CDE-Learning: Camera Deviation Elimination Learning for Unsupervised Person Re-identification

Jinjia Peng[1], **Songyu Zhang**[1, *], **Huibing Wang**[2, *]
[1]School of Cyber Security and Computer, Hebei University, Hebei, China
Hebei Machine Vision Engineering Research Center, China
[2]College of Information Science and Technology, Dalian Maritime University
pengjinjia@hbu.edu.cn, zhangsongyu@stumail.hbu.edu.cn, huibing.wang@dlmu.edu.cn

Abstract

Unsupervised Person Re-identification (Re-ID) aims to identify the same person shot from non-overlapping cameras without any annotated data. In this task, attributes such as contrast, saturation, and resolution of the camera cause the deviation in the target features. Since the camera label is readily available, they are employed to achieve the constraints across cameras and smooth the deviations during the model training phase. However, features from the same camera are prone to generating false positives due to the identical camera properties, which induce camera deviations on pseudo-label assignment. To address this problem, this paper proposes a novel camera-unbiased method named Camera Deviation Elimination Learning (CDE-Learning). In the method, the Camera Deviation Compensation (CDC) module is designed to align data distributions from disparate cameras. This alignment decouples camera information from identity information during the pseudo-label allocation. Our Camera Deviation Balancing (CDB) module integrates different camera constraints in a united loss and adjusts camera constraints by constructing contrastive pairs between intra-camera and inter-camera. After explicit constraints, the Camera Attribution Auxiliary (CAA) task predicts whether a pair of images originate from the same camera to implicitly enhance the capacity to distinguish the camera deviation. We demonstrated the superior performance of the proposed CDE-Learning on benchmark datasets.

Code — https://github.com/zsszyx/CDE-Learning

Introduction

Unsupervised person re-identification (Re-ID) (Fu et al. 2022; Ge et al. 2020; Zhang et al. 2021) aims to match individuals across non-overlapping camera views. The primary challenge is the appearance variations caused by diverse camera perspectives, known as irrelevant camera deviation. To address this issue, some approaches (Wang et al. 2021; Lee et al. 2023; Zhang et al. 2023) attempt to incorporate camera labels to establish camera agents and then create cross-camera contrast pairs to mitigate the impact of inter-camera perspective variations. However, these methods neglect the significant appearance variations within the camera complexity. Taking into account this issue, different from existing clustering methods(Yao et al. 2024; Peng et al. 2024), some works (Xuan and Zhang 2021; Wu, Zheng, and Lai 2019; Wang et al. 2022, 2024, 2020) increase the number of intra-camera contrast pairs to explore varied appearance features in the same camera. They separate inter-camera and intra-camera learning as camera constraints, which unfortunately need to be modulated by hyper-parameter tuning.

On the other hand, these approaches ignore the effect of camera deviation on pseudo-label allocation during clustering. The features for clustering inevitably contain deviation noise induced by cameras. Consequently, features in the same camera may lead to false positive samples, affecting the quality of pseudo-labels and accumulating wrong prior knowledge. In addition, the hyperparameter setting to balance camera constraints requires manual parameter tuning to achieve optimal performance in different scenarios, reducing their flexibility and applicability.

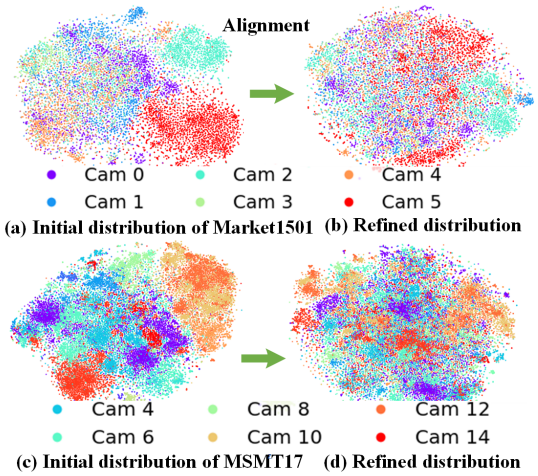

Figure 1: Comparison of feature distributions before (a,c) and after (b,d) alignment. The camera deviation leads to more concentrated features under the same camera. After alignment, the feature distribution becomes more uniform and thus can avoid the false positive samples influenced by the camera deviation.

* Corresponding author

To address the camera deviation in unsupervised Re-ID, this paper proposes a camera-unbiased framework: Camera Deviation Elimination Learning. Our approach consists of three components: the Camera Deviation Compensation module, the Camera Deviation Balancing module, and the Camera Attribution Auxiliary task.

In order to reduce the deviation in clustering, the Camera Deviation Compensation (CDC) module aligns feature distributions of various cameras, optimizing the inferior pseudo-labels. Specifically, we first construct a camera domain which is formed by the features described as the same camera label. The CDC module reduces the gap between the centroids of camera domains to eliminate the interference of potential camera deviation. Figure 1 illustrates the aligned feature distributions of different datasets. As shown in Figure 1, our method improves the uniformity of the distribution and reduces the camera deviation, leading to more accurate clustering results.

