International Journal of Computer Vision (2022) 130:2924–2939
https://doi.org/10.1007/s11263-022-01680-y

Unsupervised Person Re-Identification via Multi-Label Classification

Dongkai Wang[1] · Shiliang Zhang[1,2]

Received: 19 August 2021 / Accepted: 27 August 2022 / Published online: 20 September 2022

Abstract

The challenge of unsupervised person re-identification (ReID) lies in learning discriminative features without true labels. Most of previous works predict single-class pseudo labels through clustering. To improve the quality of generated pseudo labels, this paper formulates unsupervised person ReID as a multi-label classification task to progressively seek true labels. Our method starts by assigning each person image with a single-class label, then evolves to multi-label classification by leveraging the updated ReID model for label prediction. We first investigate the effect of precision and recall rates of pseudo labels to the ReID accuracy. This study motivates the Clustering-guided Multi-class Label Prediction (CMLP), which adopts clustering and cycle consistency to ensure high recall rate and reasonably good precision rate in pseudo labels. To boost the unsupervised learning efficiency, we further propose the Memory-based Multi-label Classification Loss (MMCL). MMCL works with memory-based non-parametric classifier and integrates local loss and global loss to seek high optimization efficiency. CMLP and MMCL work iteratively and substantially boost the ReID performance. Experiments on several large-scale person ReID datasets demonstrate the superiority of our method in unsupervised person ReID. For instance, with fully unsupervised setting we achieve rank-1 accuracy of 90.1% on *Market-1501*, already outperforming many transfer learning and supervised learning methods.

Keywords Person re-identification · Unsupervised learning · Multi-label classification

1 Introduction

Recent years have witnessed the great success of person re-identification (ReID), which learns discriminative features from labeled images with deep Convolutional Neural Network (CNN) (Zheng et al., 2016; Krizhevsky et al., 2012; He et al., 2016). Because it is expensive to annotate person images across multiple cameras, recent research efforts start to focus on unsupervised person ReID. The challenge of unsupervised person ReID lies in learning discriminative features without true labels. To conquer this challenge, most of recent works (Lin et al., 2018; Wang et al., 2018; Yu et al., 2019; Wei et al., 2018; Deng et al., 2018; Zhong et al., 2018a, 2019; Chen et al., 2019; Wu et al., 2019; Li et al., 2019; Qi et al., 2019; Zhang et al., 2019) define unsupervised person ReID as a transfer learning task, which leverages labeled data on other domains for model initialization or label transfer. Among them, some works minimize the discrepancy between source and target domain to a common feature space (Wei et al., 2018; Zhong et al., 2019; Lv et al., 2018; Lin et al., 2018; Wang et al., 2018). Some others estimate the pseudo labels for the target dataset (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a). Detailed review of existing methods will be presented in Sect. 2.

Thanks to the above efforts, the performance of unsupervised person ReID has been significantly boosted. However, there is still a considerable gap between supervised and unsupervised person ReID performance. Meanwhile, the setting of transfer learning leads to limited flexibility. As discussed in many works (Long et al., 2015; Yan et al., 2017; Wei et al., 2018), the performance of transfer learning is closely related to the domain gap between source and target datasets, *e.g.*, large domain gap degrades the performance on target dataset. It is non-trivial to estimate the domain gap and select suitable

Communicated by Rynson W.H. Lau.

✉ Shiliang Zhang
slzhang.jdl@pku.edu.cn

Dongkai Wang
dongkai.wang@pku.edu.cn

[1] School of Computer Science, Peking University, Beijing, China

[2] Peng Cheng Laboratory, Shenzhen, Guangdong, China

source datasets for transfer learning in unsupervised person ReID.

This paper targets to boost the performance of unsupervised person ReID without leveraging any manual annotations. Previous works mainly adopt single-class label of contrastive learning (Wu et al., 2018) or clustering to generate pseudo labels (Lin et al., 2019). Single-class labels cannot represent sample similarity well. Iterative clustering suffers from noise accumulation. To address those issues, our method uses classification model to predict multi-class pseudo labels. As illustrated in Fig. 1, we treat each unlabeled person image as a class, and train the ReID model to assign each image with a multi-class label. In other words, the ReID model is trained to classify each image to multiple positive classes belonging to the same identity. As each person has multiple images, this framework effectively identifies images of the same identity and differentiates images of different identities. This in-turn facilitates the ReID model to optimize inter and intra class similarities. Compared with previous methods (Lin et al., 2019; Wu et al., 2018), which classify each image into a single class, the multi-label classification has potential to exhibit better efficiency and accuracy for person feature learning. It also avoids the noise accumulation in sample clustering by leveraging the updated ReID model for label prediction.

As illustrated in Fig. 1, the proposed framework involves two components: 1) multi-class label prediction to assign pseudo labels for each person image, and 2) multi-label classification loss to measure the discrepancies between predicted multi-class labels and classification outputs. To enable a positive feedback for CNN training, predicted pseudo labels are expected to be more accurate than classifier outputs. Meanwhile, the multi-label classification loss should be fast to compute, because treating each image as a class leads to a large number of classes. Those two components make it critical to design accurate multi-class label prediction algorithm and effective loss function.

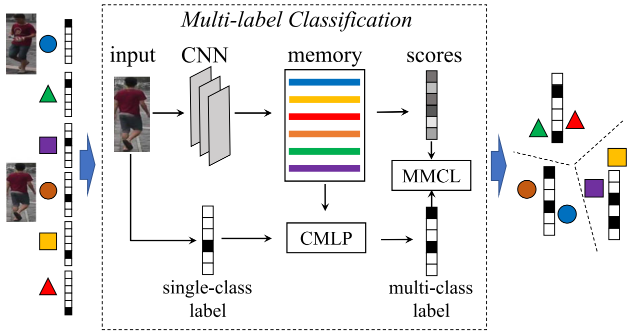


Fig. 1 Illustrations of the proposed multi-label classification for unsupervised person ReID. We target to assign each unlabeled person image with a multi-class label reflecting the person identity. This is achieved by iteratively running CMLP for multi-class label prediction and MMCL for multi-label classification loss computation. This procedure guides CNN to produce discriminative features for ReID

We investigated different algorithms for multi-class pseudo label generation. For instance, the multi-class label of an image can be predicted by treating its visually similar images as positive classes using K-Nearest Neighbor search (*KNN*) (Zhong et al., 2019), Similarity Score comparison (*SS*) (Fan et al., 2018D), or Clustering (*Clustering*) (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a), respectively. Fig. 2 illustrates precision and recall rate curves of predicted positive classes by different algorithms, as well as their ReID mAP (mean Average Precision) optimized by existing multi-label classification loss functions (Zhang & Zhou, 2013; Durand et al., 2019). It can be observed that, those prediction algorithms present different precision-recall patterns. *Clustering* tends to find more positive classes, thus produces high recall and low precision rates. *SS* gets higher precision and lower recall rates. Different precision-recall patterns correspond to different performance, *i.e.*, the ReID mAP presents stronger correlation to the recall rate than to precision. Too low precision degrades the mAP because considerable noises are involved in loss computation. This observation is consistent with those found in noisy label learning works, *i.e.*, noisy labels lead to overfitting and significantly degrade the performance (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Han et al., 2018).

The above observation motivates us to predict pseudo labels with high recall rate and reasonably good precision rate. We propose the Clustering-guided Multi-class Label Prediction (CMLP) algorithm. CMLP first utilizes *Clustering* to adaptively predict the number of positive classes for each image. This strategy effectively ensures a high recall rate, *e.g.*, having more neighbors leads to a larger numbers

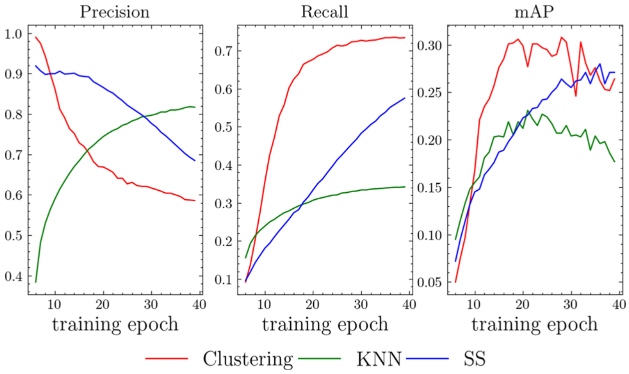


Fig. 2 Precision and recall rate curves of predicted positive classes by *KNN* (Zhong et al., 2019), *SS* (Fan et al., 2018D), *Clustering* (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a), as well as their corresponding ReID mAP on the *Market-1501* dataset. It is clear that, the ReID mAP is more sensitive to the recall rate than to the precision. Too low precision means involving considerable noises, thus degrades the mAP as the training goes on

of positive classes for an image. To further ensure a good precision rate, CMLP filters noisy classes with cycle consistency, which considers neighborhood similarity to spot and remove noisy labels.

Predicted pseudo labels allow for model training with a proper multi-label classification loss. Treating each image as a class leads to a huge number of classes, making it expensive to train classifiers like Fully Connected (FC) layers. We thus further study an efficient multi-label classification loss function. As shown in Fig. 1, we adopt the feature of each image as a classifier and store it in a memory bank. By updating the memory bank, we achieve an efficient classifier learning. Based on the memory bank, we propose the Memory-based Multi-label Classification Loss (MMCL). MMCL addresses the vanishing gradient issue in traditional multi-label classification loss (Zhang & Zhou, 2013; Durand et al., 2019) by abandoning the sigmoid function and enforcing the classification score to 1 or -1. Besides treating each class separately to compute the local loss on sampled classes, MMCL further computes a global loss to enforce the CNN output a uniform distribution on positive classes. This constraint involves all classes in loss computation and leads to a better training efficiency and model performance. MMCL facilitates the learning of discriminative features, which in turn provide high quality sample similarity for label prediction in CMLP.

