Visual Intelligence

RESEARCH Open Access

CCSD: cross-camera self-distillation for unsupervised person re-identification

Jiyuan Chen[1], Changxin Gao[1*], Li Sun[2] and Nong Sang[1]

Abstract

Existing unsupervised person re-identification (Re-ID) methods have achieved remarkable performance by adopting an alternate clustering-training manner. However, they still suffer from camera variation, which results in an inconsistent feature space and unreliable pseudo labels that severely degrade the performance. In this paper, we propose a cross-camera self-distillation (CCSD) framework for unsupervised person Re-ID to alleviate the effect of camera variation. Specifically, in the clustering phase, we propose a camera-aware cluster refinement mechanism, which first splits each cluster into multiple clusters according to the camera views, and then refines them into more compact clusters. In the training phase, we first obtain the similarity between the samples and the refined clusters from the same and different cameras, and then transfer the knowledge of similarity distribution from intra-camera to cross-camera. Since the intra-camera similarity is free from camera variation, our knowledge distillation approach is able to learn a more consistent feature space across cameras. Extensive experiments demonstrate the superiority of our proposed CCSD against the state-of-the-art approaches on unsupervised person Re-ID.

Keywords: Unsupervised learning, Person re-identification, Knowledge distillation, Deep learning

1 Introduction

Person re-identification (Re-ID) aims to identify the same person in a non-overlapping camera system. Thanks to the rapid development of fully supervised Re-ID methods [1], the performance on existing datasets has improved impressively. However, they are heavily dependent on laborious and expensive data annotation, which limit their application in the real world. Therefore, researchers recently begin to focus on the unsupervised person Re-ID study.

Existing state-of-the-art methods for unsupervised Re-ID mainly apply an alternative two-phase pipeline [2–4], i.e., a clustering phase and a training phase. In the clustering phase, pseudo labels for each sample are first generated by clustering methods. In the training phase, the pseudo labels are used to train the Re-ID model in a supervised manner. Clustering and training are performed alternately until the model converges. Although these two-phase unsupervised Re-ID methods have achieved impressive accuracy, they still suffer from camera variation, which causes inconsistency in the feature space. The inconsistent feature space across cameras leads to unreliable pseudo labels which further mislead the model to learn a worse feature embedding space. This error amplification will severely reduce the performance of unsupervised Re-ID.

Due to the inconsistent feature space caused by camera variation, the feature distributions under different cameras have a large discrepancy, as shown in Fig. 1(a). This makes unsupervised person Re-ID over all the cameras a difficult task. In contrast, within a specific camera, it becomes a much easier task, because it is free from camera variation. To this end, we split the person Re-ID task into several camera-specific sub-tasks and then align them across cameras, called the split-and-alignment strategy. In this way, cross-camera alignment is the key issue. We observe that the similarity between the samples and the clusters across different cameras is less reliable than that within the same camera, mainly because of camera variation, as

*Correspondence: cgao@hust.edu.cn
[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
Full list of author information is available at the end of the article

Springer

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

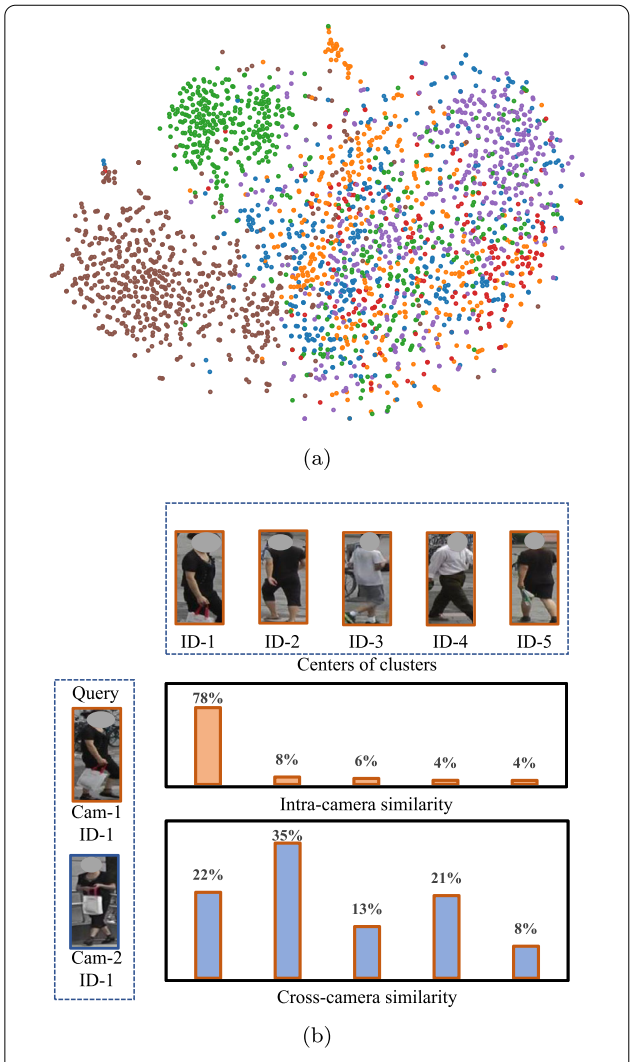


Figure 1 Illustration of camera variations in person Re-ID. (a) shows the t-SNE visualization [5] of feature distribution from Market-1501. The feature distributions under different cameras, indicated by different colors, have a large discrepancy. (b) shows the intra-camera and cross-camera similarities, where samples and clusters from the same and different cameras, respectively. Due to the inconsistent feature space, intra-camera similarity is more reliable than cross-camera one

illustrated in Fig. 1(b). Based on this observation, we introduce self-distillation to transfer the "dark knowledge" [6] of the intra-camera similarity distribution to the cross-camera distribution to align the feature spaces across cameras.