In addition, our Camera Deviation Balancing (CDB) module unifies camera constraints and achieves an adaptive balancing of them. The camera constraints are united within an integral loss to eliminate extra parameter tuning and enhance the adaptability of CDE-Learning. With the fact that a person may be captured by multiple cameras or only a few cameras and requiring different camera constraints, our approach utilizes a multi-prototype dictionary to dynamically generate intra-camera or inter-camera pairs to adjust camera constraints based on the situation of the individual. The Camera Attribute Auxiliary task (CAA) offers an implicit approach to the model to focus on distinguishing camera backgrounds while simultaneously learning to recognize pedestrian identities. This strategy further mitigates camera deviation during the training process. In summary, our method makes the following innovations:

- This paper proposes a camera deviation elimination learning framework. A camera deviation compensation module is designed to align the feature distributions from different cameras, which mitigates the camera deviation at the clustering stage and generates reasonable pseudo-labels.
- In the camera deviation balancing module, it integrates intra and inter-camera learning in a united loss by adaptively constructing contrast pairs of camera constraints, avoiding complex parameter tuning.
- The camera attribute auxiliary task enhances the model's understanding of subjects and backgrounds to improve its learning capacity under unsupervised conditions.

Related Works

Unsupervised Person Re-ID

Unsupervised person re-identification (He et al. 2024; Huang et al. 2024; Chen et al. 2023; Liu et al. 2023; Li, Li, and Guo 2022; Peng, Jiang, and Wang 2023) can be broadly divided into two categories: Unsupervised Domain Adaptation (UDA) and Purely Unsupervised Re-ID (Cho et al. 2022a; Isobe et al. 2021; Ji et al. 2021). UDA approaches aim to develop a re-ID model by incorporating labeled source data and unlabeled target data for training. In contrast, Purely Unsupervised Re-ID methods focus on training models exclusively with unlabeled data with an initial pre-training phase. A series of studies primarily focus on alignment to mitigate the distribution shift between cameras or domains at the pixel level. Wang et al. (Wang and Zhang 2020) iteratively predict multi-class labels and update the network with a memory-based multi-label classification loss. MGS (Li and Qi 2023) introduces a self-paced clustering method guided by multi-label learning for unsupervised pedestrian re-identification, aiming to improve the quality and accuracy of pseudo-labels. Our method is dedicated to enhancing the quality of prior knowledge within the context of purely unsupervised person re-identification.

Contrastive Learning

Recent advancements in unsupervised representation learning have conceptualized the learning process as analogous to a dictionary lookup, where the goal is to identify and retrieve relevant information from a vast pool of unlabelled data. ProCo (Du et al. 2024) is a novel approach to long-tail visual recognition that models class distributions for sampling contrastive pairs, overcoming the limitation of batch size in supervised contrastive learning. Huang (Huang, Yi, and Zhao 2023) provides a theoretical interpretation of how models trained with contrastive self-supervision can be applied to downstream tasks. MoCo (Chen et al. 2020) introduces momentum contrast, a technique that capitalizes on the momentum of a moving target to create a more robust and scalable self-supervised learning framework, facilitating improved feature extraction from unlabelled datasets. Furthermore, Barlow Twins (Houlsby et al. 2021) stands out as a self-supervised learning method that effectively reduces redundancy in the representations learned by two identical networks. Our approach integrates augmented instances and cluster centers within consistency regularization constraints, thereby significantly enhancing the efficacy of contrastive learning.

Learning to Eliminate Camera Deviation

Overview

Learning to eliminate camera deviation aims to decouple the feature distribution and diminish camera deviation during the pseudo-label assignment and cross-camera feature learning. Figure 2 presents a total illustration of the proposed CDE-Learning. It comprises three main components: the Camera Deviation Compensation (CDC) module, the Camera Deviation Balancing (CDB) module, and the Camera Attribute Auxiliary (CAA) task.

To eliminate the camera deviation in the pseudo-label assignment, the CDC module aligns distributions from different cameras. In particular, the individual features under the same camera i form a camera domain F_i, and each camera domain has its unique distribution, containing identity feature information and irrelevant camera deviation information. The camera deviation in the specific camera domain leads to an increased number of false positive samples in clustering. Our method addresses these flawed clustering results by reducing the gap between the centroid s_i of camera domains. By mitigating the impact of camera deviation, our

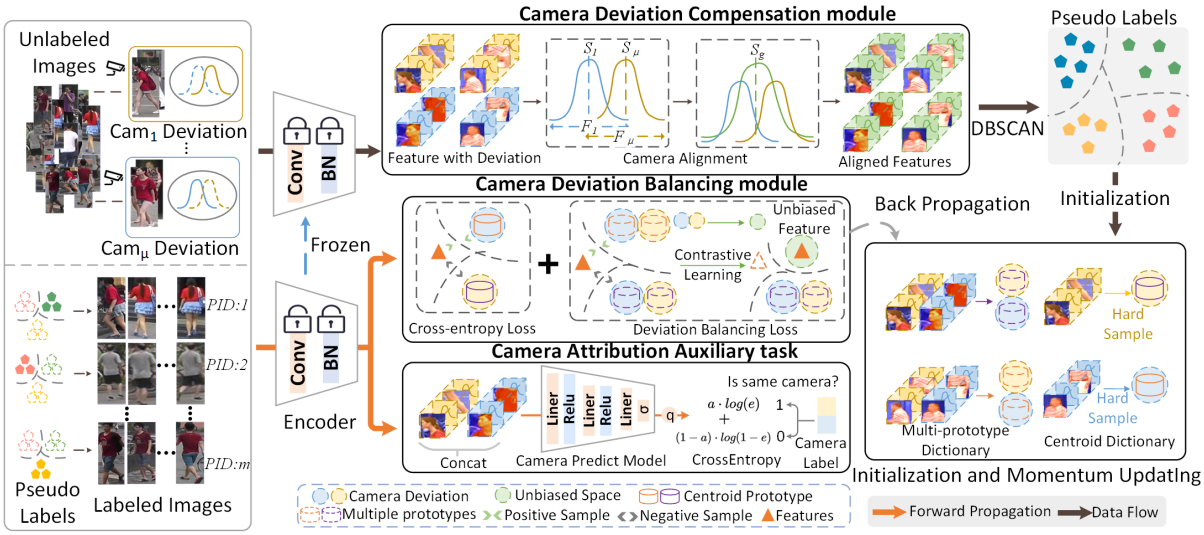

Figure 2: Illustration of a common framework for CDE-Learning. Our method optimizes the encoder by continuous iteration between clustering and training. The CDC module and CDB module optimize camera deviation in both pseudo-label allocation and cross-camera invariant feature learning. The parameters of the encoder remain fixed during clustering. Our method achieves a camera-unbiased invariant feature learning.

