Staged Encoder Training for Cross-Camera Person Re-Identification

zhi Xu (xuzhi@guet.edu.cn)
School of Computer Information and Security, Guilin University of Electronic Technology
Jiawei Yang
School of Computer Information and Security, Guilin University of Electronic Technology
Yuxuan Liu
School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology
Longyang Zhao
School of Computer Information and Security, Guilin University of Electronic Technology
Jiajia Liu
School of Institude of Electronic and Electrical Engineering, Civil Aviation Flight University of China

Research Article

Keywords: Camera variation, Contrastive learning, Unsupervised, Person Re-identification

Posted Date: November 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3511084/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on February 16th, 2024. See the published version at https://doi.org/10.1007/s11760-023-02909-0.

Staged Encoder Training for Cross-Camera Person Re-Identification

Zhi Xu[1] · **Jiawei Yang**[1] · **Yuxuan Liu**[2] · **Longyang Zhao**[1] · **Jiajia Liu**[3]

Received: date / Accepted: date

Abstract As a cross-camera retrieval problem, person Re-identification (ReID) suffers from image style variations casued by camera parameters, lighting and other reasons, which will seriously affect the model recognition accuracy. To address this problem, this paper proposes a two-stage contrastive learning method to gradually reduce the impact of camera variations. In the first stage, we train an encoder for each camera using only images from the respective camera. This ensures that each encoder has better recognition performance on images from its respective camera while being unaffected by camera variations. In the second stage, we encode the same image using all trained encoders to generate a new combination code that is robust against camera variations. We also use Cross-Camera Encouragement [12] distance that complements the advantages of combined encoding to further mitigate the impact of camera variations. Our method achieves high accuracy on several commonly used person ReID datasets, e.g., achieces 90.8% rank-1 accuracy and 85.2% mAP on the Market1501, outperforming the recent unsupervised works by 12+%. Code is available at https://github.com/yjwyuanwu/SET.

Keywords Camera variation · Contrastive learning · Unsupervised · Person Re-identification

1 Introduction

Given a query image, person Re-identification(ReID) aims to match the person across multiple non-overlapping cameras [12,20]. In ReID scenarios, each identity may be recorded by multiple cameras with different parameters and environments, these factors change the appearance of the image, making it challenging to recognize the same identities.

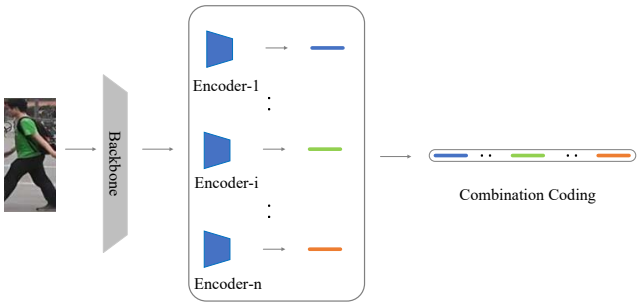

Fig. 1 An illustration: the generation of combination coding in the inter-contrast learning stage using encoders trained in the intra-contrast learning stage

In previous studies, researchers have addressed the above challenges through supervised methods, mainly focusing on finding appropriate mapping functions based on the data distribution of images captured by different cameras [14,9]. However, such approaches require annotated training samples to learn the camera transfer model and are only applicable to small datasets. In recent years, researchers have focused on studying unsupervised domain adaptation (UDA) methods [3,18,28,10,5,23] and purely unsupervised methods [11,17,19,27,1] to address this issue. UDA is complex to train and requires that the difference between the source and target domains is not significant. In this paper, we focus on the fully unsupervised approach, which uses only unlabeled data in the target domain and is trained using the generated pseudo-labels.

In research on fully unsupervised methods, it is common to use data augmentation to make the model robust to camera variations [27]. Alternatively, in the training step, sam-

⋆ corresponding author: Zhi Xu e-mail: xuzhi@guet.edu.cn
[1] School of Computer Information and Security, Guilin University of Electronic Technology, China.
[2] School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, China.
[3] School of Institude of Electronic and Electrical Engineering, Civil Aviation Flight University of China.

ples are clustered and pseudo-labelled, and then a model is designed to extract features that are robust to camera variations [11, 17, 19, 1]. Unlike previous methods, this paper focuses on the pseudo-label prediction step in the fully unsupervised setting. Most pseudo-label prediction algorithms follow a similar process, which includes feature extraction, similarity computation, and assigning the same label to similar samples for training. The feature similarity calculation is a crucial step in this process. However, camera variations lead to an increase in the inter-class distance for the same identity, which significantly affects the reliability of the similarity results.