We test our approach on several large-scale person ReID datasets, *i.e.*, *Market-1501* (Zheng et al., 2015) and *DukeMTMC-reID* (Ristani et al., 2016) without leveraging labeled data. Extensive ablation studies show that multi-class labels predicted by CMLP present high recall rate and reasonable good precision rate. By iteratively computing CMLP and MMCL, our training procedure is stable and addresses the mAP degradation issue in Fig. 2. Besides that, our method is easy to tune and repeat, *e.g.*, it is not sensitive to the selection of parameters, and generalizes well on different datasets. Comparisons with recent works show our method achieves promising performance. For instance, we achieve rank-1 accuracy of 90.1% on *Market-1501*, significantly outperforming the recent SSL (Lin et al., 2020) and HCT (Zeng et al., 2020) by 19.4% and 10.1%, respectively. Our performance is also better than the AD-Cluster(Zhai et al., 2020) and ECN++(Zhong et al., 2020), which use *DukeMTMC-reID* (Ristani et al., 2016) for transfer learning.

In summary, our method iteratively runs multi-class label prediction and computes multi-label classification loss to seek true labels for multi-label classification and CNN training. As shown in experiments, our algorithm, although does not leverage any labeled data, achieves competitive performance. The maintained memory bank reinforces both label prediction and classification. To the best of our knowledge, this is an original work treating unsupervised person ReID as a multi-label classification task. Compared with our conference version, this journal version investigates the effect of precision and recall rates of predicted labels to the ReID accuracy. This study leads to the updated CMLP and MMCL for label prediction and loss computation. CMLP produces multi-class pseudo labels with high recall rate and reasonably good precision. MMCL addresses issues in traditional MCL and integrates local loss and global loss to achieve high optimization efficiency. Those new algorithms further exploit the promising performance of this multi-label classification framework, *e.g.*, they significantly boost the rank-1 accuracy on *Market-1501* from 80.3 to 90.1%. Our work also shows that, unsupervised training has the potential to achieve better flexibility and accuracy than existing transfer learning and supervised learning strategies for person ReID model training.

2 Related Work

This work is related with unsupervised person ReID, unsupervised feature learning, and noisy label learning. This section briefly reviews related works in those three categories, respectively.

Unsupervised Person Re-identification learns ReID models on unlabeled target domains. Most of related works can be summarized into two categories, *i.e.*, fully unsupervised ReID methods, and transfer learning based methods, respectively. The first category only utilizes unlabeled data for model learning. Traditional fully unsupervised ReID methods include designing hand-crafted features (Liao et al., 2015; Zheng et al., 2015), exploiting localized salience statistics (Zhao et al., 2013; Wang et al., 2014) or dictionary learning based methods (Kodirov et al., 2015), respectively. These methods stem from the hand-crafted features, which are not discriminative enough to exploit complex scenarios. Recently, there appear some CNN-based unsupervised ReID methods. Lin *et al.* (Lin et al., 2019) propose a bottom-up clustering method to firstly merge similar samples, then learn discriminative features with deep network. Ding *et al.* (Ding et al., 2019) improve previous work by proposing a better clustering criterion. CNN-based methods have significantly outperformed traditional hand-crafted feature based methods.

The second category makes use of the knowledge of source labeled domain to train model on the target unlabeled domain, *e.g.*, (Lin et al., 2018; Wang et al., 2018; Yu et al., 2019; Wei et al., 2018; Deng et al., 2018; Zhong et al., 2018a, 2019; Chen et al., 2019; Wu et al., 2019; Li et al., 2019; Qi et al., 2019; Zhang et al., 2019). These works can be further divided into two groups: 1) reducing the source-target discrepancy in a common feature space, and 2) assigning pseudo labels for the target dataset. Methods of the first group aim to align the distribution of source and target domains in a common space. PTGAN (Wei et al., 2018) and SPGAN (Deng et al.,

2018) minimize the discrepancy by generating images with source identity and target domain style to train the model. HHL (Zhong et al., 2018a) minimizes target intra-domain discrepancy by generating images under different cameras. ECN (Zhong et al., 2019) utilizes transfer learning and minimizes the target invariance. Some other works (Lin et al., 2018; Wang et al., 2018) minimize the attribute-level discrepancy by utilizing extra attribute annotations. The other group of methods starts by initializing a ReID model on the labeled dataset, then adopts different strategies to assign pseudo labels for the target unlabeled dataset, including clustering (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a) and associating with source labeled dataset (Yu et al., 2019). To avoid noisy labels in clustering, Ge *et al.* (Ge et al., 2020a) propose a mutual mean teacher method to relieve the effect of noisy labels.

Unsupervised Feature Learning aims to learn from data without human-provided labels. Recent works on unsupervised feature learning mainly falls into three categories. The first category utilizes generative models, which aim to reconstruct the distribution of data. Classical generative models include Restricted Bolztmann Machines (RBM) (Hinton et al., 2006; Tang et al., 2012), and auto-encoder (Vincent et al., 2008; Le, 2013). The latent features of generative models could be used for downstream tasks such as image classification. Recently, there appear some more advanced generative models, including the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and variational auto-encoder (Kingma & Welling, 2013). They can approximate the data distribution better and produce more realistic samples.

The second category exploits data internal structures to infer supervision cues. Therefore, it is also called as self-supervised learning. For example, Komodakis *et al.* (Komodakis & Gidaris, 2018) predict the rotation of image to learn a good representation. Li *et al.* (Li et al., 2018) use motion and view as supervision to learn an initialization for action recognition. Wu *et al.* (Wu et al., 2018) regard each image as a single class, and propose a non-parametric classifier to train CNN. He *et al.* (He et al., 2019) propose momentum updated encoder and dynamic queue for contrastive learning and obtain superior performance to supervised initialization. Self-supervised learning usually learns generalized features and models, which can be used as an initialization for downstream tasks.

The third category aims to learn discrimintive features directly for classification or retrieval tasks, thus can relieve the requirement on labeled data. Iscen *et al.* (Iscen et al., 2018) utilize manifold learning to seek positive and negative samples to compute the triplet loss. Ye *et al.* (Ye et al., 2019) learn instance-aware features in a mini-batch using the newly-updated features. Our work shares certain similarity with (Wu et al., 2018; Ye et al., 2019), in that we also treat each image as a single class. However, we consider multi-label classification, which is important in identifying images of the same identity as well as differentiating different identities.

Learning with Noisy Label has been exploited by deep learning community in image classification, object detection and tracking. A line of methods aims to design robust loss functions against label noises. Ghosh *et al.* (Ghosh et al., 2017) design a Mean Absolute Error (MAE) loss to train model with noisy labels. Zhang *et al.* (Zhang & Sabuncu, 2018) analyze the drawbacks of MAE and cross-entropy loss in noisy label learning and further propose a Generalized Cross Entropy (GCE) loss. Label Smoothing Regularization (LSR) (Szegedy et al., 2016) is also a way to address noisy labels. Li *et al.* (Li et al., 2020) propose a cleanliness score based on classification and regression outputs to serve as the soft label and reweighting factor for object detector training. This method avoids noisy assignment in traditional IoU based metrics. Wu *et al.* (Wu et al., 2021) propose a noise-robust loss to deal with unsupervised tracking learning.

Another kind of methods focuses on refining the training strategies for noisy label learning. Co-teaching (Han et al., 2018) and co-training (Ma et al., 2017) train two collaborative models and each model chooses high confidence training samples for the other. These training strategies can avoid noise accumulation. Probabilistic modeling in uncertainty learning is also a way to deal with label noises. Danelljan *et al.* (Danelljan et al., 2020) model the label noise and ambiguities through a probabilistic distribution and propose a novel model to estimate an interpretational score for tracking. Neverova *et al.* (Neverova et al., 2019) explicitly use a higher-order uncertainty model to model the aleatoric uncertainty in annotations and use two models for densepose estimation.

Our approach uses unlabeled data for learning, thus belongs to the fully unsupervised ReID. It differs with previous methods in that, we formulate unsupervised ReID as a multi-label classification task. Multi-label classification is designed to cope with multi-class labels (Zhang & Zhou, 2013; Durand et al., 2019; Wang et al., 2018; Lin et al., 2018). Durand *et al.* (Durand et al., 2019) deal with multi-label learning based on partial labels and utilize graph neural network to predict missing labels. Wang *et al.* (Wang et al., 2018; Lin et al., 2018) use multi-label classification to learn attribute features. This paper utilizes multi-label classification to predict multi-class labels for learning person identity features. To the best of our knowledge, this is an original work introducing multi-label classification into unsupervised person ReID. It potentials have been demonstrated by the promising performance of this work.

3 Formulation

Given a person image dataset $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ without ID annotation, our goal is to train a person ReID model on $\mathcal{X}$. For any query person image q, the person ReID model is expected to produce a feature vector to retrieve image g containing the same person from a gallery set G. In other words, the ReID model should guarantee q share more similar feature with g than with other images in G. We could conceptually denote the goal of person ReID as,

$$g^* = \arg\min_{g \in G} \text{dist}(f_g, f_q), \tag{1}$$

where $f \in \mathbb{R}^d$ is a d-dimensional L2-normalized feature vector extracted by the person ReID model. dist$(\cdot)$ is the distance metric, *e.g.*, the L2 distance.

To make training on $\mathcal{X}$ possible, we start by treating each image as an individual class and assign x_i with a label y_i. This pseudo label turns $\mathcal{X}$ into a labeled dataset, and allows for the ReID model training. y_i is initialized to a two-valued vector, where only the value at index i is set to 1 and the others are set to -1, *i.e.*,

$$y_i[j] = \begin{cases} 1 & j = i \\ -1 & j \neq i \end{cases} \tag{2}$$

Since each person may has multiple images in $\mathcal{X}$, the initial label vector is not valid in representing person identity cues. Multi-class label prediction is required to assign multi-class labels to each image, which can be used for ReID model training with a multi-label classification loss. Labels of x_i can be predicted by matching its feature f_i to features of other images, and finding consistent feature groups. With multi-class labels, we can compute the loss by comparing predicted labels and classifier outputs. Due to the huge number of image classes in $\mathcal{X}$, it is hard to train multi-label classifiers. One efficient solution is to use the f_i as the classifier for the i-th class. This computes the classification score for any image x_j as,

$$c_j[i] = f_i^\top \times f_j, \tag{3}$$

where c_j denotes the multi-label classification score for x_j.