In this paper, a cross-camera self-distillation(CCSD) framework is proposed to alleviate the effect of camera variation by performing splitting and alignment in the clustering and training phases, respectively. In the clustering phase, we design a camera-aware cluster refinement mechanism to obtain reliable camera-aware clusters. Specifically, we first obtain global clusters by camera-agnostic clustering [7] and then split them into multiple camera-aware clusters. Note that camera-aware clusters are expected to be accurate enough to reduce the damage from noise pseudo labels. Therefore, we evaluate the compactness of the camera-aware clusters and split the loose ones. We can obtain the refined camera-aware clusters and reserve the association across cameras at the same time in this way. In the training phase, we propose a CCSD module to align the features from different cameras. We first build a memory bank for each camera, which is initialized by the centers of the refined clusters. Then, the intra-camera or cross-camera similarities can be calculated between the sample and the clusters from the same or different cameras. Since the intra-camera similarity is more reliable than the cross-camera similarity, we propagate the knowledge from the former to the latter. Owing to the generalization and ensemble abilities of self-distillation, our method is able to learn a more consistent feature space across cameras. We conduct extensive experiments on three public datasets, and those experimental results show that the proposed approach performs favorably against the state-of-the-art unsupervised person Re-ID approaches.

The main contributions of this work can be summarized as follows:

1) We propose a cross-camera self-distillation framework, termed as CCSD, for unsupervised person re-identification. With a split-and-alignment strategy, our method can effectively alleviate the effect of camera variation.
2) We propose applying a self-distillation method to align the feature space among different cameras. By transferring the knowledge of similarity distribution from intra-camera to cross-camera, our method learns a more consistent feature space.
3) We conduct extensive experiments and component studies to demonstrate the superiority of the proposed method for unsupervised person Re-ID.

2 Related work

2.1 Unsupervised person re-identification

Unsupervised Re-ID methods can be grouped into purely unsupervised or unsupervised domain adaptation (UDA) based categories according to whether external labeled datasets are used or not. Our work belongs to the purely unsupervised category without using any labeled data.

In purely unsupervised Re-ID methods, Fan et al. [8] attempted to use clustering methods. Song et al. [2] set up the base framework for clustering-based unsupervised person Re-ID. It proposed a novel self-training scheme to iteratively make guesses for unlabeled target data based on an encoder and train the encoder based on the guessed labels. More importantly, the way they generated pseudo labels that combined the k-reciprocal Jaccard distance [9] and DBSCAN had been adopted by many later works. Fu

et al. [10] built multiple clusters from different views to train the encoder. Ge et al. [3] proposed a novel self-paced contrastive learning framework with hybrid memory to make full use of all valuable information. Dai et al. [11] presented cluster contrast which stored feature vectors and computes contrast loss at the cluster level. Although these methods have achieved impressive accuracy, they still suffer from camera variation.

In recent years, some approaches have been proposed to address the problem of camera variation. Wu et al. [12] pointed out that the inconsistent pairwise similarity distributions degraded the matching performance. Xuan & Zhang [13] and Yang et al. [14] had similar statements. However, they adopt various specific schemes to solve this problem. In particular, Wu et al. [12] designed a camera-aware similarity consistency loss from the feature distribution level; Xuan & Zhang [13] proposed a novel intra-inter camera similarity for pseudo-label generation from the metric level; Yang et al. [14] introduced meta-learning to make the model more generalized to feature distortion caused by camera variation from the optimization level. In our work, we try to eliminate the inconsistency across cameras in a self-distillation manner.

2.2 Self-distillation method

Self-distillation methods have attracted much attention in recent years by transferring knowledge during a single network. Hinton et al. [6] proposed the concept of "dark knowledge" which could be transferred from teacher models to student models to improve the performance of the students. In self-distillation, the teacher models and student models use the same model. Zhang et al. [15] proposed a new framework to distill knowledge from the deeper sections of the network into its shallow sections. Yun et al. [16] applied class-wise self-knowledge distillation to match the output distributions of the training model between intra-class samples and augmented samples within the same source with the same model. Ge et al. [17] presented batch knowledge ensembling to produce refined soft targets for anchor images by propagating and ensembling the knowledge of the other samples in the same mini-batch. In this work, different from the existing self-distillation, we take account of the camera views of each class (i.e., identity in Re-ID). We are the first to employ self-distillation to address the camera variation issue in person Re-ID to the best of our knowledge.

