tag{11}$$

where $\mathcal{B}_i = \{\boldsymbol{X} | \boldsymbol{X} \in \mathcal{B} \text{ and } \boldsymbol{X} \in \mathcal{S}_i\}$, and ξ is a momentum coefficient.

G. Discussion

In this subsection, we will discuss the proposed DKD-MPL framework from a theoretical perspective and its potential to alleviate supervision deviation caused by pseudo labels, focusing specifically on the GKD and SKD loss functions.

1) GKD Loss: The GKD loss, as defined in Eqn. (4), is related to previous approaches that have explored distance-based classification in the context of few-shot classification [48]. These methods commonly employ a point-to-point metric, such as Euclidean distance or cosine distance, to establish similarity between samples. In contrast, the proposed GKD module introduces a loss function that leverages the Mahalanobis distance, a point-to-set metric. This approach enables the learning of a robust embedding space by estimating the data distribution of each cluster, effectively combining distance-based classification.

Algorithm 1 Training procedure of DKD-MPL framework for unsupervised person Re-ID.

Input: the unlabeled training set $\mathbb{X}$, momentum coefficient ξ for Eqn. (11), tradeoff parameter α, score parameter r for Eqn. (6), learning rate η for Eqn. (10), number of epochs E, number of iterations T, and an encoder $\mathcal{F}_\theta$ with parameters θ pretrained on Image-Net.
for epoch = 1 to E **do**
  Initialize another encoder $\mathcal{F}_{\hat{\theta}}$ as the encoder $\mathcal{F}_\theta$ via $\hat{\theta} = \theta$.
  Perform clustering using DBSCAN algorithm on the low-dimensional embeddings of training features given by $\hat{\theta}$, resulting in k clusters.
  Initialize the classifier via Eqn. (8) and estimate the concentration level of each cluster via Eqn. (9).
  for iteration = 1 to T **do**
    Sample $P \times K$ instances from $\mathbb{X}$ to generate a mini-batch $\mathcal{B}$.
    Find class-wise anchor points by averaging embeddings with the same labels in the current mini-batch $\mathcal{B}$.
    Compute the overall loss function via Eqn. (7).
    Update the encoder parameters θ via Eqn. (10) and the weight vectors of the classifier via Eqn. (11).
  end for
end for
Output: A desirable embedding function $\mathcal{F}_\theta$.

To explain the GKD loss further, we focus on $\hat{\mathbf{y}}_i$. It can be checked that both $\mathbf{z}$ and $\mathbf{w}$ are ℓ_2-normalized, i.e., $||\mathbf{z}-\mathbf{w}||^2 = 2 - 2\mathbf{z}\cdot\mathbf{w}$, we have:

$$\begin{aligned}\hat{\mathbf{y}}_i[j] &= \frac{\exp(-||\mathbf{z}_i - \mathbf{w}_j||^2/s_j^2)}{\sum_{j=1}^k \exp(-||\mathbf{z}_i - \mathbf{w}_j||^2/s_j^2)} \\ &\propto \frac{\exp(2\mathbf{z}_i\cdot\mathbf{w}_j/s_j^2)}{\sum_{j=1}^k \exp(2\mathbf{z}_i\cdot\mathbf{w}_j/s_j^2)}\end{aligned} \tag{12}$$

Combining Eqn. (12) with Eqn. (4), we rewrite the minimization of Eqn. 4) by fixing the term $\mathbf{y}_i$ as:

$$\min \mathcal{L}_{\text{GKD}} \iff \min -\sum_{i=1}^{|\mathcal{B}|} \log \frac{\exp(\mathbf{z}_i\cdot\mathbf{w}_i/\tau_i)}{\sum_{j=1}^k \exp(\mathbf{z}_i\cdot\mathbf{w}_j/\tau_j)}. \tag{13}$$

Previous clustering-guided contrastive losses [19], [30] share a similar form to Eqn. (4), where $\tau \propto s^2$ and $\mathbf{w}_i$ is the positive weight vector corresponding to the sample $\boldsymbol{X}_i$. Hence, these losses can be interpreted as a special case of our approach, where different clusters are assumed to have the same fixed concentration level, denoted as $\tau_j = \tau$ for $j = 1, 2, \cdots, k$.

2) SKD Loss: Our proposed SKD loss utilizes self-knowledge distillation in a novel way, by suppressing label noise in unsupervised or self-training settings. We now provide a detailed analysis of how self-distillation can handle noisy labels, by demonstrating that it is equivalent to smoothing hard labels for refinement with a dynamically estimated, class-specific prior distribution.