It is easy to infer that, both label prediction and multi-label classification require features of images in $\mathcal{X}$. We hence introduce a $n \times d$ sized memory bank $\mathcal{M}$ to store those features, where $\mathcal{M}[i] = f_i$. With $\mathcal{M}$, we propose the Clustering-guided Multi-class Label Prediction (CMLP) for label prediction and Memory-based Multi-label Classification Loss (MMCL) for ReID model training, respectively.

CMLP takes a single-class label as input and outputs the updated multi-label prediction $\bar{y}_i$ based on memory bank $\mathcal{M}$, *i.e.*,

$$\bar{y}_i = \text{CMLP}(y_i, \mathcal{M}), \tag{4}$$

where $\bar{y}$ is the multi-class label.

MMCL takes the image feature f, predicted multi-class label $\bar{y}$, and the memory bank $\mathcal{M}$ as inputs. The computed loss $\mathcal{L}$ can be represented as,

$$\mathcal{L} = \sum_{i=1}^{n} \text{MMCL}(\mathcal{M}, f_i, \bar{y}_i), \tag{5}$$

where f_i, $\bar{y}_i$ denote the image feature and pseudo multi-class label of image x_i.

$\mathcal{M}$ is updated after each training iteration as,

$$\mathcal{M}[i]^t = \alpha \cdot \mathcal{M}[i]^{t-1} + (1 - \alpha) \cdot f_i, \tag{6}$$

where the superscript t denotes the t-th training epoch, α is the updating rate. $\mathcal{M}[i]^t$ is then L2-normalized by $\mathcal{M}[i]^t \leftarrow ||\mathcal{M}[i]^t||_2$. Both label prediction and loss computation require robust features in $\mathcal{M}$ to seek reliable labels and classification scores, respectively. We use many data argumentation techniques to reinforce $\mathcal{M}$. In other words, each $\mathcal{M}[i]$ combines features of different augmented samples form x_i, it hence presents better robustness. More details are given in Sect. 5.2.

With $\mathcal{M}$, CMLP uses both clustering and cycle consistency to predict $\bar{y}_i$, making it more accurate than the classification score. This ensures the loss computed by MMCL valid in boosting the ReID model, which in-turn produces positive feedbacks to $\mathcal{M}[i]$ and label prediction. This $\mathcal{M}$-CMLP-MMCL loop makes it possible to train discriminative ReID models on unlabeled dataset. Implementations to CMLP and MMCL can be found in the following parts.

4 Proposed Methods

4.1 Clustering-guided Multi-class Label Prediction

The accuracy of predicted positive classes in $\bar{y}_i$ by Eq. (4) can be measured with the precision and recall rates, *i.e.*, finding more true positive classes leads to a high recall, and eliminating false positive classes improves the precision. It is hard to guarantee 100% precision and recall rates in $\bar{y}_i$. We thus need to study a proper optimization objective for label prediction, *e.g.*, emphasizing on the promotion of recall, precision, or both. This part first investigates a reasonable optimization objective for label prediction, hence presents details of our Clustering-guided Multi-class Label Prediction (CMLP).

Analysis of existing methods: Most of label prediction algorithms share similar spirit, *i.e.*, assigning similar images

Table 1 Illustration to the effect of precision and recall rates of positive classes to ReID mAP. *+/-x%* denotes the operation of increasing/decreasing the percentage of predicted positive classes.

Method	Operation	Avg. *P*	Avg. *R*	Avg. *mAP*
KNN	-	66.16	27.03	41.8
	+20%	55.19	32.61	40.30
	-20%	76.57	16.77	30.5
SS	-	80.61	31.94	44.4
	+20%	68.42	39.84	45.9
	-20%	89.48	9.81	27.0
Clustering	-	70.36	54.14	57.0
	+20%	62.89	57.28	57.9
	-20%	74.24	42.84	53.4

P and *R* denote precision and recall rate, respectively. Experiments are conducted on *Market-1501*

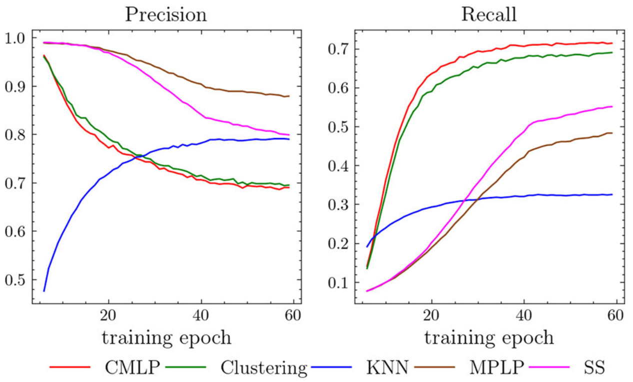


Fig. 3 Precision and recall curves of predicted positive classes by different algorithms on *Market-1501*. Tested label prediction algorithms include *KNN* (Zhong et al., 2019), *SS* (Fan et al., 2018D), *Clustering* (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a), the *MPLP* (Wang & Zhang, 2020) in our conference version, and our *CMLP*

with similar labels. One challenge in multi-class label prediction is confirming the number of similar images. *KNN* and *SS* select positive classes according to a fixed threshold, *i.e.*, the nearest neighbour number and a similarity, respectively. *Clustering* adaptively selects positive classes according to the cluster size. Figure 2 indicates that, the promotion of recall rate is more beneficial to boost the ReID mAP than improving precision rate. We conduct another experiment in Table 1 to further verify this observation. Table 1 increases/decreases the number of positive classes of existing label prediction algorithms. Intuitively, introducing more positive classes degrades the precision, but improves the recall and consistently boosts the mAP. Similar conclusion can be observed for different label prediction algorithms.

It is not hard to explain the above observations. Increasing the recall rate assigns more diverse samples with the same label, which in-turn boosts the power of ReID model in identifying visually different samples as the same person. Simply increasing the recall rate may lead to very low precision and substantial performance degradation, especially for large training epoch as illustrated in Fig. 2. This phenomenon can be explained by the "memorization" effect of deep networks (Arpit et al., 2017). Namely, when trained with noisy labels, deep networks will learn easy patterns in initial epochs (Arpit et al., 2017; Zhang et al., 2016), resulting in a steady performance promotion. As the the training goes on, deep networks will tend to overfit to noisy labels by the hard constraint of training loss, which causes performance degradation.

CMLP is thus proposed to predict positive classes with high recall and reasonably good precision rates. This is achieved by adaptively confirming the number of positive classes by clustering image features and using the cluster size as a reference number of positive classes. The multi-class label of each image is hence predicted based on an image rank list. The following part presents the computations of the reference number and positive classes, respectively.

Reference number computation Given all image features stored in the memory bank $\mathcal{M}$, we compute the distance matrix D using Euclidean distance, *i.e.*, the distance $D_{i,j}$ of images x_i and x_j is computed as

$$D_{i,j} = ||\mathcal{M}[i] - \mathcal{M}[j]||_2. \tag{7}$$

With D, we further consider *Rerank* to refine the feature distance. Specifically, *Rerank* refines the distance matrix using k-reciprocal encoding (Zhong et al., 2017),

$$D'_{i,j} = 1 - \frac{\sum_{m=1}^{n} \min(\mathrm{MK}(i,m), \mathrm{MK}(j,m))}{\sum_{m=1}^{n} \max(\mathrm{MK}(i,m), \mathrm{MK}(j,m))}, \tag{8}$$

where $\mathrm{MK}(i,m)$ is an indicator function, showing whether images x_i and x_m are mutual k-nearest neighbors.

The updated distance matrix D' can be adopted to generate image clusters. We choose the widely used DBSCAN (Ester et al., 1996) as the clustering algorithm. For an image $x_i \in C_k$, where C_k is a cluster, we can obtain the reference number of its positive classes n_i as the size of C_k, *i.e.*, $n_i = |C_k|$.

Multi-class label prediction For an image x_i, CMLP computes an image rank list R_i by sorting the distance stored in $D'_{i,[1:n]}$. Top ranked indexes in R_i can be selected as candidate positive classes of x_i. To boost the recall rate, CMLP tends to select more candidate positive classes from R_i. This can be simply achieved by increasing the reference positive class number as $n'_i = n_i + e$, where e is a parameter related to the recall rate. Based on n'_i and R_i, we select the initial positive classes for image x_i as $P_i = R_i[1 : n'_i]$.

To ensure a reasonably good precision rate, CMLP adopts a filtering strategy to spot and delete false positive classes from P_i. Inspired by k-reciprocal nearest neighbor(Jegou et al., 2007; Zhong et al., 2017), we assume that, if two images

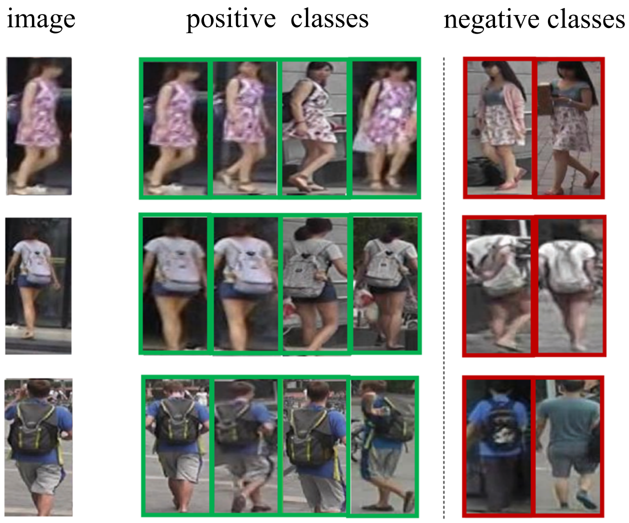


Fig. 4 Visualization of predicted positive and negative classes by *CMLP* on *Market-1501*. True positive and true negative are annotated by green and red bounding boxes, respectively (Color figure online)

belong to the same class, their neighboring image sets should also be similar. In other words, two images should be mutual neighbor for each other if they can be assigned with similar labels. With this intuition, we filter hard negative classes in P_i by referring to the cycle consistency.