3 Methods

Given a Re-ID dataset $\mathcal{D} = \{(x_i, c_i)\}$, where x_i is the ith sample in $\mathcal{D}$ and c_i is the camera index of x_i. Our objective is to train a Re-ID model f embedding images into feature space where images of the same identification can be matched. Our CCSD framework is based on traditional two-phase unsupervised person Re-ID methods. As Fig. 2 shows, in the clustering phase, a camera-aware cluster refinement mechanism is designed to generate refined clusters; in the training phase, we train a better intra-camera similarity distribution and align the similarity distribution from the intra-camera to cross-camera distribution.

3.1 Camera-aware cluster refinement mechanism

In this section, we first introduce our camera-aware cluster refinement mechanism which refines camera-aware clusters for computing the accurate similarity distribution. It contains four steps as demonstrated in Fig. 2.

Global clustering We first employ the camera-agnostic clustering method named DBSCAN [7] on the whole dataset $\mathcal{F} = \{f(x_i)|x_i \in \mathcal{D}\}$, where $f(x_i)$ denotes the feature of x_i extracted by f. All noise samples discovered by the clustering algorithm are discarded and the reserved samples form a new dataset $\mathcal{D}' = \{(x_i, y_i, c_i)|y_i \in \{1, 2, \ldots, N_y\}\}$, where y_i is the global pseudo label of x_i and N_y is the number of global clusters.

Split by cameras Although clustering samples within the same camera such as in Ref. [13] could avoid camera variation and generate compact clusters within one camera, it will lose the cross-camera association, which is indispensable in the cross-camera self-distillation module. Hence, we simply split the global clusters according to cameras after global clustering. A global cluster $\{(x_i, y_i, c_i) \in \mathcal{D}'|y_i = m\}$ denoted as $\mathcal{C}_m$ can be split into a set of camera-aware clusters $\{\mathcal{C}_m^s|s \in \mathcal{S}_m\}$, where $\mathcal{S}_m$ is the set of the cameras in $\mathcal{C}_m$. We use both the camera index s and global pseudo label m to distinguish different camera-aware clusters.

Compute compactness Naturally, there are still a few imperfect camera-aware clusters, which are distorted by noisy samples. The main reason is that the global clusters are produced by the clustering method with hyper-parameters adapted to discover the cross-camera association on the whole dataset. The compactness of the camera-aware clusters $\mathcal{C}_m^s$ in camera s can be evaluated by $m\text{Sil}(\mathcal{C}_m^s)$ which is the average value of Silhouette score [18], which measures how well a sample belongs to its own cluster. Note that we only use the sample within the same camera in the calculation of the Silhouette scores. This means that $m\text{Sil}(\mathcal{C}_m^s)$ only indicates the compactness of $\mathcal{C}_m^s$ in camera s rather than in the whole dataset.

Split by compactness If $m\text{Sil}(\mathcal{C}_m^s)$ is less than a default threshold λ, we employ the hierarchical clustering method on $\mathcal{C}_m^s$ and split it into two refined camera-aware clusters $\mathcal{C}_m^{s\,\prime}$ and $\mathcal{C}_m^{s\,\prime\prime}$. For convenience, we label the sample $x_i \in \mathcal{C}_m^{s\,\prime}$ with camera-aware pseudo labels denoted as $\tilde{y}_i \in \{1, 2, \ldots, N_s\}$ in camera s, where N_s is the number of refined camera-aware clusters in camera s. $\mathcal{K}_n^s$ denotes a cluster