method can enhance the results of clustering and improve the quality of pseudo-labels.

To address the camera deviation in cross-camera feature learning, our CDB module proposes a united loss to incorporate camera constraints, avoiding the extra manual parameter tuning. Specifically, our approach designs a multi-prototype dictionary containing both intra and inter-camera features to generate contrast pairs adaptively. According to the camera constraints required by different feature learning, our method constructs contrast pairs with corresponding camera information. The contrast pairs utilized in the proposed Deviation Balancing Loss contain diverse camera knowledge, establishing an effective camera constraint that allows the model to learn camera-unbiased invariant features. Thus, our method is able to mitigate the impact of inter-camera viewpoint variation and intra-camera appearance feature variation simultaneously.

To enhance the model's understanding of the camera background during the training process, the CAA task constructs feature pairs from the extracted image features and employs a prediction model composed of multiple MLPs to predict camera attributes. The camera labels are utilized to their fullest extent, contributing not only to the optimization of pseudo-labels and the imposition of explicit camera constraints, but also serving as supervisory information for the auxiliary task, thereby improving the model's generalization capabilities across diverse multi-camera scenarios. The training details are presented in Algorithm 1.

Camera Deviation Compensation Module

While generating pseudo-labels, individual features are extracted using a pre-trained encoder, where $F \in \mathbb{R}^{n\times d}$ represent extracted features, n represents the total number of images, and d is the dimension of features. The extracted features often encompass redundant camera deviation, leading to false positive samples and affecting the quality of pseudo-labels. Our CDC module mitigates camera deviation by aligning the distributions of diverse cameras. Specifically, our method obtains the camera domain F_i by the camera labels, and $F = \{F_1, F_2, F_3, \ldots, F_\mu\}$, where μ represents the number of cameras, each camera domain F_i contains features under the same camera. Subsequently, the centroids for each camera are obtained through:

$$s_i = \frac{\sum_{j=1}^{n} f_j \cdot w_j}{\sum_{j=1}^{n} w_j}, w_j = \begin{cases} 1 & f_j \in F_i, \\ 0 & f_j \notin F_i. \end{cases} \tag{1}$$

where $f_j \in \mathbb{R}^d$ means the j-th feature in F, and w_j is the weight scalar. The global centroid is obtained by:

$$s_g = \frac{\sum_{i=1}^{\mu} s_i}{\mu} \tag{2}$$

Finally, the refined features F' can be obtained by:

$$F' = \bigcup_{i=1}^{\mu} F'_i, F'_i = \bigcup_{f_j \in F_i} f_j - (s_i - s_g) \tag{3}$$

The refined features are utilized in Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for pseudo-label generation. By aligning the feature distributions of different cameras, the CDC module mitigates the camera deviation and reduces false positive samples due to divergent features within the same camera. Conversely, reducing the gap between centroids does not alter the distribution within the intra-camera domain, thereby preventing distortions in identity similarity that may arise from abrupt

changes. Our method considers the influence of camera deviation in the pseudo-label assignment, which increases the accuracy of the pseudo-label.

Camera Deviation Balancing Module

To diminish the camera deviation in cross-camera feature learning, our method integrates camera constraints into a united loss. Specifically, our method proposes a multi-prototype dictionary to dynamically generate contrast pairs containing different camera information to provide customized camera constraints. Initially, an improvement to the traditional PK sampling method is introduced. The proposed Camera PK sampling strategy ensures that each minibatch contains p distinct identities, with each identity comprising k features selected from various cameras as far as possible. It means that when only a few cameras capture an individual, the intra-camera features are repeatedly chosen, resulting in a balanced sample selection. Our multi-prototype dictionary builds a library for each cluster, which denotes by $X = \{X_1, X_2, X_3, \ldots, X_m\}, X_i \in \mathbb{R}^{k \times d}$. Each X_i presents a library of the cluster, containing k prototypes, and m is the total number of clusters. The proposed Deviation Balancing Loss is performed by:

$$L_d = \frac{1}{p \times k} \sum_{q_i \in Q}^{p \times k} -log \frac{\sum_{x_\epsilon^+ \in X_+}^{k} exp(S(q_i, x_\epsilon^+ / \tau))}{\sum_{x_j \in X}^{m \times k} exp(S(q_i, x_j / \tau))} \quad (4)$$

where $q_i \in \mathbb{R}^d$ represents the i-th feature in the minibatch Q, $S(\cdot)$ is the cosine distance function. The x_ϵ^+ represents the prototype in X_+, X_+ means the library with the same pseudo-label as q_i, and x_j denotes common prototypes in the multi-prototype dictionary. After backpropagation, multi-prototype memory is updated through:

$$X_+ \leftarrow \bigcup_{\epsilon=1}^{k} \alpha \cdot x_\epsilon + (1-\alpha) q_\epsilon, x_\epsilon \in X_+, q_\epsilon \in Q_+ \quad (5)$$

where q_ϵ represents the ϵth feature in Q_+, Q_+ means features with the same identity as X_+, and α is the momentum updating factor. Our Deviation Balancing Loss utilizes samples from the prototype dictionary X to construct contrastive pairs. The camera knowledge contained in these contrastive pairs automatically adapts to the individual requirements for camera constraints, which are determined by the updating strategy and the sampling strategy. The sampling strategy ensures the inclusion of diverse camera perspectives for each identity. In cases where inter-camera diversity is insufficient, intra-camera diversity is increased. The updating strategy guarantees that the knowledge acquired from these prototypes is integrated into the memory, facilitating contrastive learning.