Specifically, for an image x_j with $j \in P_i$, CMLP computes its top-n_i' nearest neighbors according D'. If x_i is one of the top-n_i' nearest neighbors of x_j, j is considered as a reliable positive class for x_i. Otherwise, it is treated as a hard negative class. We filter out these hard negative classes and obtain the final positive class set P_i^*. As P_i^* contains l classes, x_i would be assigned with a multi-class label $\bar{y}_i$ with l positive classes, *i.e.*,

$$\bar{y}_i[j] = \begin{cases} 1 & j \in P_i^* \\ -1 & j \notin P_i^* \end{cases} \tag{9}$$

Figure 3 compares the precision-recall patterns of different label prediction algorithms. It is clear that, CMLP achieves the highest recall rate, meanwhile it maintains a reasonable precision rate larger than 0.7. Compared with *Clustering*, CMLP gets more accurate predictions, *i.e.*, better recall and similar precision. Visualization of predicted positive and negative classes are illustrated in Fig. 4. More experimental evaluations will be presented in Sect. 5.

4.2 Memory-based Multi-label Classification Loss

The predicted multi-class labels can be used for training the ReID model with a multi-label classification loss. This part first discusses issues in traditional multi-label classification loss, then proceeds to introduce our Memory-based Multi-label Classification Loss (MMCL). Different from our conference version (Wang & Zhang, 2020), the improved MMCL involves a local and global loss to enhance the efficiency of optimization. As shown in our experiments, the updated MMCL brings substantial performance gains. More details can be found in Sect. 5.4

Traditional multi-label classification loss In traditional multi-label classification methods, sigmoid and logistic regression loss is a common option (Zhang & Zhou, 2013; Durand et al., 2019; Wang et al., 2018; Lin et al., 2018). For a task with n classes, it adopts n independent binary classifiers for classification. The loss of classifying image x_i to class j can be computed as,

$$\ell(j|x_i) = \log(1 + \exp(-\bar{y}_i[j] \times s_i[j])), \tag{10}$$

where $s_i[j] = \mathcal{M}[j]^\top \times f_i$ computes the classification score of image x_i for the class j. $\bar{y}_i[j]$ is the label of image x_i for class j. With the loss at a single class, we can obtain the Multi-Label Classification (MCL) loss, *i.e.*, $\mathcal{L}_{MCL}$,

$$\mathcal{L}_{MCL} = \sum_{i=1}^{n}\sum_{j=1}^{n} \ell(j|x_i), \tag{11}$$

where n is the number of images in the dataset $\mathcal{X}$, which equals to the class number in our setting.

Because the $\mathcal{M}[j]^\top$ and f_i are L2 normalized, the classification score is restricted between $[-1, 1]$. This limits the range of sigmoid function in Eq. (11), making the loss non-zero even for correct classifications. This issue can be addressed by introducing a scalar τ on the classification score. This updates Eq. (11) as,

$$\ell_\tau(j|x_i) = \log(1 + \exp(-\bar{y}_i[j] \times s_i[j]/\tau)). \tag{12}$$

We denote the corresponding MCL loss as $\mathcal{L}_{MCL-\tau}$. The gradient of $\mathcal{L}_{MCL-\tau}$ can be computed as,

$$\frac{\partial \mathcal{L}_{MCL-\tau}}{\partial f_i} = -\frac{\exp(-\bar{y}_i[j]s_i[j]/\tau)}{1 + \exp(-\bar{y}_i[j]s_i[j]/\tau)} \frac{\bar{y}_i[j]\mathcal{M}[j]}{\tau}. \tag{13}$$

With Eq. (13), we illustrate the gradient of $\mathcal{L}_{MCL-\tau}$ with different values of τ when $\bar{y}_i[j] = 1$ in Fig. 5. It is clear that, the updated MCL loss still suffers from substantial vanishing gradient issue as the classification score larger than 0.25 or smaller than -0.25. Another issue is that, our task involves a large number of classes, making the positive and negative classes unbalanced. Treating those negative classes equally in Eq. (11) may cause the issue of model collapse.

Memory-based Multi-label Classification Loss MMCL is proposed to ensure an efficient optimization to the ReID model and the enhanced robustness to noisy labels. The two issues in traditional MCL are addressed by computing the local loss on positive classes and hard negative classes. To

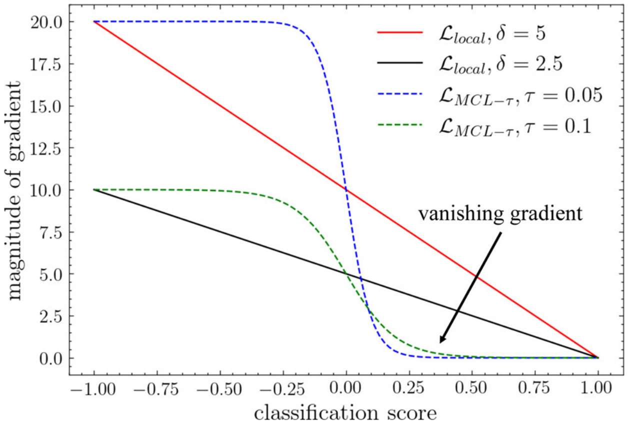


Fig. 5 Gradient Analysis for $\mathcal{L}_{MCL-\tau}$ and $\mathcal{L}_{local}$. It is clear that, $\mathcal{L}_{local}$ does not suffer from the vanishing gradient issue

further enhance the efficiency of optimization, MMCL considers a global loss computed on all classes. As illustrated in Fig. 6, the above intuitions lead to a local loss and a global loss, respectively. We hence could denote the MMCL as,

$$\mathcal{L}_{MMCL} = \lambda \mathcal{L}_{local} + (1 - \lambda)\mathcal{L}_{global}, \tag{14}$$

where the λ is weighting parameter to balance the two losses.

Local Loss $\mathcal{L}_{local}$ is proposed to address two issues in traditional MCL. For the first issue, since the score is bounded by $[-1, 1]$, we can abandon the sigmoid function and directly compute the loss by regressing the classification score to 1 and -1. This simplifies the loss computation and improves the training efficiency. The loss of classifying image x_i to class j can be updated as,

$$\ell^*(j|x_i) = ||s_i[j] - \bar{y}_i[j]||^2, \tag{15}$$

where $s_i[j] = \mathcal{M}[j]^\top \times f_i$ is the classification score.

The second issue is the imbalance between positive and negative classes. $\mathcal{L}_{local}$ introduces hard negative class mining to solve it. This is inspired by the sample mining in deep metric learning (Wu et al., 2017), where hard negative samples are more informative for training. Similarly in our multi-label classification, hard negative classes are more meaningful than easy negative ones.

For x_i, its negative classes can be denoted as $R_i \backslash P_i^*$. We rank them by their classification scores and select the top $r\%$ classes as the hard negative classes. The collection of hard negative classes for x_i can be denoted as N_i, $|N_i| = (n - |P_i^*|) \cdot r\%$.

The $\mathcal{L}_{local}$ is computed on positive classes and sampled hard negative classes as follows,

$$\mathcal{L}_{local} = \sum_{i=1}^{n} \frac{\delta}{|P_i^*|} \sum_{j \in P_i^*} \ell^*(j|x_i) + \frac{1}{|N_i|} \sum_{j \in N_i} \ell^*(j|x_i), \tag{16}$$

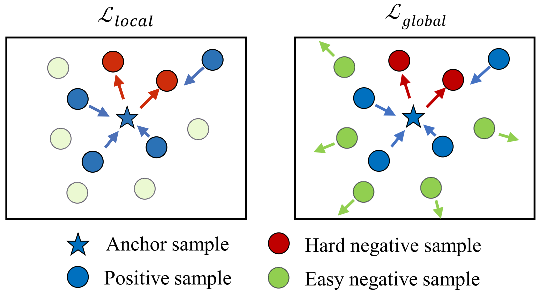


Fig. 6 Illustration of $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ for feature distance learning. $\mathcal{L}_{local}$ focuses on hard negative classes. $\mathcal{L}_{global}$ takes all negative classes into consideration

where δ is a coefficient measuring the importance of positive class loss and negative class loss, which will be tested in experiments.

We also illustrate the gradients of $\mathcal{L}_{local}$ when $\bar{y}_i[j] = 1$ in Fig. 5, where the gradient of $\mathcal{L}_{local}$ can be computed as,

$$\partial \mathcal{L}_{local} / \partial f_i = 2\delta \times (\mathcal{M}[j]^\top \times f_i - \bar{y}_i[j]) \times \mathcal{M}[j]. \tag{17}$$

Global Loss As $\mathcal{L}_{local}$ focuses on sampled classes and considers loss on each class separately, $\mathcal{L}_{global}$ is computed on all classes to improve the efficiency of CNN training. For an image x_i, given its classification score $s_i = \mathcal{M}^\top \times f_i$ and multi-class label $\bar{y}_i$, we first compute a probability vector v_i by applying the softmax operation on s_i,

$$v_i[j] = \frac{\exp(s_i[j]/\tau)}{\sum_{k=1}^{n} \exp(s_i[k]/\tau)}, \tag{18}$$

where τ is a temperature scalar, which is set to 0.05 according to (Zhong et al., 2019).

Traditional cross entropy loss enforces the probability on positive classes to 1. This conflicts with the training objective in multi-label classification, *i.e.*, the ideal probability v_i should exhibit uniform distribution over positive classes. Inspired by (Arazo et al., 2019; Tanaka et al., 2018), we aim to push the probability on each positive label to $1/|P_i^*|$, where $|P_i^*|$ is the number of positive classes of x_i. $\mathcal{L}_{global}$ is hence computed by modifying the cross entropy loss as,

$$\mathcal{L}_{global} = -\sum_{i=1}^{n} \frac{1}{|P_i^*|} \sum_{j \in P_i^*} \log(\frac{v_i[j]}{\max(1/|P_i^*|, v_i[j])}), \tag{19}$$

where $\max(1/|P_i^*|, v_i[j])$ supervises the model to learn the uniform distribution. For instance, if the classification probability on a positive class j is less than $1/|P_i^*|$, the loss will refine the model to output probability approaching to $1/|P_i^*|$. Benefited from the probability representation in Eq. (18), $\mathcal{L}_{global}$ can be efficiently computed since Eq. (19) is only computed on positive classes.