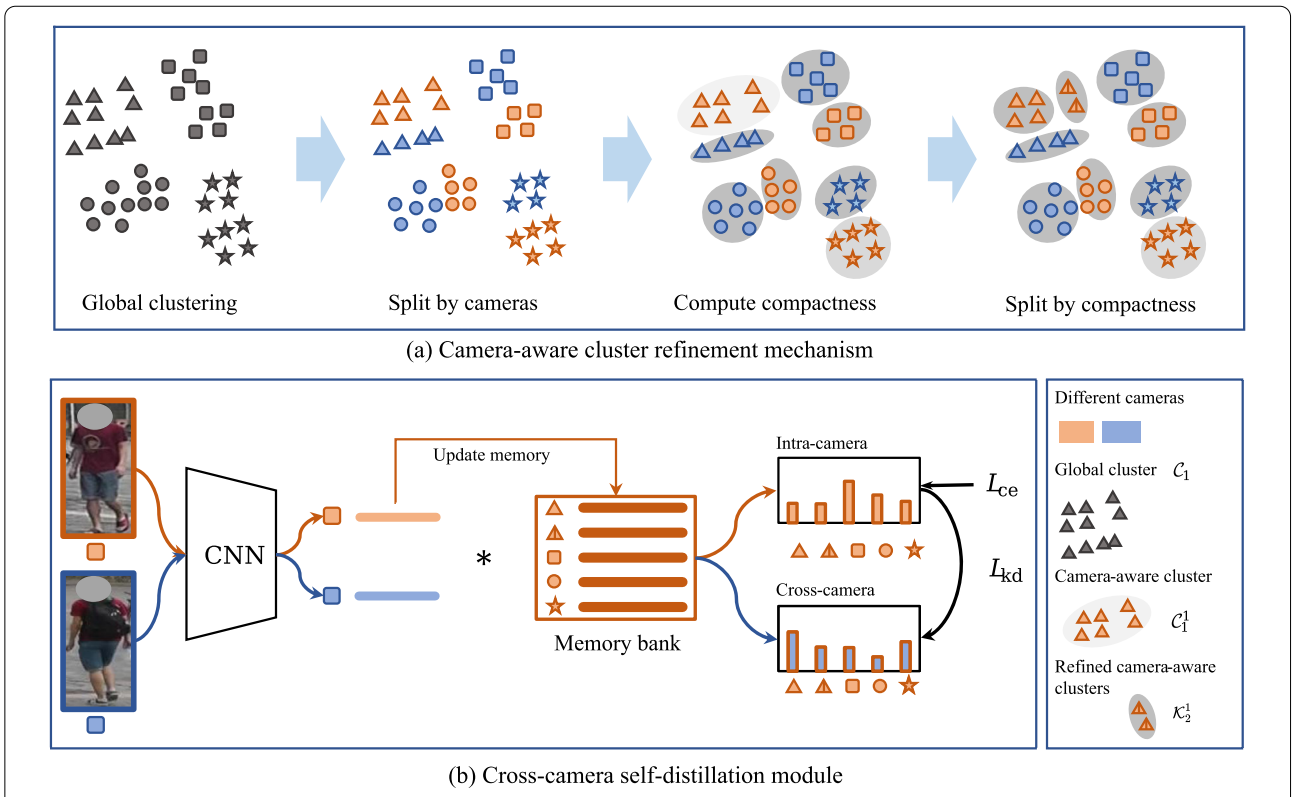


Figure 2 Overview of our CCSD framework. In the clustering phase, we first adopt camera-agnostic clustering to generate global clusters and split them into camera-aware clusters by camera views and compactness successively. In the training phase, a memory bank is built for each camera to calculate intra-camera and cross-camera similarities. We train the model to learn a reliable intra-camera similarity distribution by L_{ce} and distill the knowledge of similarity distribution from intra-camera to cross-camera by L_{kd}

with camera-aware pseudo labels n in camera s. Note that the former global pseudo labels of $\mathcal{K}_n^s$ are still reserved to indicate the cross-camera relations.

3.2 Cross-camera self-distillation module

With the help of the camera-aware cluster refinement mechanism, we compute the intra-similarity or cross-similarity distributions by the similarity between the samples and refined camera-aware clusters from the same or different cameras.

Specifically, we construct a memory bank $\mathcal{M}^s$ in camera s. The memory bank is initialized by

$$\mathcal{M}^s[j] = \frac{1}{|\mathcal{K}_j^s|} \sum_{x \in \mathcal{K}_j^s} f(x), \tag{1}$$

where $\mathcal{M}^s[j]$ denotes the jth entry of $\mathcal{M}^s$ storing the cluster center of $\mathcal{K}_j^s$. $\mathcal{K}_j^s$ is a refined camera-aware cluster with camera-aware pseudo labels j in camera s.

The similarity distribution of samples x in camera s is defined as

$$p^s(j|x) = \frac{\exp(\frac{\langle \mathcal{M}^s[j], f(x) \rangle}{\tau})}{\sum_{i=1}^{N_s} \exp(\frac{\langle \mathcal{M}^s[i], f(x) \rangle}{\tau})}, \tag{2}$$

where $p^s(j|x)$ is the similarity between x and $\mathcal{K}_j^s$, τ is a temperature hyper-parameter, $\langle \cdot \rangle$ denotes the cosine similarity, and N_s denotes the number of refined camera-aware clusters. If x comes from the camera s, $p^s(x)$ denotes the intra-camera similarity distribution of x. Otherwise, $p^s(x)$ is the cross-camera similarity distribution of x.

To learn a more reliable intra-camera similarity distribution, we shorten the distance between the samples and their respective camera-aware clusters by a memory-based non-parametric cross-entropy loss defined as:

$$L_{ce} = -\lg p^s(n|x), \tag{3}$$

where $x \in \mathcal{K}_n^s$. $p^s(n|x)$ denotes the similarity between x and $\mathcal{M}^s[n]$ which is the cluster center of $\mathcal{K}_n^s$.

Additionally, we align the intra-camera and cross-camera similarity distributions in a self-distillation manner. Given a sample $x \in \mathcal{K}_n^s$, we transfer its knowledge of the intra-camera similarity distribution to another sample $\tilde{x}$ from the same global cluster but different cameras by Kullback-Leibler divergence loss defined as

$$L_{\text{kd}} = -\sum_{j\in\{1,2,\ldots,N_s\}} p^s(j|x)\lg p^s(j|\tilde{x}). \tag{4}$$

In return, the intra-camera similarity distribution of $\tilde{x}$ in its own camera can be transferred to the cross-camera similarity distribution of x.