Cross-entropy loss is used to perform contrastive learning. Specifically, the centroid dictionary stores cluster centroids, denoted by $C = \{c_1, c_2, c_3, \ldots, c_m\}$. It is important to note that the centroid dictionary and the prototype dictionary utilize distinct memory allocations. The centroid cross-entropy loss L_c is obtained by:

$$L_c = \frac{1}{|Q|} \sum_{q_i \in Q}^{|Q|} -log \frac{exp(S(q_i, c_+)/\tau)}{\sum_{c_j \in C}^{m} (exp(S(q_i, c_j)/\tau))} \quad (6)$$

Algorithm 1: CDE-Learning

Inputs: Unlabeled dataset with camera labels
Parameters: Sampling parameter p and k
Output: The fine-tuned encoder and checkpoints
Start: Initialize Epoch parameter *num_epochs* and iterations parameter *num_iters*

```
while epoch in [1, num_epochs] do
    Extract the features F with the encoder
    Construct camera domain {F_1, F_2, F_3, ..., F_μ}
    Obtain camera centroids {s_1, s_2, s_3, ..., s_μ} by Eq.1
    Obtain the global centroid s_g by Eq.2
    Align features F_i to get refined features F_i' by Eq.3
    Clustering F' = {F_1', F_2', F_3', ..., F_μ'} into m clusters with DBSCAN
    while iter in [1, num_iters] do
        Sample p × k queries from pseudo labeled dataset
        Extract the minibatch features Q through the encoder
        Compute loss L = L_c + L_d + L_t
        Back propagation
        Update parameters of the encoder by optimizer
        Update multi-prototype memory by Eq. 5
        Update centroid memory by Eq. 7
    end while
end while
```

where $c_+ \in \mathbb{R}^d$ represents the cluster centroid with the same pseudo-label as q_i. $|Q|$ represents the total number of features in the minibatch. The centroid dictionary is updated through:

$$\forall q_i \in Q_+, c_+ \leftarrow \alpha c_+ + (1-\alpha) \cdot q_{argmax(S(q_i, c_+))} \quad (7)$$

where $q_{argmax(S(q_i, c_+))}$ is the positive sample with the farthest cosine distance, c_+ is the centroid with the same pseudo label of Q_+. This approach enhances the identification of hard samples and deeply mines identity consistency features.

Camera Attribute Auxiliary Task

Explicit camera constraints often necessitate the adjustment of hyperparameters to manage the influence of various regularization conditions, which can diminish the model's generalization performance. The proposed work introduces a camera discrimination auxiliary task aimed at enhancing the model's understanding of camera-related knowledge within feature distributions. Specifically, the auxiliary prediction module is constructed as a multi-layer MLP architecture T, producing a scalar output $e \in [0, 1]$ that indicates the probability that the current feature pair originates from the same camera. The input to T is derived by concatenating feature pairs obtained from the encoder-extracted features, as illustrated:

$$\forall q \in Q, e = T(concat(q_{2k}, q_{2k+1})), k \in \mathbb{Z} \quad (8)$$

The loss for the auxiliary task is obtained by:

$$L_t = \frac{1}{2|Q|} \sum [a \cdot log(e) + (1-a) \cdot log(1-e)], a \in \{0, 1\} \quad (9)$$

Methods	Market-1501				
	Source	mAP	R1	R5	R10
Camera-aware Unsupervised Method					
JVTC	ECCV'20	47.5	80.0	89.3	91.2
IICS	CVPR'21	72.5	89.3	95.8	97.1
CAP	AAAI'21	79.3	91.0	96.1	97.5
ICE	ICCV'21	82.2	93.4	97.2	98.3
PPLR	CVPR'22	84.2	94.5	97.5	98.2
CDR	ICCV'23	84.5	93.3	97.5	-
NPSS	TIFS'23	82.5	93.8	97.5	98.2
HCACE	TMM'24	83.4	93.7	97.5	98.1
CGC	TIFS'24	85.2	94.3	97.5	98.2
CDE	This paper	**87.4**	**94.7**	**98.2**	**98.8**
Purely Unsupervised Method					
ICE	ICCV'21	78.3	93.1	97.2	98.1
MCRN	AAAI'22	80.6	92.7	-	-
SECRET	AAAI'22	81.5	93.1	-	-
CCL	ACCV'22	83.5	93.4	97.4	98.2
ISE	CVPR'22	84.5	94.7	97.8	98.2
DCMIP	ICCV'23	86.5	94.3	98.2	98.7

Table 1: Comparison of Rank1-5(%) and mAP(%) performance with state-of-the-arts on Market-1501.