Algorithm 1 The pipeline of our framework.

```
Input: Unlabeled dataset $\mathcal{X} = \{x_1, x_2, ..., x_n\}$. ReID model $\Phi(\cdot)$. Memory bank $\mathcal{M}$. Total training epoch $num\_epochs$.
Output: Updated ReID model $\Phi(\cdot)$.

Initialize $\Phi(\cdot)$ with ImageNet-pretrained weights.
Initialize $\mathcal{M}$ to all zero;
for $epoch\_id$ in $1...num\_epochs$ do
    Compute distance matrix $D^{'}$ based on $\mathcal{M}$ use Eq. (8);
    Cluter on $D^{'}$ to obtain the reference number $n_i$ for each image $x_i$;
    for $x_i$ in $\mathcal{X}$ do
        Extract image feature by ReID model $f_i = \Phi(x_i)$;
        Initialize the single-class label $y_i$ by Eq. (2);
        /* CMLP computation */
        Compute image rank list $R_i$ based on $D^{'}$;
         Expand $n_i$ and select initial positive classes $P_i = R_i[1 : n_i^{'}]$;
         Use cycle consistency to filtr out hard negative classes to obtain $P_i^*$;
         Construct predicted multi-class label by Eq. (9);
        /* MMCL computation */
         Compute local loss on sampled classes by Eq. (16);
         Compute global loss on all classes by Eq. (19);
         Compute final loss $\mathcal{L}$ by Eq. (14);
         Update ReID model $\Phi(\cdot)$ based on loss $\mathcal{L}$ and SGD.
         Update Memory bank with feature $f_i$ by Eq. (6);
     end for
end for
return Final ReID model $\Phi(\cdot)$.
```

Discussions Comparison between the $\mathcal{L}_{MCL-\tau}$ and $\mathcal{L}_{local}$ in Fig. 5 clearly shows that, the vanishing gradient issue is effectively addressed by $\mathcal{L}_{local}$. Because of vanishing gradient, $\mathcal{L}_{MCL-\tau}$ won't enforce the classifier to output positive scores larger than 0.25. Note that, our framework uses feature similarity computed by Eq. (3) as classification scores. Low positive classification score indicates small feature similarity with positive classes, which is harmful for decreasing the intra-class variance. Compared with $\mathcal{L}_{MCL-\tau}$, $\mathcal{L}_{local}$ is more effective than $\mathcal{L}_{MCL-\tau}$ in optimizing the ReID model. Figure 5 also shows that, δ controls the magnitude of the gradient of $\mathcal{L}_{local}$. As discussed in (Zhang & Sabuncu, 2018), mean square loss is inferior to log-based loss (*e.g.* cross entropy) when classification score is near the decision boundary. This issue is addressed by introducing δ to scale the gradient magnitude of $\mathcal{L}_{local}$.

The global loss $\mathcal{L}_{global}$ improves the local loss by considering all labels into loss computation. As shown in Fig. 6, $\mathcal{L}_{local}$ only focuses on sampled labels, thus is not effective in optimizing the feature distance. $\mathcal{L}_{global}$ involves all labels in to computation and can effectively improve the performance of learned ReID feature. We will test the effects of $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ in Sect. 5. Algorithm 1 summarizes the whole pipeline of the proposed method.

5 Experiment

5.1 Datasets and Evaluation Metrics

We evaluate the proposed approach on two large scale public person ReID datasets: *Market-1501* (Zheng et al., 2015) and *DukeMTMC-reID* (Ristani et al., 2016).

Market-1501 Zheng et al. (2015) is made up of 32,368 pedestrian images taken by six manually configured cameras. It contains 1,501 person IDs. On average, there are 3.6 images for each person under each camera. Because Market1501 has provided the training set and testing set, we use images in the training set for unsupervised learning and report the ReID performance on the testing set.

DukeMTMC-reID Ristani et al. (2016) is composed of 1,812 pedestrians and 36,411 bounding boxes, with 1,404 pedestrians appear in more than one camera. This dataset is captured by 8 cameras. 16,522 bounding boxes of 702 pedestrians are used for training. The other identities are included in the testing set.

Evaluation Metrics. We follow the standard settings in (Zheng et al., 2015; Ristani et al., 2016) to conduct experiments. Performance is evaluated by the Cumulative Matching Characteristic (CMC) and mAP.

5.2 Implementation Details

All experiments are implemented using PyTorch framework. We use ResNet-50 (He et al., 2016) as backbone to extract the feature and initialize it with parameters pre-trained on ImageNet (Deng et al., 2009). After pooling-5 layer, we remove subsequent layers and add a batch normalization layer (Ioffe & Szegedy, 2015), which produces a 2048-dim feature. During testing, we also extract the pooling-5 feature to calculate the distance. For multi-label classification, we allocate a memory bank to store L2 normalized image features. The memory bank is initialized to all zeros. As mentioned in Sect. 3, we leverage CamStyle (Zhong et al., 2018b) as a data augmentation strategy for unlabeled images. Strategies like random crop, random rotation, color jitter, and random erasing are also used to improve the feature robustness. To further ensure the quality of predicted multi-class labels, we incorporate temporal information (Li & Zhang, 2020) to refine the distance in Eq. 8.

The input image is resized to 256×128. We use SGD to optimize the model, the learning rates for ResNet-50 base layers are 0.01, and others are 0.1. The memory updating rate α starts from 0 and grows linearly to 0.6. We train the model for 40 epochs, and the learning rate is divided by 10 after every 30 epochs. The batch size for model training is 64. We do clustering and predict multi-class label every epoch. For multi-label classification loss, we set the weight δ to 5, hard negative mining ratio $r\%$ to 0.1% and loss weight $\lambda = 0.3$

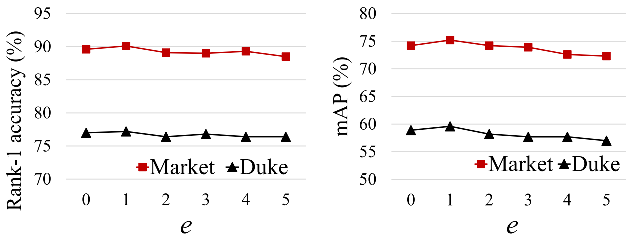


Fig. 7 Evaluation of e for positive class selection

Table 2 Evaluation of parameter δ in Eq. (16)

δ	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
1	88.1	71.8	75.7	57.3
2	89.7	74.2	77.0	59.4
3	89.0	74.1	76.3	59.0
4	89.9	74.7	77.2	58.9
5	**90.1**	**75.2**	**77.2**	**59.6**
6	89.5	75.1	76.1	58.5

Bold indicates the best performance in the table

for balancing local and global loss in MMCL through the parameter analysis in Sect. 5.3.

5.3 Parameter Analysis

This section investigates hyper-parameters introduced by our method, including the e for positive class selection, coefficient δ, hard negative mining ratio $r\%$ and weighting parameter λ. Each experiment varies one hyper-parameter while keeping others fixed. All experiments are conducted with unsupervised ReID setting on both *Market-1501* and *DukeMTMC-reID*.

***e* for positive class selection** Larger e selects more candidate positive classes, thus boosts the recall rate. Figure 7 reports the performance of different e in CMLP. We can observe that increasing e improves the performance. For example, $e = 1$ performs better than $e = 0$ in terms of both mAP and rank-1 accuracy. However, too large e is harmful for model learning. According to Fig. 7, we set $e = 1$, which corresponds to good performance on both datasets.

Coefficient δ Table 2 reports the analysis on coefficient δ of MMCL. As discussed in Sect. 4.2, δ plays a role to scale the gradient of $\mathcal{L}_{local}$. $\delta = 1$ does not scale the gradient. In this case, the $\mathcal{L}_{local}$ cannot produce large gradients to pull positive samples together, leading to bad performance. For example, the rank-1 accuracy is dropped to 88.1% on *Market-1501* and 75.7% on *DukeMTMC-reID*. As δ becomes larger, $\mathcal{L}_{local}$ effectively improves the similarity of positive samples, leading to better performance. However, too large δ may make the training unstable. According to Table 2, we set $\delta = 5$, which performs best on different datasets.

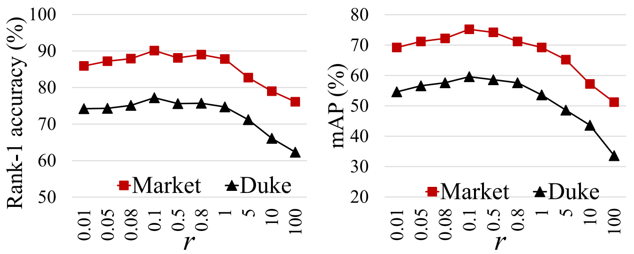


Fig. 8 Evaluation of hard negative mining ratio r

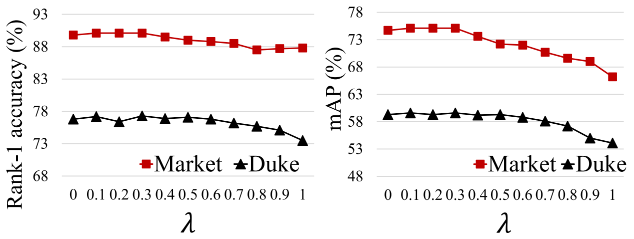


Fig. 9 Evaluation of weighting parameter λ in Eq. (14)

Hard negative mining ratio $r\%$ Fig. 8 shows effects of hard negative mining ratio $r\%$ in $\mathcal{L}_{local}$. r=100 means using all negative classes for loss computation, which is harmful for both accuracy and efficiency. This implies that, not all of the negative classes are helpful for unsupervised ReID training, if they are treated independently with equal weight. As r becomes smaller, hard negative mining would be activated and it boosts the accuracy. Too small r selects too few negative classes, hence is also harmful for the performance. We finally set $r = 0.1$ a reasonable tradeoff between accuracy and efficiency.

Loss weighting parameter λ Fig. 9 investigates the effect of λ, which weights the local and global loss in Eq. (14). We vary λ from 0.0 to 1.0. By setting λ to 1.0, only local loss is taken into consideration. As discussed in Sect. 4.2, global loss could effectively enlarge inter-class variation and improve the ReID performance. Those figures clearly show that, setting smaller λ boosts the performance. Proper weights hence should be set to balance the loss on local loss and global loss. According to this experiments, we set $\lambda = 0.3$, a reasonable tradeoff between local and global losses.