The total loss of our method is:

$$L_{\text{total}} = L_{\text{ce}} + \beta L_{\text{kd}}, \tag{5}$$

where β is the hyper-parameter to balance L_{ce} and L_{kd}.

When we update the model parameters with L_{total} by gradient descent during back-propagation, the memory bank $\mathcal{M}^s$ is updated by

$$\mathcal{M}^s[j]_{(t)} \leftarrow \mu\mathcal{M}^s[j]_{(t-1)} + (1-\mu)f(x_i), \tag{6}$$

where $x_i \in \mathcal{K}_j^s$, μ is the updating rate of the memory banks.

4 Experiments

4.1 Datasets and evaluation metrics

We evaluate our methods on two common person Re-ID benchmarks, i.e., Market1501 [19] and MSMT17 [20].

Market1501 consists of 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing. All images of Market1501 are captured from 6 cameras.

MSMT17 is the largest and most challenging dataset which contains 4101 identities and 126,441 bounding boxes, captured by 12 outdoor cameras, 3 indoor cameras, and during 12 time slots.

In the training phase, we only use images and camera labels without any annotation information about identity. In testing, performance is evaluated by cumulative matching characteristics (CMC) Rank1, Rank5, Rank10 accuracies and mean average precision (mAP).

4.2 Implementation details

Clustering settings To generate global pseudo labels, we first use the metric method named k-reciprocal Jaccard distance to calculate the distance matrix on the whole dataset, and k is set to 30. In the clustering approach DBSCAN [7], we set the distance threshold to 0.4 and the minimum number of neighbors in a core point to 4. In the camera-aware cluster refinement approach, we set the threshold of splitting by compactness λ = mean(mSil) – 0.5std(mSil), where mean(·) denotes the function to take the mean and std(·) denotes the function to calculate the standard deviation. Therefore, the threshold is adjusted adaptively according to the statistics of the current clusters. The hyper parameter β in Equ. (5) is experimentally set to 0.8.

General training settings We employ ResNet50 as our backbone of the encoder with the parameters pre-trained on ImageNet. All sub-module layers after layer-4 are replaced with a series of global average pooling (GAP), batch normalization layer and L2-normalization layer, which will produce 2048-dimensional features. We use an Adam optimizer with a weight decay of 5×10^{-4} to train our Re-ID model. The learning rate is set as 3.5×10^{-4} at the beginning and is reduced to 1/10 of its previous value every 20 epochs for a total of 60 epochs. The updating rate μ of the memory banks is set to 0.1. In the training phase, we sample 256 images of 16 global pseudo labels (16 instances for each label). Specially, we sample the instances from at least 2 cameras if it is possible.

Data augmentation All images are uniformly resized to 256×128. Following Ref. [11], we adopt random horizon flipping, random padding with 10 pixels, random cropping, and random erasing for each image for data augmentation.

4.3 Comparison with state-of-the-art methods

The existing state-of-the-art unsupervised Re-ID methods can also be categorized into purely unsupervised and unsupervised domain adaptation (UDA) methods. The former trains a Re-ID model without any annotations, while the latter transfers the knowledge on a labeled source dataset to an unlabeled target dataset. For a fair comparison, we group them into unsupervised domain adaptation and fully unsupervised in Table 1. In addition, we also evaluate our method on the supervised Re-ID task.

Comparison with UDA methods We first compare our method with state-of-the-art UDA methods including AD-Cluster [21], DG-Net++ [22], SpCL [3], GLT [23], and UMDA [24]. They utilize extra labeled data to boost the performance and usually have better performance than purely unsupervised methods. As Table 1 demonstrates, our method exceeds all the UDA methods, which even utilize extra labeled data.

Comparison with purely unsupervised methods We enumerate the performances of eight most recent purely unsupervised methods including HCT [25], SpCL [3], DSCE [14], RLCC [26], IICS [13], CAP [4], ICE [27], CCL [11], ISE [28], PPLR [29] for comparison. As Table 1 indicates, our method outperforms all of them by a great amount which proves the superiority of our method. Specifically, the mAP of our method surpasses ISE by 0.7% on Market1501 and PPLR by 0.2% on MSMT17. This demonstrates the effectiveness of our method.

Table 1 Comparison with state-of-the-art methods. w\GT denotes performing the unsupervised Re-ID methods with true labels. The best results are marked in bold