In this paper, we address the above issues by investigating a more reasonable distance computation for generating pseudo-labels. Since it is easier to identify pedestrians with the same identity in the same camera than in different cameras, as shown in Fig. 2, we decompose the distance calculation between sample encodings into two stages, gradually searching for reliable pseudo-labels. These stages are trained alternately to jointly optimize the backbone network. In the first stage, i.e., intra-contrast learning stage, multiple branches are trained together, with branch k using samples from camera k for training. Since the samples of each branch come from a single camera and are not affected by camera variations, the similarity computation in this stage is performed directly using the encodings obtained from the backbone network and the encoder. The contrast learning method used for training is discussed in detail in Sec. 3.2.

In the second stage, i.e., inter-contrast learning stage, we use all samples in the training set to jointly train an additional encoder. Since the samples in the training set come from different cameras, we must take camera variations into account during this stage. Inspired by studies such as [19, 4], which show that the classification probability is more robust to the domain gap than raw features, we consider the feature obtained from the backbone as "raw feature". As shown in Fig. 1, the encoders trained in the first stage for each camera are used to obtain the combined encoding of the samples as "classification". Furthermore, to avoid misidentifying samples from different identities as the same identity when their combined encodings are close, we further explicitly reduced the sample distance between different cameras using the Cross-Camera Encouragement [12]. The distance between sample encodings in the second stage is composed of the original encoding distance ($\boldsymbol{d_1}$), the combined encoding distance ($\boldsymbol{d_2}$), and the Cross-Camera Encouragement distance ($\boldsymbol{d_3}$). We also employed contrastive learning for training in this stage. d_2 and d_3 will be introduced in Sec. 3.5.

The proposed method decomposes the distance calculation between sample encodings into two stages, and gradually finds reliable pseudo-labels. This method is more reliable than directly predicting pseudo-labels across cameras in that, and effectively alleviates the impact of camera variations. Code is available at https://github.com/yjwyuanwu/SET

Our contributions can be summarized as follows:

- We propose a two-stage comparative learning framework to optimise the image coding extraction process, where the two stages mutually promote each other's performance.
- The proposed method for similarity computation effectively alleviates the challenge of camera variations, in which d_2 and d_3 have complementary advantages.
- At the stage of pseudo-labelling, we present a method for reprocessing pseudo-labels to address the issue of over-labeling.
- Our method achieves high accuracy on three commonly used person re-identification datasets. It provides insights into improved similarity calculation for fully unsupervised person ReID.

2 Related work

The proposed method is inspired by domain adaptation methods and effectively mitigates the impact of camera variations in a fully unsupervised setting. The work on these two topics will be introduced in the following two subsections.

2.1 Domain adaptation

Domain adaptation can be summarized into three categories: GAN-based style transfer, finding features that are robust to camera varitions, and mutual training. Zhong et al.[26] proposed a triplet training sample construction method using style transfer and non-overlapping person ReID datasets. Wei et al. [18] introduced a GAN-based approach that transfers task images to match the style of the target domain dataset while preserving the label information from the source domain. For research on finding robust features, Zheng et al. [25] proposed a method to separate features into appearance and structural features, and Zou et al. [28] explored domain adaptation using appearance features as domain-invariant features. There are also studies [19,4] showing that the classification probability is more robust to the domain gap than raw features, and our work was inspired by this research result. Other methods, such as MMT [5] and NRMT [23], focus on reducing the impact of low-quality pseudo-labels through mutual training [22] to improve the model's recognition accuracy.

2.2 Fully unsupervised person ReID

Fully unsupervised methods related to mitigating camera variations mainly focus on three aspects: data augmentation, extracting features that are robust to camera variations,