The above experiments on two datasets also indicate that, our parameters are easy to tune, *i.e.*, selected parameters on one dataset show consistent performance on different datasets. We hence set $e = 1$, $\delta = 5$, $r = 0.1$, and $\lambda = 0.3$ for following experiments.

5.4 Ablation Study

Baselines To conduct ablation studies, we first implement two baselines in Table 3 for comparison. The first baseline

Table 3 Test of validity of proposed CMLP and MMCL.

Method		Market-1501				DukeMTMC-reID			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Supervised		87.1	94.8	96.9	68.3	75.9	86.9	90.5	57.9
Pre-trained		8.1	17.5	23.6	2.2	5.6	11.5	14.9	1.6
CMLP	Single-class Label	32.2	52.8	62.1	10.5	29.0	44.7	51.7	10.6
	Clustering	88.6	95.2	96.8	72.2	77.0	87.5	90.1	57.6
	CMLP	**90.1**	**95.6**	**97.2**	**75.2**	**77.2**	**87.6**	**90.6**	**59.6**
MMCL	$\mathcal{L}_{local}$	87.6	92.9	95.6	67.1	73.4	85.5	87.9	53.3
	$\mathcal{L}_{global}$	89.7	95.4	96.9	74.5	76.7	86.8	90.5	58.8
	$\mathcal{L}_{local}\&\mathcal{L}_{global}$	**90.1**	**95.6**	**97.2**	**75.2**	**77.2**	**87.6**	**90.6**	**59.6**

"Supervised" denotes baseline trained with annotated labels and Cross Entropy loss. "Pre-trained" denotes baseline pre-trained on ImageNet
Bold indicates the best performance in the table

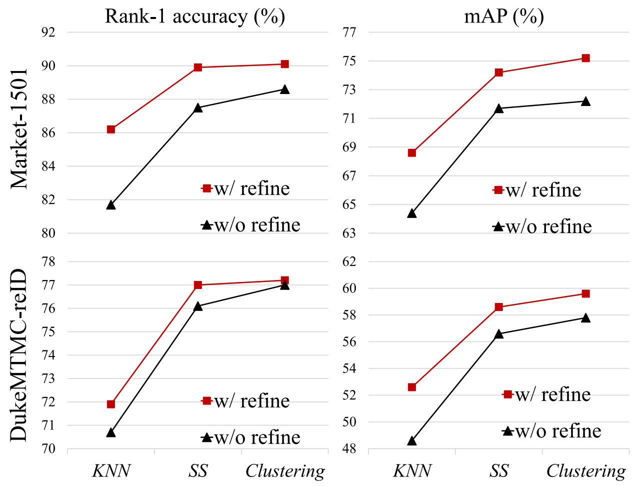


Fig. 10 Evaluation of proposed positive label expansion and refinement strategy in CMLP for different multi-class label prediction strategies

trains the ReID model in supervised learning, we denote it as "Superivsed". This supervised learning achieves high accuracy on two datasets, *e.g.*, 87.1% and 75.9% rank-1 accuracy on *Market-1501* and *DukeMTMC-reID*, respectively. The second baseline is denoted as "Pre-trained", which directly uses parameters pre-trained on ImageNet for testing. As shown in Table 3, the bad performance of Pre-trained is mainly caused by the domain gap between ImageNet and ReID datasets. Those two baselines can be viewed as the upper bound and lower bound of unsupervised person ReID.

Validity of CMLP We first analyze the effectiveness of our label prediction algorithm. As shown in Table 3, we first show the result achieved without any label prediction, which is denote as "Single-class Label". It improves the lower bound baseline on both datasets. Specifically, the rank-1 accuracy improves from 8.1 to 32.2% and 5.6% to 29.0% when tested on *Market-1501* and *DukeMTMC-reID*, respectively. This demonstrates that treating each image as a single class is a feasible way to improve the baseline performance. Further applying label prediction substantially boosts the performance. For example, when adopting CMLP w/o refine, we obtain 88.6% rank-1 accuracy and 72.2% mAP on *Market-1501*, which outperform Single-class Label by 56.4% and 61.7%, respectively. This verifies our motivation that multi-class label is better than single-class label in unsupervised ReID. Further incorporating the positive label expansion and label refinement strategies, CMLP boosts the ReID performance to 90.1% rank-1 accuracy and 75.2% mAP on *Market-1501*. This shows the effectiveness of each component in CMLP.

Our positive label expansion and refinement strategy proposed in Sect. 4.1 is effective as shown in above experiments. Figure 10 further tests it with other label prediction methods, *e.g.*, *KNN* (Zhong et al., 2019), *SS* (Fan et al., 2018D), and *Clustering* (Fu et al., 2019; Zhang et al., 2019; Ge et al., 2020a), respectively. From Fig. 10, we can observe that this strategy brings consistent performance gains for different label prediction algorithms. For example, it brings about 4% gains in mAP for *KNN* on *Market-1501*. A similar result can be observed on *DukeMTMC-reID*. We hence could conclude that, our method is simple and works well.

Validity of MMCL This part evaluates the proposed MMCL by first testing $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ separately. As reported in Table 3, both $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ achieve good performance on two dataset, *e.g.*, on *Market-1501*, $\mathcal{L}_{local}$ obtains 67.1% mAP while $\mathcal{L}_{global}$ obtains 74.5%. Combining those two losses leads to substantial performance gains, *i.e.*, 75.2% mAP on *Market-1501*. A similar results can be observed on *DukeMTMC-reID*, *i.e.*, combining two losses achieves 59.6% mAP. This implies that, those two losses are complementary to each other. Trained with CMLP and MMCL, our ReID model even gets better performance than the "Supervised" model, *i.e.*, the performance upper bound of Cross Entropy (CE) loss. This is mainly because MMCL is superior to CE loss, as shown in Table 4. It indicatesing

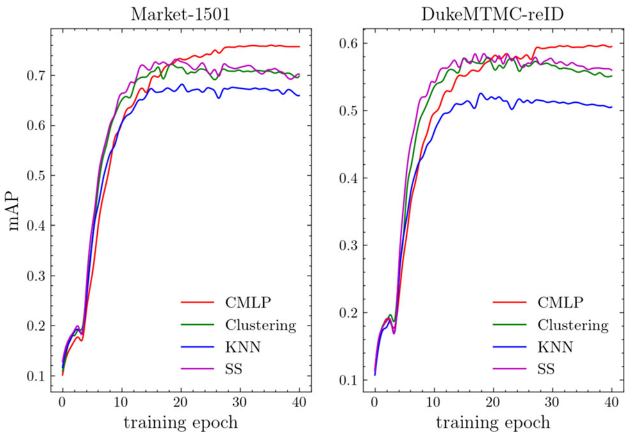


Fig. 11 The mAP curves achieved by different label prediction algorithms with MMCL on two datasets

Table 4 Test of MMCL and Cross Entropy (CE) loss with different labels on *Market-1501* and *DukeMTMC-reID*, respectively

Loss	Label Type	Market-1501		DukeMTMC-reID	
		Rank-1	mAP	Rank-1	mAP
CE	Single-class	31.4	9.9	31.5	11.8
CE	Ground Truth	87.1	68.3	75.9	57.9
MMCL	Single-class	32.2	10.5	29.0	10.6
MMCL	Ground Truth	89.9	76.2	79.0	59.8

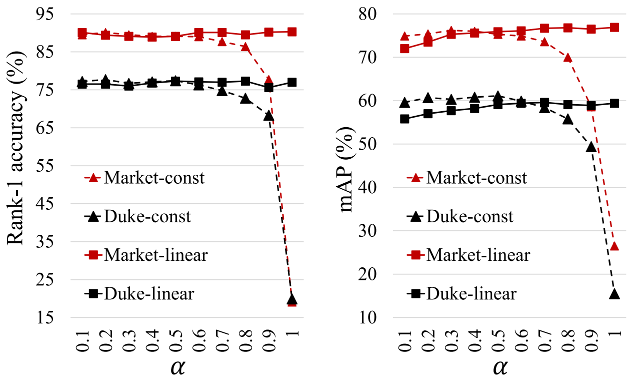


Fig. 12 Comparison of different memory updating methods subject to varied maximum updating rate *α* in Eq. (6) on *Market-1501* and *DukeMTMC-reID*, respectively

that, the proposed multi-label classification framework iss an effective method in to learning discriminative person features in under the unsupervised setting.

Figure 11 further tests MMCL with different label prediction algorithms. Compared with mAP curves in Fig. 2, which are computed with traditional multi-label classification loss, the ones in Fig. 11 exhibit better stability and performance. Combined with the CMLP, the mAP degradation issue in Fig. 11 is effectively addressed. This demonstrates our MMCL is more robust to noisy labels. It is also clear that, our CMLP gets the best performance compared with other label prediction algorithms.

To further show the potential of MMCL, we compare it with Cross Entropy (CE) loss with different labels including single-class label and ground truth label. Experimental results are summarized in Table 4. Trained with single-class labels, MMCL gets 32.2% rank-1 accuracy and 10.5% mAP on Market-1501, slightly better than those of CE. We further test them using ground truth labels. MMCL still performs better than CE on both datasets. This indicates that MMCL is superior to CE. We can also observe that the proposed CMLP+MMCL achieves similar performance to its upper bound using ground truth labels. CMLP+MMCL obtains 75.2% and 59.6% mAP on *Market-1501* and *DukeMTMC-reID*, which is only lower than its upper bound by 1.0% and 0.2%. It is also interesting to observe is that, MMCL+CMLP performs better than MMCL + Ground Truth in Rank-1 on *Market-1501*, *e.g.*, 90.1 v.s. 89.9. This can be partially explained by the label noises in ground truth annotation, which can be relieved to some extent by unsupervised learning.

Ablation study on different memory bank updating strategies we further evaluate different updating methods of the memory bank, as discussed in Sect.3. We test two updating methods. One increases the memory updating rate linearly during training. The other uses constant updating rate. For both methods, we vary the maximum updating rate from 0.1 to 1.0 and summarize the results in Fig. 12. Our results show that, linear increasing is better than the other. This indicates that an adaptive updating rate is a more reasonable choice for our task. We also observe that, linear increasing is not sensitive to the maximum updating rate, *e.g.*, too large updating rate significantly degrades its competitor, but does not degrade the performance of linear increasing. This property makes our method easier to tune.