Method	References	Market1501				MSMT17			
		mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10
Unsupervised domain adaptation									
AD-Cluster [21]	CVPR20	68.3	86.7	94.4	96.5	–	–	–	–
DG-Net++ [22]	ECCV20	63.8	78.9	87.8	90.4	22.1	48.8	60.9	65.9
SpCL [3]	NeurIPS20	76.7	90.3	96.2	97.7	26.5	53.1	65.8	70.5
GLT [23]	CVPR21	79.5	92.2	96.5	97.8	27.7	59.5	70.1	74.2
UMDA [24]	CVPR21	81.5	92.9	97.6	98.4	34.9	64.7	76.2	80.2
Fully unsupervised									
HCT [25]	CVPR20	56.4	80.0	91.6	95.2	–	–	–	–
SpCL [3]	NeurIPS20	72.6	87.7	95.2	96.9	–	–	–	–
DSCE [14]	CVPR21	53.8	73.8	84.2	–	15.5	35.2	48.3	–
RLCC [26]	CVPR21	77.7	90.8	96.3	97.5	27.9	56.5	68.4	73.1
IICS [13]	CVPR21	72.9	89.5	95.2	97.0	26.9	56.4	68.8	73.4
CAP [4]	AAAI21	79.2	91.4	96.3	97.7	36.9	67.4	78.0	81.4
ICE [27]	ICCV21	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4
CCL [11]	Arxiv21	82.1	92.3	96.7	97.9	27.6	56.0	66.8	71.5
ISE [28]	CVPR22	84.7	94.0	97.8	**98.8**	35.0	64.7	75.5	79.4
PPLR [29]	CVPR22	84.4	**94.3**	97.8	98.6	42.2	73.3	**83.5**	**86.5**
CCSD	Ours	**85.4**	93.9	**98.0**	98.7	**42.4**	**73.3**	82.8	85.8
Supervised									
CAP (w\GT)	AAAI21	85.1	93.3	97.5	98.4	53.7	77.1	87.4	90.8
ICE (w\GT)	ICCV21	86.6	95.1	98.3	98.9	50.4	76.4	86.6	90.0
CCSD (w\GT)	Ours	87.1	94.9	98.2	99.0	53.3	78.1	88.4	91.4

Table 2 Generalization of CCSD with other modules on Market1501. IBN denotes that the backbone applies IBN-ResNet50. GeM represents the GeM pooling layer

Methods	Market1501			
	mAP	Rank1	Rank5	Rank10
ResNet-50	85.4	93.9	98.0	98.7
ResNet-50+GeM	85.0	94.0	97.3	98.5
IBN-ResNet-50+GeM	86.2	94.2	98.0	98.9

Comparison with supervised methods Most purely unsupervised methods can be converted to fully supervised methods by directly training the Re-ID model with ground truth labels rather than clustering-based pseudo labels. Compared to other purely unsupervised methods, e.g. ICE and CAP, our method maintains competitiveness and a dominant position under the supervised setting.

Generalization of CCSD with other modules To evaluate the generalization of our method, we also test our method with other common modules, i.e., instance-batch normalization (IBN) and generalized mean pooling (GeM). As visualized in Table 2, the performance of our model is further improved with an IBN-ResNet50 backbone network and GeM pooling layer, which shows the good generalization ability of our method.

4.4 Ablation study

In this section, we study the components of CCSD to validate their effectiveness, i.e., split by camera (SCam), split by compactness (SCom) and cross-camera self-distillation mechanism (L_{kd}). The baseline of our method is set to an ordinary clustering-based method with a memory-based non-parametric softmax loss, that is, our method without SCam, SCom and L_{kd}. The comparison results are reported in Table 3, and some visualization results are displayed in Fig. 3. We analyze them in detail as follows.

Evaluation of SCam We evaluate the SCam by adding it to baseline, denoted as CCSD_1 in Table 3. We can see that CCSD_1 without cross-camera alignment performs much worse than the baseline. The reason is that, without considering the cross-camera constraint, the model has difficulty in matching the images across cameras. In Fig. 3, we can also observe that the samples of CCSD_1 with the same ID are more dispersive than those of the baseline. This means cross-camera alignment is necessary for our method.

The effectiveness of L_{kd} We add L_{kd} to CCSD_1, termed CCSD_2, to verify the effectiveness of the cross-camera self-distillation mechanism. The performance will sharply rebound and surpass the baseline, as presented in Table 3. It demonstrates that the cross-camera self-distillation can distill the dark knowledge within one camera to each other adequately. From Fig. 3, we can observe that the samples of

Table 3 Evaluation of the components of CCSD. SCam and SCom denote the procedures of splitting by cameras and splitting by compactness in the camera-aware cluster refinement mechanism, respectively. L_{kd} denotes the cross-camera self-distillation module

Methods	Components			Market1501				MSMT17			
	SCam	SCom	L_{kd}	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10
Baseline				80.6	91.3	96.4	97.6	19.9	43.3	53.3	58.5
CCSD_1	✓			54.1	74.0	87.3	91.3	20.8	45.3	57.9	63.4
CCSD_2	✓		✓	83.3	92.8	97.1	98.3	40.6	72.2	82.1	85.3
CCSD_3	✓	✓	✓	85.4	93.9	98.0	98.7	42.4	73.3	82.8	85.8