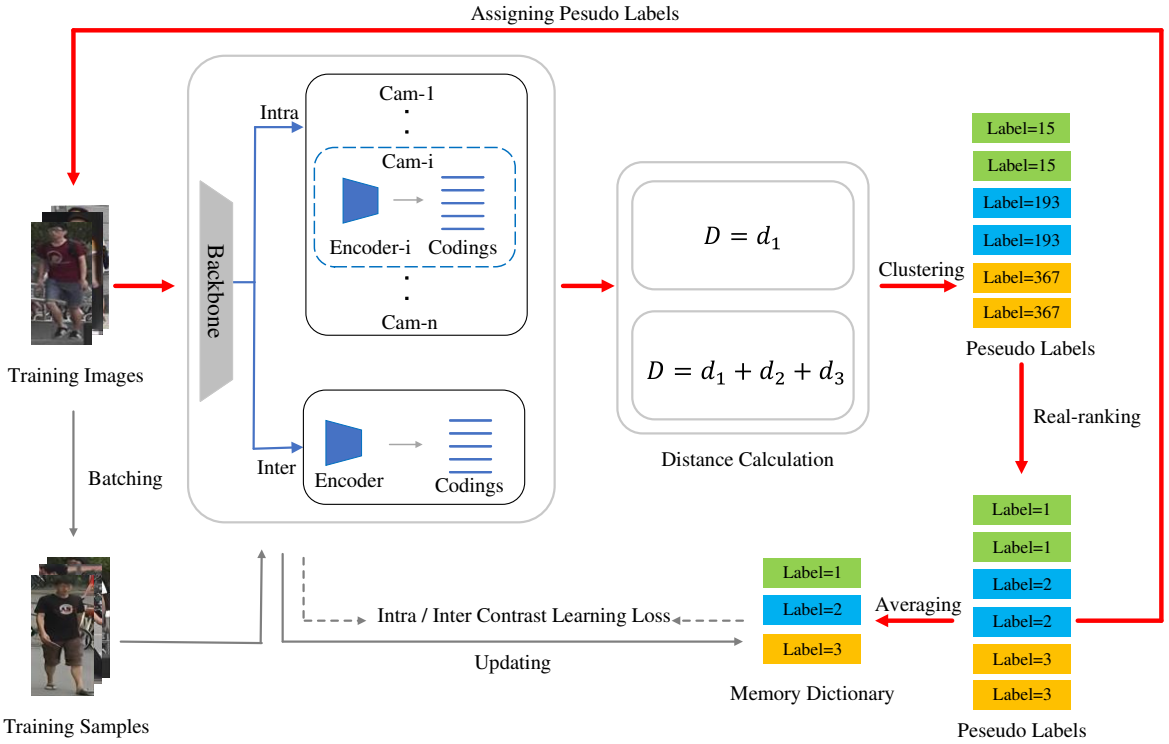

Fig. 2 overall flowchart .The whole training process is divided into two parts, intra contrast learning and inter contrast learning, which share the backbone structure, and are represented by the upper and lower parts in the box, respectively. Both parts undergo two stages of training sequentially: (1)Initialization Stage (indicated by the red line): The clustering results of image encodings are used for dictionary feature initialization and pseudo-label initialization of the samples. (2)Training Stage (indicated by the thin arrow line): The thin solid arrow line updates features in the dictionary, while the thin dashed arrow line calculates the loss for the current stage and updates the backbone and encoder. During testing, we encode the images using the backbone and encoder from the inter-contrast learning stage. We compute the Euclidean distance between the encodings to obtain the final query results

and generating reliable pseudo-labels. Zhong et al. [27] proposed a method to improve model accuracy through data enhancement and using label smoothing regularization (LSR) loss. Chen et al. [1] extracted features from the statistical information of different camera images and performed feature fusion to generate cross-camera invariant features. For research on generating reliable pseudo-labels, Lin et al. [11] considered each image as an individual sample and gradually grouped them based on sample similarity. Wang et al. [17] formulated ReID as a multi-classification problem and employed optimized similarity computation to enhance the accuracy of pseudo-label prediction. The work most similar to our study is [19], which produces feature vectors that withstand differences in cameras by utilizing classification outcomes from various camera classifiers to mitigate the camera disparity issue. In contrast, we use image encoding to produce camera-robust composite encodings directly. Additionally, we use d_3 (the Cross-Camera Encouragement distance) to compensate for the shortcomings of d_2 (the combined encoding distance) and improve the model's optimization by using a memory dictionary rather than a classifier, resulting in a better reduction of intra-class distance of the samples whilst expanding the inter-class distance. our method is proven to be more effective on multiple datasets.

3 Proposed Method

3.1 Formulation

Given an unlabelled dataset χ, we can consider it to consist of multiple subdatasets, denoted as $\chi = \{\chi^c\}$, $c = 1 : C$, where the superscript c indicates that all of the images in this subdataset are from camera c and C represents the total number of cameras. Our task is to train a model on χ, such that for each query image q, this ReID model generates a feature encoding to retrieve the pedestrian images in gallery set G that contain the same identity. in other words, the feature encoding of q should have a smaller distance to the encoding of a gallery image g with the same identity as q compared to the distances to other images in G. The task can be defined as follows:

$$g^* = \arg\min_{g \in G} dist(r_g, r_q) \tag{1}$$

where r represents the image encoding extracted by the ReID model, and $dist(\cdot)$ is the distance metric.