5.5 Comparison with the State of the Art

We compare our method against state-of-the-art unsupervised learning and transfer learning approaches on *Market-1501* (Zheng et al., 2015) and *DukeMTMC-reID* (Ristani et al., 2016). Table 5 summarize the comparison.

We compare two types of methods, including unsupervised learning methods: LOMO (Liao et al., 2015), BOW (Zheng et al., 2015), BUC (Lin et al., 2019), DBC (Ding et al., 2019), MMCL (Wang & Zhang, 2020), SSL (Lin et al., 2020), HCT (Zeng et al., 2020) and SpCL (Ge et al., 2020b), *etc*, and transfer learning based approaches: PUL (Fan et al., 2018D), PTGAN (Wei et al., 2018), SPGAN (Deng et al., 2018), CAMEL (Yu et al., 2017), MMFA (Lin et al., 2018), TJ-AIDL (Wang et al., 2018), HHL (Zhong et al., 2018a), ECN (Zhong et al., 2019), MAR (Yu et

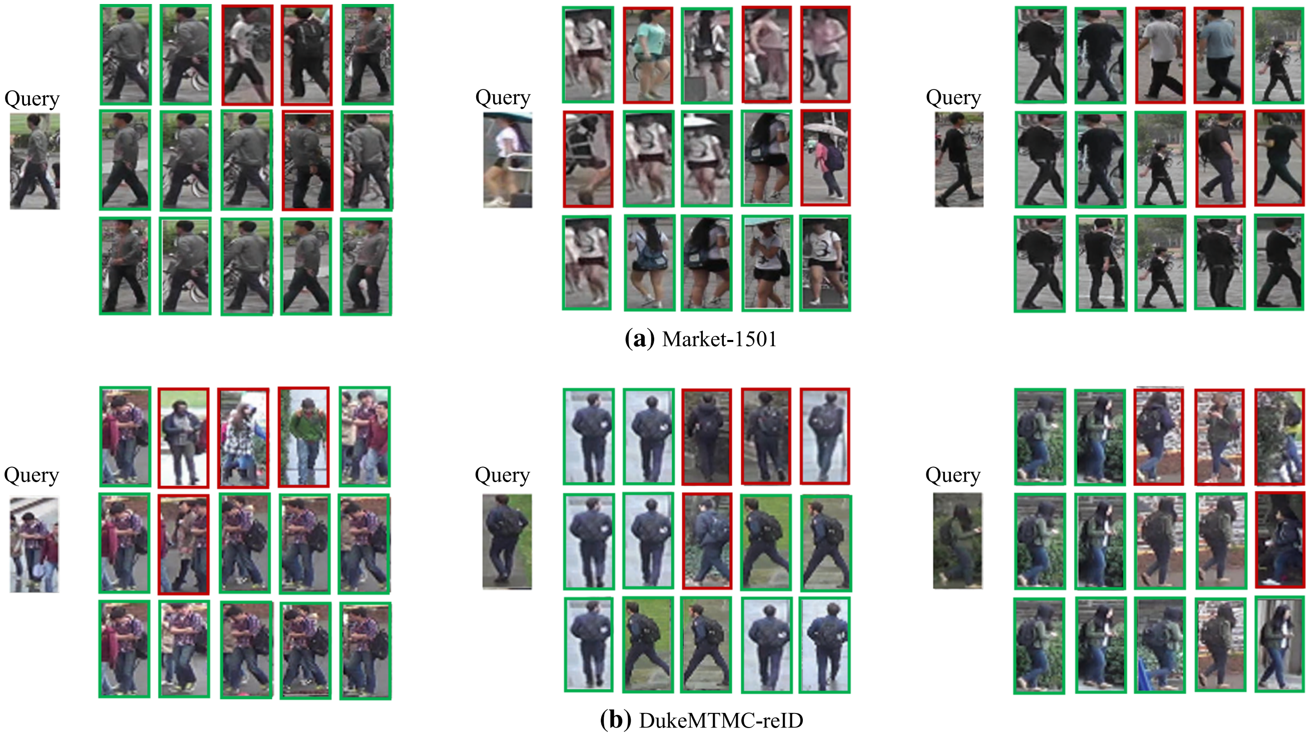
Table 5 Unsupervised person ReID performance comparison with state-of-the-art methods on *Market-1501* and *DukeMTMC-reID*

Method	Reference	Market-1501					DukeMTMC-reID				
		Source	R-1	R-5	R-10	mAP	Source	R-1	R-5	R-10	mAP
PUL(Fan et al., 2018D)	TOMM18	Duke	45.5	60.7	66.7	20.5	Market	30.0	43.4	48.5	16.4
PTGAN(Wei et al., 2018)	CVPR18	Duke	38.6	-	66.1	-	Market	27.4	-	50.7	-
SPGAN(Deng et al., 2018)	CVPR18	Duke	51.5	70.1	76.8	22.8	Market	41.1	56.6	63.0	22.3
CAMEL(Yu et al., 2017)	ICCV17	Multi	54.5	-	-	26.3	-	-	-	-	-
MMFA(Lin et al., 2018)	BMVC19	Duke	56.7	75.0	81.8	27.4	Market	45.3	59.8	66.3	24.7
TJ-AIDL(Wang et al., 2018)	CVPR18	Duke	58.2	74.8	81.1	26.5	Market	44.3	59.6	65.0	23.0
HHL(Zhong et al., 2018a)	ECCV18	Duke	62.2	78.8	84.0	31.4	Market	46.9	61.0	66.7	27.2
DECAMEL(Yu et al., 2018)	TPAMI19	Multi	60.2	76.0	-	32.4	-	-	-	-	-
ECN(Zhong et al., 2019)	CVPR19	Duke	75.1	87.6	91.6	43.0	Market	63.3	75.8	80.4	40.4
MAR(Yu et al., 2019)	CVPR19	MSMT	67.7	81.9	-	40.0	MSMT	67.1	79.8	-	48.0
PAUL(Yang et al., 2019)	CVPR19	MSMT	68.5	82.4	87.4	40.1	MSMT	72.0	82.7	86.0	53.2
SSG(Fu et al., 2019)	ICCV19	Duke	80.0	90.0	92.4	58.3	Market	73.0	80.6	83.2	53.4
CR-GAN(Chen et al., 2019)	ICCV19	Duke	77.7	89.7	92.7	54.0	Market	68.9	80.2	84.7	48.6
CASCL(Wu et al., 2019)	ICCV19	MSMT	65.4	80.6	86.2	35.5	MSMT	59.3	73.2	77.5	37.8
PDA-Net(Li et al., 2019)	ICCV19	Duke	75.2	86.3	90.2	47.6	Market	63.2	77.0	82.5	45.1
UCDA(Qi et al., 2019)	ICCV19	Duke	64.3	-	-	34.5	Market	55.4	-	-	36.7
PAST(Zhang et al., 2019)	ICCV19	Duke	78.38	-	-	54.62	Market	72.35	-	-	54.26
MMCL(Wang & Zhang, 2020)	CVPR20	Duke	84.4	92.8	95.0	60.4	Market	72.4	82.9	85.0	51.4
pMR-SADA(Wang et al., 2020)	CVPR20	Duke	83.0	91.8	94.1	59.8	Market	74.5	85.3	88.7	55.8
AD-Cluster(Zhai et al., 2020)	CVPR20	Duke	86.7	94.4	96.5	68.4	Market	72.6	82.5	85.5	54.1
ECN++(Zhong et al., 2020)	TPAMI20	Duke	84.1	92.8	95.4	63.8	Market	74.0	83.7	87.4	54.4
GCL(Chen et al., 2021)	CVPR21	Duke	89.1	95.0	96.6	73.4	Market	77.2	86.2	88.4	60.4
LOMO(Liao et al., 2015)	CVPR15	None	27.2	41.6	49.1	8.0	None	12.3	21.3	26.6	4.8
BOW(Zheng et al., 2015)	ICCV15	None	35.8	52.4	60.3	14.8	None	17.1	28.8	34.9	8.3
BUC(Lin et al., 2019)	AAAI19	None	66.2	79.6	84.5	38.3	None	47.4	62.6	68.4	27.5
DBC(Ding et al., 2019)	BMVC19	None	69.2	83.0	87.8	41.3	None	51.5	64.6	70.1	30.0
SSL(Lin et al., 2020)	CVPR20	None	71.7	83.8	87.4	37.8	None	52.5	63.5	68.9	28.6
MMCL(Wang & Zhang, 2020)	CVPR20	None	80.3	89.4	92.3	45.5	None	65.2	75.9	80.0	40.2
HCT(Zeng et al., 2020)	CVPR20	None	80.0	91.6	95.2	56.4	None	69.6	83.4	87.4	50.7
JVTC+(Li & Zhang, 2020)	ECCV20	None	79.5	89.2	91.9	47.5	None	74.6	82.9	85.3	50.7
SpCL(Ge et al., 2020b)	NeurIPS20	None	88.1	95.1	97.0	73.1	-	-	-	-	-
GCL(Chen et al., 2021)	CVPR21	None	83.7	91.6	94.3	63.4	None	72.4	82.0	84.9	53.3
MetaCam(Yang et al., 2021)	CVPR21	None	83.9	92.3	-	61.7	None	73.8	84.2	-	53.8
Ours	-	None	**90.1**	**95.6**	**97.2**	**75.2**	None	**77.2**	**87.6**	**90.6**	**59.6**

Bold indicates the best performance in the table

al., 2019), PAUL (Yang et al., 2019), SSG (Fu et al., 2019), CR-GAN (Chen et al., 2019), CASCL (Wu et al., 2019), PDA-Net (Li et al., 2019), UCDA (Qi et al., 2019), PAST (Zhang et al., 2019), MMCL (Wang & Zhang, 2020) and ECN++ (Zhong et al., 2020), *etc*.