According to [49], label smoothing involves a weighted combination of the one-hot hard label vector and a pre-defined fixed distribution. The network output is denoted as:

$$\begin{aligned}\mathcal{L}_{\text{smoothing}} &= \sum_{i=1}^{|\mathcal{B}|} H(\widetilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i) \\ &= \sum_{i=1}^{|\mathcal{B}|} \alpha H(\mathbf{y}_i, \hat{\mathbf{y}}_i) + (1-\alpha) H(\mathbf{u}, \hat{\mathbf{y}}_i),\end{aligned} \tag{14}$$

where $\mathbf{u}$ is generally chosen as the uniform distribution and $\widetilde{\mathbf{y}}_i = \alpha \mathbf{y}_i + (1-\alpha)\mathbf{u}$ represents the smoothed label vector for the sample $\boldsymbol{X}_i$.

Looking back at Eqn. (6), we see that $\hat{\mathbf{y}}_i^+ = \mathcal{C}(\mathbf{q}_i^+ | \boldsymbol{W})$ is a fixed constant and does not contribute to the gradient computation during optimization. As a result, we can reformulate the overall loss function defined in Eqn. (7) as follows:

$$\begin{aligned}\mathcal{L}_{\text{DKD}} &= \alpha \mathcal{L}_{\text{GKD}} - (1-\alpha) \sum_{i=1}^{|\mathcal{B}|} \sum_{j \in \mathcal{P}_r(\hat{\mathbf{y}}_i^+)} \hat{\mathbf{y}}_i^+[j] \log \hat{\mathbf{y}}_i[j] \\ &= \sum_{i=1}^{|\mathcal{B}|} \alpha H(\mathbf{y}_i, \hat{\mathbf{y}}_i) + (1-\alpha) H(\mathbf{v}_i, \hat{\mathbf{y}}_i),\end{aligned} \tag{15}$$

where $\mathbf{v}_i$ is explained as a distribution vector derived from the positive anchor point $\mathbf{q}_i^+$, which corresponds to the sample $\boldsymbol{X}_i$. Specifically, the distribution vector $\mathbf{v}_i$ is formulated using Eqn. 16):

$$\mathbf{v}_i[j] = \begin{cases} \mathcal{C}(\mathbf{q}_i^+ | \boldsymbol{W})[j], & \text{if } j \in \mathcal{P}_r(\mathbf{q}_i^+) \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

By comparing Eqn. (15) with Eqn, (14), we observe that they share a similar form. Consequently, the SKD loss in our framework can be interpreted as a generalized form of label smoothing in two aspects: 1) The prior distribution $\mathbf{v}$ is dynamically estimated instead of being pre-defined, and 2) The prior distribution $\mathbf{v}$ is class-wise, introducing confidence diversity among classes, unlike the uniform distribution applied across samples in the mini-batch.

IV. Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the DKD-MPL approach. We aim to answer several research questions (**RQs**) as follows:

- **RQ1:** How does the DKD-MPL perform compared to state-of-the-art baselines?
- **RQ2:** How does the DKD-MPL perform with different trade-off parameter α values between the GKD loss and SKD loss?
- **RQ3:** How does the DKD-MPL perform with different choices of positive anchor points, specifically in terms of the hyperparameter r in the SKD loss?
- **RQ4:** How does the momentum coefficient affect the person Re-ID performance of the DKD-MPL?
- **RQ5:** How does the batch size affect the person Re-ID performance of the DKD-MPL?
- **RQ6:** How does the choice of backbone network employed in the DKD-MPL framework affect the person Re-ID performance?

A. Datasets and Evaluation Metrics

1) Datasets: To evaluate the re-identification performance of the proposed method, we use four public Re-ID benchmarks in the experiments including Market-1501 [50], DukeMTMC-ReID [51], MSMT17 [52] and VeRi-776 [53], which are also commonly used by most previous works. The details of these benchmarks are stated as follows.

Market-1501 [50] includes 32,668 labeled person images of 1,501 identities collected from six camera views. For evaluation, the dataset is divided into 12,936 images of 751 identities for training, 3,368 query images and 19,732 images of 705 identities for testing.

DukeMTMC-reID is a subset of the DukeMTMC dataset [51]. The dataset is divided with 16,522 images of 702 identities for training, 2,228 query images of 702 identities and 17,611 gallery images for testing.

MSMT17 [52] is composed of 126,411 person images from 4,101 identities collected by a 15-camera system. The training set consists of 32,621 images of 1,041 identities, and the testing set contains 11,659 images as the query and 82,161 images as the gallery.

VeRi-776 [53] collects vehicle images in the real-world urban surveillance scenario. The training set has 575 vehicles with 37,746 images and the testing set has 200 vehicles with 11,579 images, captured by 20 cameras.