Methods	PersonX				
	source	mAP	R1	R5	R10
MMT	CVPR'20	78.9	90.6	96.8	98.2
SPCL	NIPS'20	78.5	91.1	97.8	98.0
UP	CVPR'21	81.4	93.2	96.6	98.3
CCL	ACCV'22	84.7	94.4	**98.3**	**98.4**
CDE	This paper	**86.2**	**95.1**	96.7	98.0

Table 2: Comparison of Rank1-5(%) and mAP(%) performance with state-of-the-arts on PersonX.

where a is derived from camera labels, indicating whether the current features originate from the same camera. The final loss is formulated as follows:

$$L = L_c + L_d + L_t \tag{10}$$

Experiments

Datasets and Implementation

Datasets. Our proposed method is evaluated on re-identification benchmarks, namely Market-1501 (Zheng et al. 2015), MSMT17 (Wei et al. 2018), PersonX (Sun and Zheng 2019), and CUHK03 (Li et al. 2014). Market-1501 and MSMT17 are widely employed in real-world re-identification tasks. Developed based on unity, PersonX contains 1,266 carefully designed 3D mannequins.

Implementation Details. A Resnet50 pre-trained on Imagenet is employed as the encoder of CDE-Learning. Data augmentation from the method (Dai et al. 2021) is utilized to enhance the robustness of the learning process. Adam optimizer is utilized with weight decay 5e-4 to train our re-ID model. The initial learning rate is set to 3.5e-4 for the first ten epochs with a warm-up scheme, after which it is decreased to 1/10 of its previous value every 20 epochs for 80 epochs. Our method is trained on an Nvidia A4000 GPU under the PyTorch framework.

Methods	MSMT17				
	source	mAP	R1	R5	R10
Camera-aware Unsupervised Method					
JVTC	ECCV'20	15.8	39.2	53.2	57.4
IICS	CVPR'21	18.4	44.3	58.2	63.2
CAP	AAAI'21	36.1	66.2	78.4	80.8
ICE	ICCV'21	38.5	68.2	80.4	85.3
PPLR	CVPR'22	41.2	71.8	79.7	**86.9**
CDR	ICCV'23	38.2	67.8	78.8	-
NPSS	TIFS'23	40.5	**72.1**	80.6	84.9
HCACE	TMM'24	41.6	72.0	80.8	84.9
CGC	TIFS'24	39.7	71.3	80.1	83.1
CDE	This paper	**43.4**	70.5	**80.8**	84.8
Purely Unsupervised Method					
ICE	ICCV'21	28.4	58.7	71.2	77.8
MCRN	AAAI'22	32.8	62.4	70.1	75.6
SECRET	AAAI'22	33.9	64.2	72.3	76.1
CCL	ACCV'22	32.8	61.5	72.4	76.9
ISE	CVPR'22	37.6	68.1	78.3	81.2
DCMIP	ICCV'23	40.9	68.4	75.8	82.6

Table 3: Comparison of Rank1-5(%) and mAP(%) performance with state-of-the-arts on MSMT17.

Methods	CUHK03				
	source	mAP	R1	R5	R10
TAUDL	ECCV'18	44.7	31.7	-	-
UTAL	PAMI'19	42.3	56.3	-	-
UP	CVPR'21	79.6	81.9	-	-
WSP+BDB	CVPR'22	82.3	84.7	-	-
CDE	This paper	**85.4**	**86.3**	94.1	94.5

Table 4: Comparison of Rank1-5(%) and mAP(%) performance with state-of-the-arts on CUHK03.

Comparison with State-of-the-arts

To validate the effectiveness of our method, we compare with the current state-of-the-art unsupervised learning methods, such as JVTC (Li and Zhang 2020), IICS (Xuan and Zhang 2021), CAP (Wang et al. 2021), ICE (Chen, Lagadec, and Bremond 2021), PPLR (Cho et al. 2022b), CDR (Lee et al. 2023), NPSS (Wang et al. 2023), HCACE (Luo et al. 2024), CGC (Miao et al. 2024), ICE(Chen, Lagadec, and Bremond 2021), MCRN (Wu et al. 2022), SECRET (He et al. 2022), CCL (Dai et al. 2022), ISE (Zhang et al. 2022), DCMIP (Zou et al. 2023), MMT (Ge, Chen, and Li 2020), SPCL (Ge et al. 2020), UP (Yang et al. 2021), TAUDL (Li, Zhu, and Gong 2018), UTAL (Li, Zhu, and Gong 2019), WSP+BDB (Fu et al. 2022). As presented in Table 1,3,2,4, our results indicate that the proposed method outperforms existing purely unsupervised methods. Specifically, our method achieves a 0.9%, 2.5%, 1.5% and 3.1% higher mean average precision (mAP) on Market-1501 (Zheng et al. 2015), MSMT17 (Wei et al. 2018), PersonX (Sun and Zheng 2019) and CUHK03 (Li et al. 2014) datasets when compared to the best method. In particular, our method achieves an mAP of 87.4% on Market-1501, outperforming the current state-of-the-art methods. Moreover, the proposed method also achieves an mAP of 43.4% in MSMT17.

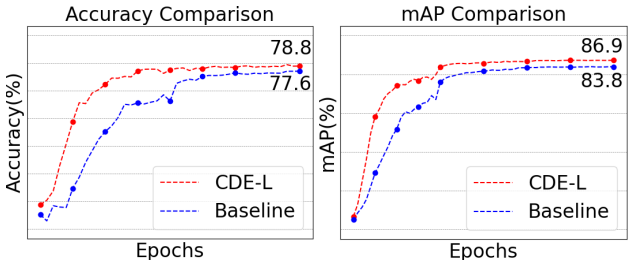

Figure 3: Comparison of pseudo-label accuracy (left) and mAP (right).