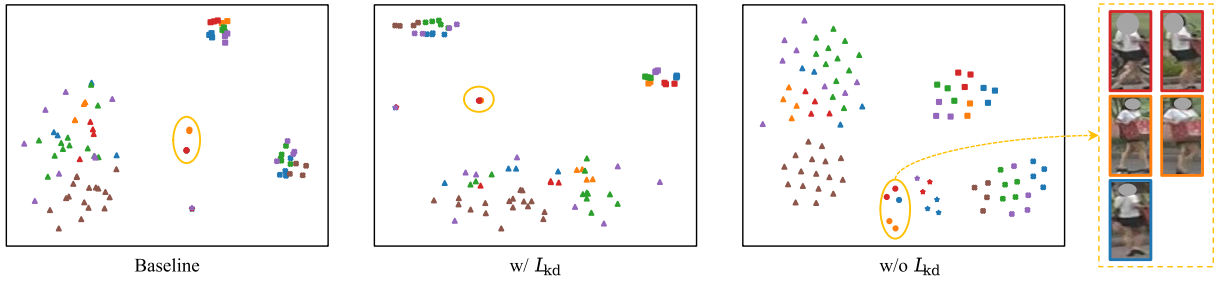


Figure 3 t-SNE visualization of features extracted by the models under different settings. Different shapes and different colors represent samples with different IDs and different cameras, respectively

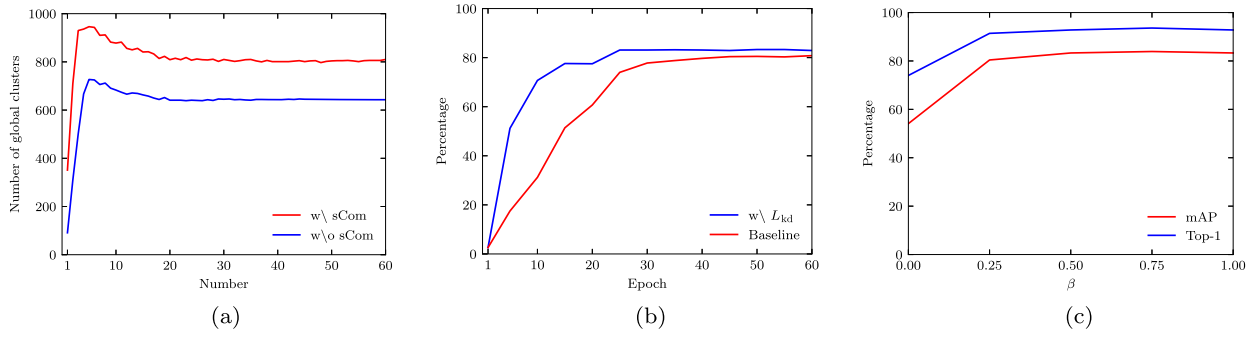


Figure 4 Evaluation of our components and parameters on Market-1501. (a) Dynamic global cluster numbers during 60 training epochs. (b) Dynamic mAP during 60 training epochs. (c) Parameter analysis of β

the same person in different cameras become more compact. We believe the reason is that the interaction between L_{ce} and L_{kd} offers a better direction for optimization than the baseline in the training progress. Figure 4(b) demonstrates that the model converges faster and better under supervision by L_{kd} and L_{ce} than baseline.

The effectiveness of SCom We evaluate the SCom by adding it to CCSD_2, termed CCSD_3. Table 3 shows that SCom can further improve the performance. This means that refining camera-aware clusters by compactness can reduce the damage from noise labels. In Fig. 4(a), we can see that the number of global clusters increases in the training phase after splitting by compactness. This shows that split by compactness is complementary to camera-aware methods.

Influence of β β is the import hyper-parameter that determines the balance between L_{ce} and L_{kd}. As Fig. 4(c) shows, the performance improves as the importance of L_{kd} increases, but it will decline slightly at last. this is because L_{kd} offers cross-camera knowledge to the model by aligning the intra-camera similarity across cameras. However, there is still discrepancy in different cameras, which harms the performance if it forces them to be overly consistent. A suitable β-value can yield the optimum result.

Analysis of the time-consuming of each stage in clustering By counting the average of the time-consuming of each

Table 4 Time-consuming of each stage in clustering

Stage	Time/s
Distance calculation	28.42028
Global clustering	1.714848
Cluster refinement	2.022478

clustering stage in all training epochs, we can see that time is mainly consumed in the distance calculation. As Table 4 presents, cluster refinement is slightly less than global clustering but much less than distance calculation. Hence, the computational burden imposed by cluster refinement is slight.

5 Conclusion

In this paper, we propose a new approach to solve the purely unsupervised Re-ID problem. It is impressive that the main obstacle in unsupervised Re-ID is the inconsistent feature space and unreliable pseudo labels caused by camera variation. To eliminate the discrepancy in feature space across cameras, we creatively propose a cross-camera self-distillation framework, which employs knowledge distillation to align the similarity distribution from intra-camera to cross-camera. The experimental results on benchmarks testify the effectiveness of our method. There is still some room for improvement in our method. The hyper-parameter β needs to be manually adjusted to maximize the performance of the model, because CCSD can only align the the feature distributions under different cameras in a global way. We will continue to investigate how to more adaptively and accurately align the distributions of different cameras.

Acknowledgements
Not applicable.

Funding
This work was supported by the National Natural Science Foundation of China (No. 62176097) and the Outstanding Youth Foundation of Hubei Province (No. 2022CFA055).