2) Evaluation Metrics: Following the standard evaluation protocol outlined in [54] and [55], the performance metrics mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) are employed in the experiments. mAP is a comprehensive evaluation metric that considers both the precision of retrieval and the recall of pedestrian images. A higher mAP value indicates better model performance. On the other hand, CMC represents the results of identifying the correct target within the top k images in the retrieved ranking. For instance, "top-5" denotes the scenario where the correct target is still included in the top 5 images in the retrieved ranking.

B. Experimental Settings

1) Network Architecture: In our approach, we make use of ResNet-50 [56] as the backbone for the encoder $\mathcal{F}_\theta$. The model is initialized with pre-trained parameters from ImageNet [57]. After layer-4, we remove all sub-module layers and add a global average pooling layer followed by a batch normalization layer and an ℓ_2-normalization layer, resulting in 2048-dimensional features. During testing, we extract the features from the global average pooling layer to calculate the distance.

2) Parameter Setting: For the Market-1501 and DukeMTMC-reID datasets, the input images are resized to 256×128, while for MSMT17 and VeRi-776 datasets, the images are resized to 224×224. Each mini-batch consists of 256 images from 16 pseudo person identities, with 16 instances per person. We employ Adam optimizer with a

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS IN TERMS OF MAP AND CMC (%) ON MARKET-1501 DATASET.

Method	Reference	mAP	top-1	top-5	top-10
JDGL [11]	CVPR19	86.0	94.8	N/A	N/A
HOReID [12]	CVPR20	84.9	94.2	N/A	N/A
TransReID [13]	ICCV21	88.4	95.0	N/A	N/A
DCAL [15]	CVPR22	87.5	94.7	N/A	N/A
MMCL [17]	CVPR20	60.4	84.4	92.8	95.0
AD-Clus.++ [40]	CVPR20	68.3	86.7	94.4	96.5
MMT [18]	ICLR20	75.6	89.3	95.8	97.5
SPCL [19]	NeurIPS20	77.5	89.7	96.1	97.6
JGCL [22]	CVPR21	66.8	87.3	93.5	95.5
IDM [23]	ICCV21	82.8	93.2	97.5	98.1
MCRN [25]	AAAI22	83.8	93.8	97.5	98.5
BUC [26]	AAAI19	38.3	66.2	79.6	84.0
SSL [27]	CVPR20	37.8	71.7	83.8	87.4
MMCL [17]	CVPR20	45.5	80.3	89.4	92.3
HCT [28]	CVPR20	56.4	80.0	91.6	95.2
SPCL [19]	NeurIPS20	73.1	88.1	95.1	97.0
RLCC [29]	CVPR21	77.7	90.8	96.3	97.5
Cluster Contrast [30]	ACCV22	82.6	93.0	97.0	98.1
PPLR [31]	CVPR22	81.5	92.8	97.1	98.1
DKD-MPL	This work	**86.0**	**94.5**	**98.0**	**98.5**

weight decay of 5e-4. The initial learning rate is set to 3.5e-4 and is reduced to $1/10$ of its previous value every 20 epochs, for a total of 50 epochs. Similar to the clustering method in [30], we use DBSCAN and Jaccard distance [58] to cluster with 30 nearest neighbors. The parameter r is set to 1.2% for Market-1501 and DukeMTMC-reID, 0.8% for MSMT17 and VeRi-776. The balancing hyperparameter α is set as 0.8 for Market-1501, and slightly tuned for other datasets. For DBSCAN, the maximum distance between two samples is set as 0.5 for Market-1501 and DukeMTMC-reID, and 0.7 for MSMT17 and VeRi-776. The minimal number of neighbors in a core point is set as 4. We do not use any post-processing techniques such as re-ranking.

3) Implementation Details: During the training stage, at the start of each epoch, we use DBSCAN for clustering to generate pseudo labels. Any unclustered outlier samples are discarded from training for simplicity. We employ various data augmentation techniques, including random horizontal flipping, padding with 10 pixels, random cropping, and random erasing [59]. The proposed method is implemented using PyTorch and trained on a single GeForce RTX 2080TI GPU with 11GB RAM. The training process takes approximately 3 hours for each of the Market-1501, DukeMTMC-reID, and Veri-776 datasets. For the larger-scale MSMT17 dataset, the training process takes around 6 hours and utilizes 4 GPUs. Once training is complete, we utilize the feature extraction network parameterized by $\mathcal{F}_\theta$, as learned from Algorithm 1, for inference.