We first compare with unsupervised learning methods. LOMO (Liao et al., 2015) and BOW (Zheng et al., 2015) utilize hand-crafted features, and show the worst performance. BUC (Lin et al., 2019) and DBC (Ding et al., 2019) treat each image as a single cluster then merge clusters, thus share certain similarity with our work. However, our method outperforms them by large margins. The reasons could be because: 1) their bottom-up clustering strategy accumulates the quantization error in partitioning the feature space, and 2) BUC tries to keep different clusters with similar size, hence suffers from the issue of imbalanced number of positive classes. HCT (Zeng et al., 2020) addresses the first issue by re-clustering at each epoch, thus gets better performance. However, as its authors state, HCT is prone to overfitting, and needs the early stopping to avoid performance degra-



(a) Market-1501

(b) DukeMTMC-reID

Fig. 13 Illustration of unsupervised person ReID results on *Market-1501* and *DukeMTMC-reID*. Each example shows top-5 retrieved images by ReID model trained with Single-class Label, our conference version (Wang & Zhang, 2020), and the method in this paper, respectively. True positive is annotated by the green bounding box. False positive is annotated by the red bounding box (Color figure online)

dation. Benefited from the CMLP and MMCL, our method obtains 90.1% rank-1 accuracy and 75.2% mAP on *Market-1501*, as well as 77.2% rank-1 accuracy and 59.6% mAP on *DukeMTMC-reID*, which outperform all of those competitors in Table 5. For instance, our method outperforms the recent SpCL (Ge et al., 2020b) by 2.0% in rank-1 accuracy and 2.1% in mAP on *Market-1501*, respectively. Our method also performs substantially better than our conference version (Wang & Zhang, 2020), *e.g.*, outperforms the conference version by 9.8% in rank-1 accuracy on *Market-1501*. Sample ReID results by our method are illustrated in Fig. 13.

6 Conclusion

This paper proposes a multi-label classification method to address unsupervised person ReID. Different from previous works, our method works without requiring any labeled data or a good pre-trained model. Good performance is achieved by iteratively predicting multi-class labels and updating the network with a multi-label classification loss. CMLP is proposed for multi-class label prediction by considering both clustering and cycle consistency. MMCL is introduced to compute the multi-label classification loss and it addresses the issues of vanishing gradient, as well as the unbalanced positive-negative class ratio. Experiments on several large-scale datasets demonstrate the promising performance of the proposed methods in unsupervised person ReID.

Acknowledgements This work is supported in part by Natural Science Foundation of China under Grant Nos. U20B2052, 61936011, in part by The National Key Research and Development Program of China under Grant Nos. 2018YFE0118400.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N., & Mcguinness, K. (2019). Unsupervised label noise modeling and loss correction. In: *Proceedings of Machine Learning Research*, 312–321.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, MS., Maharaj, T., Fischer, A., Courville, A., & Bengio, Y., et al. (2017) A closer look at memorization in deep networks. In:*Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, (pp. 233–242).

Chen, H., Wang, Y., Lagadec, B., Dantcheva, A., & Bremond, F., (2021) Joint generative and contrastive learning for unsupervised person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2004–2013)

Chen, Y., Zhu, X., & Gong, S. (2019) Instance-guided context rendering for cross-domain person re-identification. In: *ICCV*

Danelljan, M., Gool, LV., & Timofte, R. (2020) Probabilistic regression for visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 7183–7192).

Deng, J., Dong, W., Socher, R., Li, L., Kai, Li., & Li, Fei-Fei. (2009) Imagenet: A large-scale hierarchical image database. In: *CVPR*.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: *CVPR*.

Ding, G., Khan, S., Yin, Q., & Tang, Z. (2019) Dispersion based clustering for unsupervised person re-identification. In: *BMVC*.

Durand, T., Mehrasa, N., & Mori, G. (2019) Learning a deep convnet for multi-label classification with partial labels. In: *CVPR*.

Ester, M., Kriegel, H. P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, 96,* 226–231.

Fan, H., Zheng, L., Yan, C., & Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14*(4), 83.

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., & Huang, TS. (2019) Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: *ICCV*.

Ge, Y., Chen, D., & Li, H. (2020a) Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: *International Conference on Learning Representation*.

Ge, Y., Zhu, F., Chen, D., Zhao, R., & Li, h. (2020b) Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: H. Larochelle, M. Ranzato, R. Hadsell, MF. Balcan, H. Lin (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 33, (pp. 11309–11321).

Ghosh, A., Kumar, H., & Sastry, P. (2017) Robust loss functions under label noise for deep neural networks. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems,* (pp. 2672–2680).

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: *Advances in Neural Information Processing Systems,* (pp. 8527–8537)

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2019) Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In: *CVPR*.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527–1554.

Ioffe, S., & Szegedy, C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *ICML*.

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2018) Mining on manifolds: Metric learning without labels. In: *CVPR*.

Jegou, H., Harzallah, H., & Schmid, C. (2007) A contextual dissimilarity measure for accurate and efficient image search. In: *CVPR*.

Kingma, DP., & Welling, M. (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114

Kodirov, E., Xiang, T., & Gong, S. (2015) Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In: *Proceedings of the British Machine Vision Conference 2015,* (pp. 44.1–44.12).

Komodakis, N., & Gidaris, S. (2018) Unsupervised representation learning by predicting image rotations. In: *ICLR*.

Krizhevsky, A., Sutskever, I., & Hinton, GE. (2012) Imagenet classification with deep convolutional neural networks. In: *NeurIPS*.

Le, QV. (2013) Building high-level features using large scale unsupervised learning. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE,* (pp. 8595–8598).

Li, YJ., Lin, CS., Lin, YB., & Wang, YCF. (2019) Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: *ICCV*.

Li, J., & Zhang, S. (2020) Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In: *European Conference on Computer Vision,* (pp. 483–499). Springer.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. (2018) Unsupervised learning of view-invariant action representations. In: *NeurIPS*.

Li, H., Wu, Z., Zhu, C., Xiong, C., Socher, R., & Davis, LS. (2020) Learning from noisy anchors for one-stage object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (pp. 10588–10597).

Liao, S., Hu, Y., Zhu, X., & Li, SZ. (2015) Person re-identification by local maximal occurrence representation and metric learning. In: *CVPR*.

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. (2019) A bottom-up clustering approach to unsupervised person re-identification. In: *AAAI*.

Lin, S., Li, H., Li, CT., & Kot, AC. (2018) Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In: *BMVC*.

Lin, Y., Xie, L., Wu, Y., Yan, C., & Tian, Q. (2020) Unsupervised person re-identification via softened similarity learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Long, M., Cao, Y., Wang, J., & Jordan, MI. (2015) Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791

Lv, J., Chen, W., Li, Q., & Yang, C. (2018) Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: *CVPR*.

Ma, F., Meng, D., Xie, Q., Li, Z., & Dong, X. (2017) Self-paced co-training. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70,* JMLR. org, (pp. 2275–2284).

Neverova, N., Novotny, D.,& Vedaldi, A. (2019) Correlated uncertainty for learning dense correspondences from noisy labels. *Advances in Neural Information Processing Systems*, 32

Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., & Gao, Y. (2019) A novel unsupervised camera-aware domain adaptation framework for person re-identification. In: *ICCV*.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016) Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 2818–2826).

Tanaka, D., Ikami, D., Yamasaki, T., & Aizawa, K. (2018) Joint optimization framework for learning with noisy labels. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 5552–5560).

Tang, Y., Salakhutdinov, R., & Hinton, G. (2012) Robust boltzmann machines for recognition and denoising. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 2264–2271). IEEE.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, PA. (2008) Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning,* (pp. 1096–1103).

Wang, D., & Zhang, S. (2020) Unsupervised person re-identification via multi-label classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (pp. 10981–10990).

Wang, H., Gong, S., & Xiang, T. (2014) Unsupervised learning of generative topic saliency for person re-identification. In: *Proceedings of the British Machine Vision Conference 2015*.

Wang, G., Lai, JH., Liang, W., & Wang, G. (2020) Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018) Person transfer gan to bridge domain gap for person re-identification. In: *CVPR*.

Wu, CY., Manmatha, R., Smola, AJ., & Krahenbuhl, P. (2017) Sampling matters in deep embedding learning. In: *CVPR*.

Wu, Q., Wan, J., Chan, AB. (2021) Progressive unsupervised learning for visual object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 2993–3002).

Wu, Z., Xiong, Y., Yu, SX., & Lin, D. (2018) Unsupervised feature learning via non-parametric instance discrimination. In: *CVPR*.

Wu, A., Zheng, WS., & Lai, JH. (2019) Unsupervised person re-identification by camera-aware similarity consistency learning. In: *ICCV*.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., & Zuo, W. (2017) Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: *CVPR*.

Yang, Q., Yu, HX., Wu, A., & Zheng, WS. (2019) Patch-based discriminative feature learning for unsupervised person re-identification. In: *CVPR*.

Yang, F., Zhong, Z., Luo, Z., Cai, Y., Lin, Y., Li, S., & Sebe, N. (2021) Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 4855–4864).

Ye, M., Zhang, X., Yuen, PC., & Chang, SF. (2019) Unsupervised embedding learning via invariant and spreading instance feature. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 6210–6219)

Yu, HX., Wu, A., & Zheng, WS. (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In: *CVPR*.

Yu, HX., Wu, A., & Zheng, WS. (2018) Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, HX., Zheng, WS., Wu, A., Guo, X., Gong, S., & Lai, JH. (2019) Unsupervised person re-identification by soft multilabel learning. In: *CVPR*.

Zeng, K., Ning, M., Wang, Y., & Guo, Y. (2020) Hierarchical clustering with hard-batch triplet loss for person re-identification. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., & Tian, Y. (2020) Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z., & Sabuncu, M. (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: *Advances in Neural Information Processing Systems*, (pp. 8778–8788).

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016) Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530

Zhang, X., Cao, J., Shen, C., & You, M. (2019) Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: *ICCV*.

Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1819–1837.

Zhao, R., Ouyang, W., & Wang, X. (2013) Unsupervised salience learning for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3586–3593).

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

Zheng, L., Yang, Y., & Hauptmann, AG. (2016) Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984

Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017) Re-ranking person re-identification with k-reciprocal encoding. In: *CVPR*.

Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018a) Generalizing a person retrieval model hetero-and homogeneously. In: *ECCV*.

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2019) Invariance matters: Exemplar memory for domain adaptive person re-identification. In: *CVPR*.

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2020) Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. (2018b) Camera style adaptation for person re-identification. In: *CVPR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.