Method	Dataset					
	Market-1501			MSMT17		
	mAP	R1	R5	mAP	R1	R5
None	82.8	92.0	97.8	33.0	62.0	76.9
+CDC	86.1	94.2	98.2	40.4	67.3	80.3
+CDB	85.4	93.7	98.0	38.2	64.2	80.7
+CAA	84.7	92.3	97.0	36.5	64.0	79.1
Combo	87.4	94.7	98.2	43.4	70.5	80.8

Table 5: Ablation studies of effective components.

Compared with some camera-related methods, such as HCACE (Luo et al. 2024), and CDR (Lee et al. 2023), our method also has significant improvements. This is because, our approach capitalizes on camera labels not only for camera-constraint losses but also during the pseudo-label generation stage, thereby enhancing the accuracy of pseudo-labels. The comparison shows the method's effectiveness in capturing identity consistency and the enhancement in eliminating camera deviation. The results on multiple datasets demonstrate the excellent applicability of our method in scenes of different scales. In conclusion, our method outperforms existing purely unsupervised methods and achieves comparable performance with SOTA Re-ID methods.

Ablation Studies

Camera Deviation Compensation Module Table 5 demonstrates the significant impact of our proposed Camera Deviation Compensation module on person Re-ID performance. By employing the solely CDC module, there is a marked increase in the mAP, rising from 82.8% to 86.1%. The CDC module only acts on label allocation without participating in training, significantly enhancing the precision of pseudo-labels and the quality of prior knowledge. Empirical evidence suggests that optimizing for camera deviation at an earlier stage can yield significant benefits. An additional comparative experiment is implemented to substantiate the efficacy in enhancing pseudo-label. Figure 3 compares pseudo-label iteration between the original baseline and our method. The results illustrate that our pseudo-labeling approach surpasses the baseline in accuracy. Our method enhances pseudo-labels' quality during the initial phases, thereby circumventing the accumulation of wrong prior knowledge. As a result, it delivers superior performance in mAP within a limited number of epochs. By the tenth epoch, our method has achieved an mAP of 74.1%, significantly surpassing the baseline remaining at 51.7%.

Loss	Market-1501		
	mAP	R1	R5
$L_d + L_c + L_t$	87.4	94.7	98.2
$L_{inter} + L_c$	85.3	92.1	94.9
$L_{intra} + L_c$	86.1	93.5	94.6
$L_{inter} + L_{intra} + L_c$ (default)	86.5	95.2	97.8
$L_{inter} + L_{intra} + L_c$ (fine tuning)	86.9	94.7	98.0

Table 6: Comparison of different losses. L_{inter} and L_{intra} are traditional camera losses.

Concatenate Method	Market-1501		
	mAP	R1	R5
Random	86.8	94.3	98.2
Same Identity	87.4	94.7	98.2

Table 7: Comparison of the composition of feature pairs.

The CDC module accelerates convergence and outperforms the baseline in the mAP, indicating a significant enhancement in the ultimate performance. Our method supplements the previous approach by addressing the neglect of camera deviation during the clustering stage.

Camera Deviation Balancing Module Table 5 presents experimental outcomes solely employing the CDB module. The CDB module is designed to mitigate individual camera deviations by employing a set of prototypes across multiple camera views. The deviation balancing loss leads to a notable enhancement in the mAP, increasing from 82.8% to 85.4%. Upon implementing both the CDB and the CDC, our method achieves mAP of 87.4% on the Market-1501 dataset and 43.4% on the MSMT17 dataset, which underscores the efficacy of our approach. An extra analysis is deployed to assess the impact of the CDB module on optimizing the training process. Table 6 presents an analysis between the deviation balancing loss introduced in our method and the conventional losses L_{intra}, L_{inter} (Wang et al. 2022) employed for intra-camera and inter-camera learning. Our method consistently outperforms solely intra-camera or inter-camera learning. Furthermore, it improves by 0.5% in mAP compared with fine-tuning intra-camera and inter-camera learning, highlighting the advantage of our approach without the need for parameter adjustments.

Camera Attribute Auxiliary Task Table 5 demonstrates the performance improvements of our auxiliary task on both the Market-1501 and MSMT17 datasets. Specifically, the mAP on the Market-1501 dataset increased by 1.9% compared to the baseline, while on the MSMT17 dataset, it improved by 3.5%. During the training process, the effects of different concatenation strategies are also compared. The first strategy involves completely random concatenation, where a feature pair could either have the same pseudo-label or different ones. As shown in Table 7, this random approach negatively impacts the model's ability to recognize identity features. Conversely, the feature pairs concatenated on the same pseudo-label enhance the model's capability to extract pedestrian information.

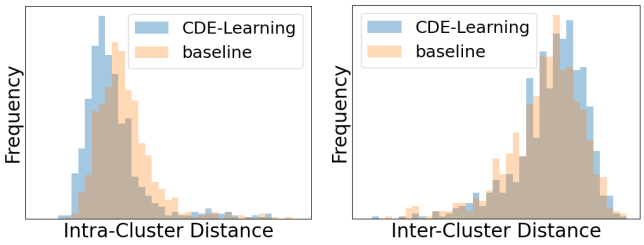

Figure 4: Comparison of intra-cluster and inter-cluster similarity between CDE-Learning and baseline.