C. Comparison with State-of-The-Arts (RQ1)

In this section, we conduct experiments to answer **RQ1**, i.e., How does the DKD-MPL perform compared with the state-of-the-art baselines. For this, we compare our method with state-of-the-art Re-ID methods. The experimental results on multiple benchmark datasets are presented in Tables II, III, IV, and V. The state-of-the-art Re-ID methods used in the experiments include:

- **Supervised Learning Methods**: Joint Discriminative and Generative Learning (**JDGL**) [11], High-Order information for occluded person Re-ID (**HOReID**) [12], Transformer for person Re-ID (**TransReID**) [13], and Dual Cross-Attention Learning (**DCAL**) [15].
- **UDA methods:** Memory-based Multi-label Classification Loss (**MMCL**) [17], Augmented Discriminative Clustering (**AD-Cluster++**) [40], Mutual Mean-Teaching (**MMT**) [18], Self-Paced Contrastive Learning (**SPCL**) [19], Joint Generative and Contrastive Learning (**JGCL**) [22], Intermediate Domain Module (**IDM**) [23].
- **PUL methods:** Bottom-Up Clustering (**BUC**) [26], Softened Similarity Learning (**SSL**) [27], Hierarchical Clustering with hard-batch Triplet loss (**HCT**) [28], **SPCL** [19], Refining pseudo Labels with Clustering Consensus (**RLCC**) [29], and **Contrast Cluster** [30], and Part-based Pseudo Label Refinement (**PPLR**) [31] for unsupervised person re-identification are employed in the experiments. Besides, Progressive Adaptation Learning (**PAL**) [60] and Unsupervised Domain Adaptive Re-ID (**UDARe-ID**) [61] are used in the experiments on VeRi-776 vehicle dataset.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS IN TERMS OF MAP AND CMC (%) ON DUKEMTMC-REID DATASET.

Method	Reference	mAP	top-1	top-5	top-10
JDGL [11]	CVPR19	74.8	86.6	N/A	N/A
HOReID [12]	CVPR20	75.6	86.9	N/A	N/A
TransReID [13]	ICCV21	81.9	91.1	N/A	N/A
DCAL [15]	CVPR22	80.1	89.0	N/A	N/A
MMCL [17]	CVPR20	51.4	72.4	82.9	85.0
MMT [18]	ICLR20	65.1	78.9	88.8	92.5
SPCL [19]	NeurIPS20	68.8	82.9	90.1	92.5
JGCL [22]	CVPR21	62.8	82.9	87.1	88.5
IDM [23]	ICCV21	70.5	83.6	91.5	93.7
MCRN [25]	AAAI22	71.5	84.5	91.7	93.8
MMCL [17]	CVPR20	40.2	65.2	75.9	80.0
HCT [28]	CVPR20	50.7	69.6	83.4	87.4
SPCL [19]	NeurIPS20	65.3	81.2	90.3	92.2
RLCC [29]	CVPR21	69.2	83.2	91.6	92.2
Cluster Contrast [30]	ACCV22	72.8	85.7	92.0	93.5
DKD-MPL	This work	**75.9**	**86.8**	**92.8**	**94.0**

1) Comparison with PUL Methods: Our proposed DKD-MPL framework achieves the best performance among all the compared methods with mAPs of 86.0% and 75.9% on the Market-1501 and DukeMTMC-reID datasets, respectively. There are noticeable improvements of 3.4% and 3.1% mAPs compared to Cluster Contrast [30], which serves as the baseline for our method. For the more challenging MSMT17 benchmark, our DKD-MPL framework contributes to an impressive mAP of 38.0%. This outperforms the state-of-the-art Cluster Contrast method by a significant improvement of 4.7% mAP. These results demonstrate the effectiveness and superiority of our proposed DKD-MPL framework in comparison to existing methods. It showcases the potential of knowledge distillation and multiview learning in mitigating

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS IN TERMS OF MAP AND CMC (%) ON MSMT17 DATASET.

Method	Reference	mAP	top-1	top-5	top-10
TransReID [13]	ICCV21	63.6	82.5	N/A	N/A
DCAL [15]	CVPR22	64.0	83.1	N/A	N/A
MMCL [17]	CVPR20	16.2	43.6	54.3	58.9
MMT [18]	ICLR20	24.0	50.1	63.5	69.3
SPCL [19]	NeurIPS20	26.8	53.7	65.0	69.8
JGCL [22]	CVPR21	21.3	45.7	58.6	64.5
IDM [23]	ICCV21	33.5	61.3	73.9	78.4
MMCL [17]	CVPR20	11.2	35.4	44.8	49.8
SPCL [19]	NeurIPS20	19.1	42.3	55.6	61.2
RLCC [29]	CVPR21	27.9	56.5	68.4	73.1
Cluster Contrast [30]	ACCV22	33.3	63.3	73.7	77.8
PPLR [31]	CVPR22	31.4	61.1	73.4	77.8
DKD-MPL	This work	**38.0**	**68.1**	**79.0**	**81.7**

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART METHODS IN TERMS OF MAP AND CMC (%) ON VERI-776 DATASET.