Abbreviations
CCSD, cross-camera self-distillation; CMC, cumulative matching characteristics; GAP, global average pooling; GeM, generalized mean pooling; IBN, instance-batch normalization; mAP, mean average precision; Re-ID, re-identification; SCam, split by camera; SCom, split by compactness; UDA, unsupervised domain adaptation.

Availability of data and materials
The datasets used in this study Market1501 and MSMT17 are available in Ref. [19] and Ref. [20] respectively. The data analyzed during the current study and the code are available from Jiyuan Chen upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed to the study conception and design. All authors contributed significantly to the analysis and manuscript preparation. All authors performed the data analyses and wrote the manuscript. JC collected data and performed the experiment. NS helped perform the analysis with constructive discussions. CG and LS revised the manuscript. All authors read and approved the final manuscript.

Author details
[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. [2]School of Communication and Electronic Engineering, East China Normal University, Shanghai, China.

Received: 3 May 2023 Revised: 25 October 2023
Accepted: 25 October 2023 Published online: 15 November 2023

References

1. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., et al. (2020). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, *22*(10), 2597–2609.
2. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., et al. (2020). Unsupervised domain adaptive re-identification: theory and practice. *Pattern Recognition*, *102*, 107173.
3. Ge, Y., Zhu, F., Chen, D., Zhao, R., & Li, H. (2020). Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID. In H. Larochelle, M. Ranzato, R. Hadsell et al.(Eds.), *Proceedings of the 34th international conference on neural information processing systems*, Red Hook: Curran Associates.
4. Wang, M., Lai, B., Huang, J., Gong, X., & Hua, X.-S. (2021). Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the 35th AAAI conference on artificial intelligence* (pp. 2764–2772). Palo Alto: AAAI Press.
5. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.
6. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint. arXiv:1503.02531.
7. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. M. Fayyad (Eds.), *Proceedings of the 2nd international conference on knowledge discovery and data mining* (pp. 226–231). Palo Alto: AAAI Press.
8. 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*, *14*(4), 1–18.
9. Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3652–3661). Piscataway: IEEE.
10. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., & Huang, T. S. (2019). Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision* (pp. 6112–6121). Piscataway: IEEE.
11. Dai, Z., Wang, G., Zhu, S., Yuan, W., & Tan, P. (2021). Cluster contrast for unsupervised person re-identification. arXiv preprint. arXiv:2103.11568.
12. Wu, A., Zheng, W.-S., & Lai, J.-H. (2019). Unsupervised person re-identification by camera-aware similarity consistency learning. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision* (pp. 6922–6931). Piscataway: IEEE.
13. Xuan, S., & Zhang, S. (2021). Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11926–11935).
14. Yang, F., Zhong, Z., Luo, Z., Cai, Y., Lin, Y., Li, S., et al. (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* (pp. 4855–4864). Piscataway: IEEE.
15. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: improve the performance of convolutional neural networks via self distillation. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision* (pp. 3713–3722). Piscataway: IEEE.
16. Yun, S., Park, J., Lee, K., & Shin, J. (2020). Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13873–13882). Piscataway: IEEE.

17. Ge, Y., Choi, C.L., Zhang, X., Zhao, P., Zhu, F., Zhao, R., et al. (2021). Self-distillation with batch knowledge ensembling improves imagenet classification. arXiv preprint. arXiv:2104.13298.
18. Silhouettes, P. J. R. (1987). A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.
19. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: a benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, Piscataway: IEEE.
20. Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Piscataway: IEEE.
21. Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., et al. (2020). Ad-cluster: augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9021–9030). Piscataway: IEEE.
22. Zou, Y., Yang, X., Yu, Z., Kumar, B. V. K. V., & Kautz, J. (2020). Joint disentangling and adaptation for cross-domain person re-identification. In A. Vedaldi, H. Bischof, T. Brox et al.(Eds.), *Proceedings of 16th European conference on computer vision* (pp. 87–104). Cham: Springer.
23. Zheng, K., Liu, W., He, L., Mei, T., Luo, J., & Zha, Z.-J. (2021). Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5310–5319). Piscataway: IEEE.
24. Bai, Z., Wang, Z., Wang, J., Hu, d., & Ding, E. (2021). Unsupervised multi-source domain adaptation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12914–12923). Piscataway: IEEE.
25. Zeng, K., Ning, M., Wang, Y., & Guo, Y. (2020). Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13657–13665). Piscataway: IEEE.
26. Zhang, X., Ge, Y., Qiao, Y., & Li, H. (2021). Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3436–3445). Piscataway: IEEE.
27. Chen, H., Lagadec, B., & Bremond, F. (2021). ICE: inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14960–14969). Piscataway: IEEE.
28. Zhang, X., Li, D., Wang, Z., Wang, J., Ding, E., Shi, J.Q., et al. (2022). Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7369–7378). Piscataway: IEEE.
29. Cho, Y., Kim, W. J., Hong, S., & Yoon, S.-E. (2022). Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7298–7308). Piscataway: IEEE.

Publisher's Note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com