τ	Market-1501			MSMT17		
	mAP	R1	R5	mAP	R1	R5
0.01	83.9	86.1	93.7	32.6	55.3	67.4
0.03	86.5	93.2	96.3	40.1	65.1	80.1
0.05	87.4	94.7	98.8	43.4	71.5	80.8
0.08	86.1	92.8	97.3	39.2	64.8	80.6
0.1	84.8	90.4	95.4	35.1	58.5	68.7
0.3	82.3	85.6	94.1	33.3	57.3	67.3

Table 8: Experimental results about the temperature parameter τ.

Parameters Analysis

Table 8 presents a detailed examination of the outcomes associated with the temperature parameter τ in our method. The results indicate that this parameter's sensitivity is crucial in distinguishing between identities. Notably, our method attains its best performance at the $\tau = 0.05$, achieving 87.4% in the mAP. In the range of $[0.03, 0.08]$, our method exhibits a degree of robustness to changes in the temperature parameter. Table 9 reflects the effects of PK sampling on the Re-ID outcomes. Our method yields the most favorable results at the parameter setting of $(16, 16)$, indicating an optimal balance for minibatch composition. When the total image count is constrained to 128, the performance with $k = 16$ is nearly equivalent to that with $k = 8$. This observation suggests that increasing the value of k does not necessarily lead to better outcomes. Excess features can not enhance task performance significantly.

Visualization

Figure 6 displays the GradCAM visualizations from our experiments, allowing for examining the image regions that the model emphasizes.

(p, k)	Market-1501			MSMT17		
	mAP	R1	R5	mAP	R1	R5
(8,8)	83.3	92.7	95.4	35.6	62.1	78.9
(8,16)	85.5	93.3	96.6	39.7	66.5	80.0
(16,8)	85.1	93.6	96.4	40.6	65.3	80.5
(16,16)	87.4	94.7	98.2	43.4	70.6	80.8

Table 9: The experimental results about PK sampling parameters. The input parameters for PK sampling are denoted as (p, k), where each minibatch comprises p identities, with each identity represented by k images.

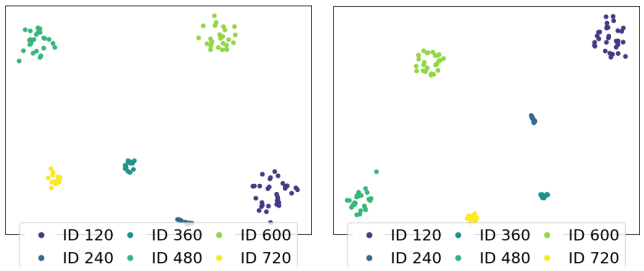

Figure 5: Visualizing feature distributions for baseline (left) and CDE-Learning (right). The feature distributions are compared by t-SNE.

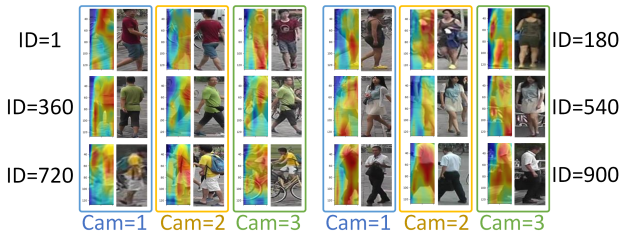

Figure 6: GradCAM visualizations of CDE-Learning.

The visualizations indicate that our method successfully focuses on relevant identity features. As depicted in Figure 5, for features corresponding to the same real ID, the distribution of our method is noticeably more clustered, indicating a more remarkable similarity among features. Figure 4 presents a probability distribution comparing CDE-Learning with the baseline method, providing a statistical analysis of the average inter-cluster and intra-cluster distances under true identities.

Conclusions

In conclusion, the proposed CDE-Learning reduces the camera deviation in unsupervised Re-ID. Specifically, our CDC module optimizes camera deviation at the pseudo-label assignment stage. Besides that, the proposed CDB module effectively balances camera deviation, avoiding redundant camera information from individual instances. Moreover, the CAA task enables the model to simultaneously focus on distinguishing background information during training. Our method has been rigorously tested on general-purpose datasets, with the experimental outcomes substantiating its efficacy.

Acknowledgments

Central Government Guides Local Science and Technology Development Fund Projects (236Z0301G); Basic Research Project of Shijiazhuang Municipal Universities in Hebei Province (241791387A); National Key Research and Development Program of China Grant (2024YFB4710800), Liaoning Provincial Natural Science Foundation Grant (2024-MS-012).

References

Chen, H.; Lagadec, B.; and Bremond, F. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

Chen, H.; Wang, Y.; Lagadec, B.; Dantcheva, A.; and Bremond, F. 2023. Learning Invariance From Generated Variance for Unsupervised Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7494–7508.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. Moco: Momentum Contrast for Unsupervised Visual Representation Learning. In *International Conference on Learning Representations (ICLR)*.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022a. Part-based Pseudo Label Refinement for Unsupervised Person Re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7298–7308.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022b. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7318.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2021. Cluster Contrast for Unsupervised Person Re-Identification. *arXiv preprint arXiv:2103.11568*.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian conference on computer vision*, 1142–1160.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fu, D.; Chen, D.; Yang, H.; Bao, J.; Yuan, L.; Zhang, L.; Li, H.; Wen, F.; and Chen, D. 2022. Large-scale pre-training for person re-identification with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2476–2486.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33: 11309–11321.

He, S.; Chen, W.; Wang, K.; Luo, H.; Wang, F.; Jiang, W.; and Ding, H. 2024. Region Generation and Assessment Network for Occluded Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 19: 120–132.