Method	Reference	mAP	top-1	top-5	top-10
MMT [18]	ICLR20	35.3	74.6	82.6	87.0
UDARe-ID [61]	PR20	35.8	76.9	85.8	N/A
SPCL [19]	NeurIPS20	38.9	80.4	86.8	89.6
BUC [26]	AAAI19	21.2	54.7	70.4	N/A
SSL [27]	CVPR20	23.8	69.3	72.1	N/A
SPCL [19]	NeurIPS20	36.9	79.9	86.8	89.9
RLCC [29]	CVPR21	39.6	83.4	88.8	90.9
PAL [60]	IJCAI20	42.0	68.2	79.9	N/A
Cluster Contrast [30]	ACCV22	42.5	87.7	91.4	93.1
PPLR [31]	CVPR22	41.6	85.6	91.1	93.4
KDK-MPL	This work	**45.8**	**89.6**	**93.3**	**94.0**

issues related to supervision deviation and improving the performance of unsupervised methods.

2) Comparison with UDA Methods: The UDA methods can make full use of the labelled source domain datasets, so they usually achieve better results than the PUL Re-ID methods with the same framework. However, as the experimental results reveal, ours outperforms the state-of-the-art UDA methods on all of the four datasets, which demonstrates the superiority of our methods. Our method could also be easily generalized to be the UDA type. Since it does not focus on how to make use of the labelled source domain dataset, the labeled source dataset does not help much.

3) Comparison with Supervised Learning Methods: We also report the performance of the state-of-the-art supervised methods on Market-1501 and DukeMCMT-reID datasets for reference. There exists a clear performance gap between previous unsupervised methods and the supervised ones. It can be observed that our DKD-MPL yields a competitive Re-ID performance on both datasets over the supervised methods. This observation clearly shows the benefits of incorporating distance-based classification and self-distillation in the task of unsupervised Re-ID.

D. Ablation Studies (RQ2, RQ3)

In the ablation studies, we investigate how different components and hyperparameters of the DKD-MPL approach affect its performance in person Re-ID tasks. Specifically, we analyze two main hyperparameters, α and r, which control the balance between the GKD loss and SKD loss and the selection of positive anchors for the SKD loss, respectively.

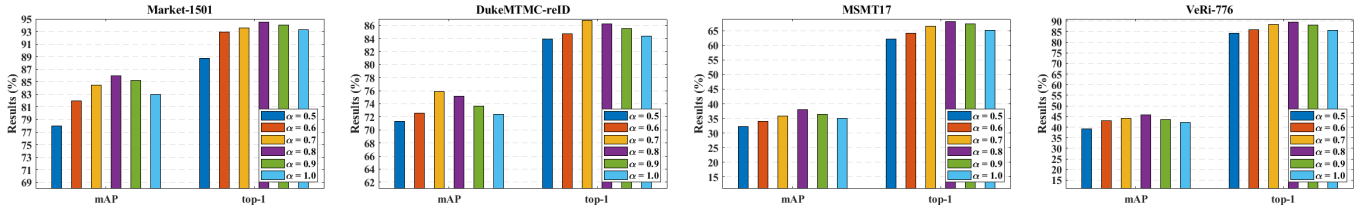
1) Trade-Off Parameter Investigation in DKD Loss (RQ2): To answer **RQ2** which concerns how does the DKD-MPL perform with different trade-off parameter α in DKD loss, we start with investigating the impact of α in Eqn. (7). To better evaluate the effect of the hyperparameter, we fixed the value of r while changing the value of α. Results on the four datasets are shown in Fig. 3. From Fig. 3, it is evident that different parameter values do have an impact on the results. For instance, when examining the results on the Market-1501 dataset, we observe that setting a larger value for α implies that self-distillation will only influence a small subset of the training samples. When α is overly small, self-distillation would dominate the training process and force each sample to have a very similar behaviour to the corresponding positive class prototype, which limits the ability and harms Re-ID performance.

2) Hyperparameter Investigation in SKD Loss (RQ3): The hyperparameter, r, determines the selection of positive anchors for the SKD loss. To address **RQ3**, which examines the performance of our method with different choices of positive anchors based on the hyperparameter r in the SKD loss, we analyze the impact of varying r values in Eqn. (6). Figure 4 illustrates the object re-identification results on the Market-1501, DukeMTMC-reID, MSMT17, and VeRi-776 datasets. From Fig. 4, we observe that the optimal performance is achieved when r is set to 1.2% on Market-1501 and DukeMTMC-reID, and 0.8% on MSMT17 and VeRi-776. Additionally, a gradual decline in performance can be observed as r increases further. This suggests that including too many classes for KL-divergence computation can be detrimental, as each training sample is unlikely to be strongly associated with a large number of clusters.

E. Training Strategy Discussion (RQ4, RQ5, RQ6)

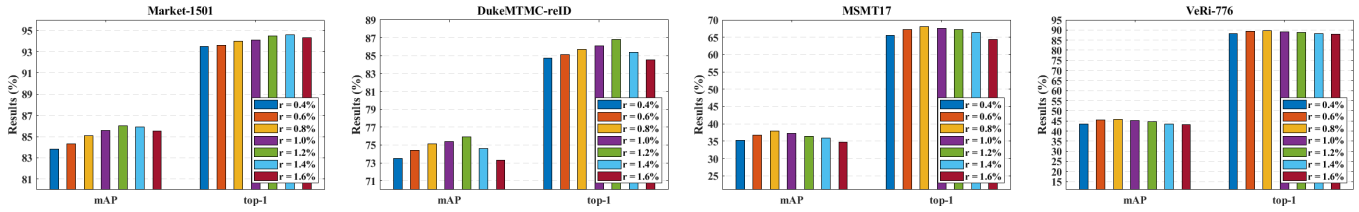
1) Impact of Momentum Coefficient (RQ4): As mentioned above, the weights of the classifier are updated via a momentum averaging formulation defined in Eqn. (9) where we introduce a monument coefficient ξ to control updating speed. Therefore, it is necessary to investigate the influence of the updating speed, which is also the concerns in **RQ4**. We report the experiment results on four datasets with a wide range of the monument coefficient ξ in Fig. 5. Note that when $\xi = 0$, the classifier would be never updated during training, which leads to inconsistency issues. Furthermore, as ξ becomes larger, the Re-ID results firstly increase and subsequently decrease, which indicates that simply increasing the updating is not always beneficial to improving the capacity of the framework.

2) Impact of Batch Size (RQ5): To explore the impact of batch size on our method and answer **RQ5**, we use different batch size settings to train our method. To explore the impact of batch size on our method, we use different batch size settings to train our method. According to the results in Table VI, our method demonstrates continuous performance improvement on Market-1501 as the batch size increases. A larger batch enables to provide more abundant samples of



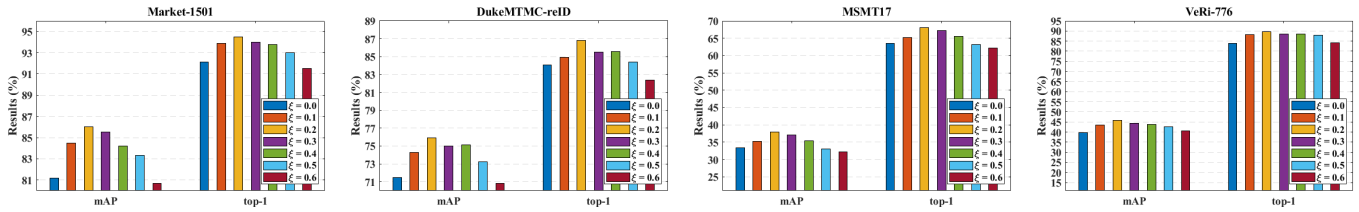
(a) (b) (c) (d)

Fig. 3. Comparison of Re-ID performance (%) on Market-1501, DukeMTMC-reID, MSMT17, and VeRi-776 datasets with varying α values.



(a) (b) (c) (d)

Fig. 4. Comparison of Re-ID performance (%) on Market-1501, DukeMTMC-reID, MSMT17, and VeRi-776 datasets with varying r values.



(a) (b) (c) (d)

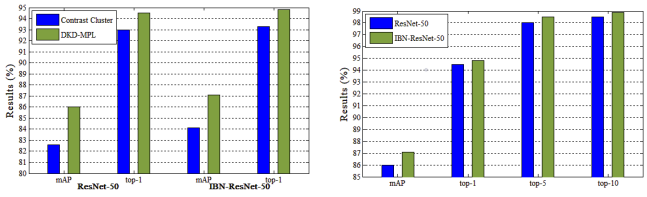
Fig. 5. Comparison of Re-ID performance (%) on Market-1501, DukeMTMC-reID, MSMT17, and VeRi-776 datasets with varying ξ values ranging from 0 to 0.6.

the same cluster which helps find more informative class prototypes to facilitate to suppress noisy supervision. Besides, it is worth mentioning that the performance of our method significantly better than the state-of-the-art Cluster Contrast [30] and SPCL [19] if the batch size is 64 or 256, which further demonstrates the superiority of our method.

TABLE VI
PERFORMANCE ON THE MARKET-1501 WITH DIFFERENT BATCH SIZES.