He, T.; Shen, L.; Guo, Y.; Ding, G.; and Guo, Z. 2022. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 879–887.

Houlsby, N.; Dosovitskiy, A.; Kipf, T.; Verbeek, J.; Unterhaltung, T.; Ilse, M.; Wierstra, D.; Abbeel, P.; and Fox, D. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *International Conference on Learning Representations (ICLR)*.

Huang, W.; Yi, M.; and Zhao, X. 2023. Towards the Generalization of Contrastive Self-Supervised Learning. *ICLR 2023*, abs/2111.00743.

Huang, Y.; Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; and Wang, L. 2024. Enhancing Person Re-Identification Performance Through In Vivo Learning. *IEEE Transactions on Image Processing*, 33: 639–654.

Isobe, T.; Li, D.; Tian, L.; Chen, W.; Shan, Y.; and Wang, S. 2021. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8526–8536.

Ji, H.; Wang, L.; Zhou, S.; Tang, W.; Zheng, N.; and Hua, G. 2021. Meta pairwise relationship distillation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3661–3670.

Lee, G.; Lee, S.; Kim, D.; Shin, Y.; Yoon, Y.; and Ham, B. 2023. Camera-Driven Representation Learning for Unsupervised Domain Adaptive Person Re-identification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11419–11428.

Li; and Qi. 2023. Unsupervised Pedestrian Re-Identification with Multi-Label Learning Guided Self-Paced Clustering. *Pattern Recognition*.

Li, J.; and Zhang, S. 2020. Joint Visual and Temporal Consistency for Unsupervised Domain Adaptive Person Re-Identification. arXiv:2007.10854.

Li, M.; Li, C.-G.; and Guo, J. 2022. Cluster-Guided Asymmetric Contrastive Learning for Unsupervised Person Re-Identification. *IEEE Transactions on Image Processing*, 31: 3606–3617.

Li, M.; Zhu, X.; and Gong, S. 2018. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, 737–753.

Li, M.; Zhu, X.; and Gong, S. 2019. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42: 1770–1782.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 152–159.

Liu, Y.; Zhou, W.; Xie, Q.; and Li, H. 2023. Unsupervised Person Re-Identification With Wireless Positioning Under Weak Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5282–5295.

Luo, X.; Jiang, M.; Kong, J.; and Tao, X. 2024. Hierarchical camera-aware contrast extension for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 26: 7636–7648.

Miao, Y.; Deng, J.; Ding, G.; and Han, J. 2024. Confidence-Guided Centroids for Unsupervised Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 19: 6471–6483.

Peng, J.; Jiang, G.; and Wang, H. 2023. Adaptive Memorization with Group Labels for Unsupervised Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Peng, J.; Yu, J.; Wang, C.; Wang, H.; and Fu, X. 2024. Adapt only once: Fast unsupervised person re-identification via relevance-aware guidance. *Pattern Recognition*, 150: 110360.

Sun, X.; and Zheng, L. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 608–617.

Wang, D.; and Zhang, S. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10981–10990.

Wang, H.; Wang, Y.; Zhang, Z.; Fu, X.; Zhuo, L.; Xu, M.; and Wang, M. 2020. Kernelized multiview subspace analysis by self-weighted learning. *IEEE Transactions on Multimedia*, 23: 3828–3840.

Wang, H.; Yang, M.; Liu, J.; and Zheng, W.-S. 2023. Pseudo-Label Noise Prevention, Suppression and Softening for Unsupervised Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 18: 3222–3237.

Wang, H.; Yao, M.; Chen, Y.; Xu, Y.; Liu, H.; Jia, W.; Fu, X.; and Wang, Y. 2024. Manifold-based Incomplete Multi-view Clustering via Bi-Consistency Guidance. *IEEE Transactions on Multimedia*.

Wang, M.; Lai, B.; Huang, J.; Gong, X.; and Hua, X.-S. 2021. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2764–2772.

Wang, M.; Li, J.; Lai, B.; Gong, X.; and Hua, X.-S. 2022. Offline-Online Associated Camera-Aware Proxies for Unsupervised Person Re-Identification. *IEEE Transactions on Image Processing*, 31: 6548–6561.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 79–88.

Wu, A.; Zheng, W.-S.; and Lai, J.-H. 2019. Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6921–6930.

Wu, Y.; Huang, T.; Yao, H.; Zhang, C.; Shao, Y.; Han, C.; Gao, C.; and Sang, N. 2022. Multi-centroid representation network for domain adaptive person re-id. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2750–2758.

Xuan, S.; and 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*, 11926–11935.

Yang, Z.; Jin, X.; Zheng, K.; and Zhao, F. 2021. Unleashing the Potential of Unsupervised Pre-Training with Intra-Identity Regularization for Person Re-Identification. arXiv:2112.00317.

Yao, M.; Wang, H.; Chen, Y.; and Fu, X. 2024. Between/Within View Information Completing for Tensorial Incomplete Multi-view Clustering. *IEEE Transactions on Multimedia*.

Zhang, G.; Zhang, H.; Lin, W.; Chandran, A. K.; and Jing, X. 2023. Camera Contrast Learning for Unsupervised Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, X.; Ge, Y.; Qiao, Y.; and 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*, 3436–3445.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J. Q.; Zhang, Z.; and Wang, J. 2022. Implicit Sample Extension for Unsupervised Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 7359–7368. IEEE.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zou, C.; Chen, Z.; Cui, Z.; Liu, Y.; and Zhang, C. 2023. Discrepant and Multi-instance Proxies for Unsupervised Person Re-identification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11024–11034.