Methods	Batch Size	mAP	top-1
SPCL [19]	64	73.1	88.1
Cluster Contrast [30]	64	80.9	91.7
DKD-MPL	64	**83.0**	**92.8**
SPCL [19]	256	50.2	72.8
Cluster Contrast [30]	256	82.6	93.0
DKD-MPL	256	**86.0**	**94.5**

3) Impact of Backbone Network (RQ6): As for the architecture of the framework, theoretically, our method does not depend on a specific backbone network as the encoder $\mathcal{F}_\theta$. In order to investigate our method's dependence on the structure of the backbone network, which is also a primary concern in **RQ6**, we conduct experiments on two versions of ResNet-50 on Market-1501 dataset and report the results in Fig. 6. Compared with ResNet-50, IBN-ResNet-50 [62] combines the advantages of Instance Normalization [63] and Batch Normalization to extract appearance-invariant and content-related features. Therefore, using IBN-ResNet can further improve performance, which suggests that the representability of the backbone network contributes to the clustering performance.



(a) (b)

Fig. 6. Performance (%) of our method (a) vs. Cluster Contrast with ResNet-50 and IBN-ResNet-50 and (b) with ResNet-50 and IBN-ResNet-50 on Market-1501.

V. CONCLUSION

In this paper, we propose a dual knowledge distillation framework on multiview pseudo labels to address the supervision deviation problem caused by the clustering process in purely unsupervised person Re-ID tasks. To generate pseudo

labels with multiple views, we achieve them in different ways including using the same inputs with differing methodologies in the GKD module, and employing different inputs in the same way in the SKD module. To facilitate the overall loss, we propose a GKD loss to learn an embedding space, encoding the semantic structure of the data discovered by using the estimated distribution around prototypes for each cluster, and a SKD loss to effectively mitigate label noise on-the-fly for higher performance. Our experiments conducted on widely-used benchmarks demonstrated the effectiveness of the proposed DKD-MPL method. In this work, we consider person images from a single modality to apply knowledge distillation in the self-training unsupervised person Re-ID, which limits the potential for distilling knowledge across modalities and addressing the challenges associated with cross-modal person Re-ID. As part of our future work, we aim to explore the application of knowledge distillation in unsupervised cross-modal person Re-ID. This exploration could enhance the model's ability to comprehend and match individuals across multiple modalities.

ACKNOWLEDGMENTS

Thanks for the resources provided by the School of Engineering, Computer and Mathematical Sciences at Auckland University of Technology New Zealand during the visiting time. We are also deeply thankful to the anonymous reviewers for their careful review and insightful suggestions, which have been instrumental in improving the quality and clarity of our paper.

REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Analys. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2022.

[2] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 1294–1305, 2023.

[3] J. Mao, Y. Yao, Z. Sun, X. Huang, F. Shen, and H.-T. Shen, "Attention map guided transformer pruning for occluded person re-identification on edge device," *IEEE Trans. Multimedia*, vol. 25, pp. 1592–1599, 2023.

[4] P. Wang, C. Ding, Z. Shao, Z. Hong, S. Zhang, and D. Tao, "Quality-aware part models for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 3154–3165, 2023.

[5] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[6] Y. Feng, J. Yu, F. Chen, Y. Ji, F. Wu, S. Liu, and X.-Y. Jing, "Visible-infrared person re-identification via cross-modality interaction transformer," *IEEE Trans. Multimedia*, pp. 1–13, 2022.

[7] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, "Clothes-changing person re-identification with RGB modality only," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1060–1069, 2022.

[8] J. Chen, W.-S. Zheng, Q. Yang, J. Meng, R. Hong, and Q. Tian, "Deep shape-aware person re-identification for overcoming moderate clothing changes," *IEEE Trans. Multimedia*, vol. 24, pp. 4285–4300, 2022.

[9] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 2014–2023, 2021.

[10] Y. Yao, X. Jiang, H. Fujita, and Z. Fang, "A sparse graph wavelet convolution neural network for video-based person re-identification," *Pattern Recognit.*, vol. 129, p. 108708, 2022.

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

[12] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6449–6458, 2020.

[13] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 15 013–15 022, 2021.

[14] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, "Person retrieval in surveillance videos via deep attribute mining and reasoning," *IEEE Trans. Multimedia*, vol. 23, pp. 4376–4387, 2021.

[15] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 4692–4702, 2022.

[16] Y. Shi, Z. Wei, H. Ling, Z. Wang, P. Zhu, J. Shen, and P. Li, "Adaptive and robust partition learning for person retrieval with policy gradient," *IEEE Trans. Multimedia*, vol. 23, pp. 3264–3277, 2021.

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

[18] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *Proc. Int. Conf. Learn. Represent.*, 2020.

[19] Y. Ge, F. Zhu, D. Chen, R. Zhao *et al.*, "Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 11 309–11 321, 2020.

[20] H. Feng, M. Chen, J. Hu, D. Shen, H. Liu, and D. Cai, "Complementary pseudo labels for unsupervised domain adaptation on person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 2898–2907, 2021.

[21] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2814–2830, 2022.

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

[23] 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," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021.

[24] A. Verma, A. Subramanyam, Z. Wang, S. Satoh, and R. R. Shah, "Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation," *IEEE Trans. Multimedia*, vol. 25, pp. 364–377, 2023.

[25] Y. Wu, T. Huang, H. Yao, C. Zhang, Y. Shao, C. Han, C. Gao, and N. Sang, "Multi-centroid representation network for domain adaptive person re-id," *Proc. Assoc. Advancement Artif. Intell.*, vol. 36, no. 3, pp. 2750–2758, 2022.

[26] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," *Proc. Assoc. Advancement Artif. Intell.*, pp. 8738–8745, 2019.

[27] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020.

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

[29] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 3435–3444, 2021.

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

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

[32] T. Si, F. He, Z. Zhang, and Y. Duan, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 4323–4334, 2023.

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

[34] 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.
[35] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Comm. ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[36] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3238–3247, 2020.
[37] 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, 2023.
[38] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 6111–6120, 2019.
[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, 2019.
[40] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian, "AD-Cluster: Augmented discriminative clustering for domain adaptive person re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 9018–9027, 2020.
[41] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 8221–8230, 2019.
[42] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proc. Int. Conf. Knowledge Discovery and Data Mining*, vol. 96, no. 34, pp. 226–231, 1996.
[43] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. H. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, vol. 31, pp. 379–391, 2022.
[44] Y. Chen, M. Liu, X. Wang, F. Wang, A.-A. Liu, and Y. Wang, "Refining noisy labels with label reliability perception for person re-identification," *IEEE Trans. Multimedia*, pp. 1–12, 2023.
[45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[46] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 3713–3722, 2019.
[47] H. Lee, S. J. Hwang, and J. Shin, "Self-supervised label augmentation via input transformations," *Proc. Int. Conf. Mach. Learn.*, pp. 5714–5724, 2020.
[48] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Metric learning for large scale image classification: Generalizing to new classes at near-zero cost," *Proc. Eur. Conf. Comput. Vis.*, pp. 488–501, 2012.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2818–2826, 2016.
[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1116–1124, 2015.
[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *Proc. Eur. Conf. Comput. Vis.*, pp. 17–35, 2016.
[52] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 79–88, 2018.
[53] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," *Proc. Eur. Conf. Comput. Vis.*, pp. 869–884, 2016.
[54] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 1116–1124, 2015.
[55] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 3754–3762, 2017.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.
[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009.
[58] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1318–1327, 2017.
[59] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proc. Assoc. Advancement Artif. Intell.*, pp. 13 001–13 008, 2020.
[60] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, "Unsupervised vehicle re-identification with progressive adaptation," *Proc. Int. Joint Conf. Artif. Intell.*, pp. 913–919, 2021.
[61] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, p. 107173, 2020.
[62] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," *Proc. Eur. Conf. Comput. Vis.*, pp. 464–479, 2018.
[63] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6924–6932, 2017.

Wenjie Zhu (Member, IEEE) received the Master's degree from Xidian University, Xi'an, China, in 2013 and the Ph.D. degree from Northeastern University, Shenyang, China, in 2019. He visited the School of Engineering, Computer & Mathematical Sciences at Auckland University of Technology (AUT) in New Zealand from 2023 to 2024. Currently, he is a lecturer in the College of Information Engineering in China Jiliang University, Hangzhou, China. His current research interests include computer vision and machine learning.

Bo Peng received the Bachelor's degrees from China Jiliang University in Hangzhou, China, and Auckland University of Technology (AUT) in Auckland, New Zealand in 2020. He received the Master's degree from The University of Queensland in 2023. His research interests include pattern recognition and machine learning.

Wei Qi Yan (Senior Member, IEEE) is with AUT computer science, his expertise covers intelligent surveillance, deep learning, robotics, computer vision, and multimedia computing. Dr Yan has served as an Associate Editor of ACM Transactions on Multimedia Computing, Communications and Applications (TOMM), an Associate Editor of Frontiers in Neuroscience, an Associate Editor of Springer Nature Computer Science, the Editor-in-Chief (EiC) of the International Journal of Digital Crime and Forensics (IJDCF), he has worked as an exchange computer scientist between the Royal Society Te Apārangi (RSNZ) and the Chinese Academy of Sciences (CAS) in China. He is a guest (adjunct) professor at the Chinese Academy of Sciences and has been a visiting professor at the University of Auckland in New Zealand and the National University of Singapore. In 2022, Dr. Yan was recognised as one of the world's top 2% cited scientists by Stanford University, USA. He currently holds the position of Chair of ACM Multimedia Chapter of New Zealand and a member of the ACM, a senior member of the IEEE and a TC member of the IEEE.