This paper is to generate more accurate pseudo-labels by reducing the effect of camera variations on the sample distance calculation during training, so as to guide the model training and enable the model to extract encodings that satisfy Eq. 1. The model comprises of two stages. As shown in Fig. 2, the first stage, intra-contrast learning stage, uses multiple branches for joint training, each branch uses only a sub-dataset for training, and the loss of branch c can be expressed as the sum of the contrast loss of all samples from camera c:

$$L^c_{\text{intra}} = \sum_{\mathrm{I}\in\chi^c, \mathrm{I}\in H^c_m} L^c_{contrast}(f,m) \tag{2}$$

where f represents the feature of image I after extraction by the backbone, m is the corresponding pseudo-label, and H represents the set of pseudo-labels generated by clustering.

In the second stage, inter-contrast learning, we share the parameters of the first stage backbone and train an additional encoder. This stage uses the whole training set for training, including images from different cameras. To minimize the impact of camera variations, we propose a combination coding that is more robust to the camera variations. As shown in Fig. 1, we use all encoders trained in the first stage to encode the images separately, and then use these encodings to generate the combination coding R. The combination coding R_i for image x_i can be denoted as:

$$R_i = \left[r_i^1, \cdots, r_i^k, \cdots, r_i^C\right] \tag{3}$$

where r_i^k is the coding of the image x_i obtained by the encoder corresponding to the k-th camera. We use d_2 (the combination coding distance) and d_3 (the Cross-Camera Encouragement distance [12]) to reduce the impact of camera variations. The distance between any two images x_i and x_j in the inter-contrast learning stage is represented as follows:

$$D(x_i,x_j) = d_1(x_i,x_j) + \mu d_2(x_i,x_j) + d_3(x_i,x_j) \tag{4}$$

where $d_1(\cdot)$ represents the Euclidean distance of the image coding. We use the clustering result H to calculate the loss in the inter-contrast learning phase to optimise the extraction of the coding r, i.e.,

$$L_{\text{inter}} = \sum_{\mathrm{I}\in H_m} L_{contrast}(r,m) \tag{5}$$

In summary, these two stages share the backbone network while having their own encoders with the same structure, and the two stages are trained alternately. The $d_2(\cdot)$ and $d_3(\cdot)$ mentioned above are explained in detail in Sec. 3.5.

3.2 Contrast learning

Both stages of the model are trained using contrast learning, including the initialization and training stages. In the initialization stage, samples are passed through the backbone network and the encoder to obtain sample encodings. Then, similarity is computed to perform clustering by assigning the same pseudo-label to samples belonging to the same cluster. After real-ranking which will be described in Sec. 3.3, the average encoding of samples with the same pseudo-label is used to initialize the memory dictionary, each of the cluster centroids stored in the memory dictionary can be represented as:

$$\phi_k = \frac{1}{|H_k|} \sum_{\phi_i \in H_k} \phi_i \tag{6}$$

where H_k represents the set of sample encodings for the k-th cluster. During the training process, the cluster centroids encoding ϕ_k in the memory dictionary is updated with the sample encoding r using Eq. 7:

$$\phi_k \leftarrow \lambda\, \phi_k + (1-\lambda) r \tag{7}$$

where $\lambda \in [0,1)$ represents the momentum update factor. λ controls the consistency between the sample coding r and the corresponding clustering mean. When λ approaches 0, the clustering mean ϕ_k is closest to the coding r of the latest training sample sample. The loss of contrast for a sample coded as r and with a pseudo-label of m can be expressed as:

$$L_{contrast}(r,m) = -\log \frac{\exp(r\cdot\phi_m/\tau)}{\sum_{k=0}^{K}\exp(r\cdot\phi_k/\tau)} \tag{8}$$

where τ is a temperature hyperparameter, $\{\phi_1,\phi_2,...,\phi_K\}$ represents the cluster centroids stored in the memory dictionary, and K represents the number of clusters. The contrast loss can reduce the intra-class distance while increasing the inter-class distance, which can improve the discriminative ability of the model. The loss of all obtained samples is used to update the backbone and encoder.

3.3 Real-ranking

We use top-down hierarchical clustering method for clustering, requiring the number of clusters M to be specified at the outset. When the number of samples is small, the resulting number of clusters K may be fewer than the specified quantity. Nevertheless, the allocation of cluster labels m is randomly assigned by the clustering algorithm within a number less than M, i.e., it may produce the problem of over-labelling:

$$m = random(M) > K \tag{9}$$

Since the cluster means in the memory dictionary are stored in order of label, when an over-label problem occurs, the cluster centroid corresponding to label m that exceeds the actual number of clusters cannot be found in the memory dictionary, which leads to the inability to compute the loss by using Eq. 8. To solve this problem, as shown in Fig. 2, we propose a method called Real-ranking to redistribute the pseudo-labels by ranking them after clustering. The ranking position of the given sample's pseudo-label is then used as its final pseudo-label, guaranteeing that no pseudo-label exceeds the actual number of classifications.

3.4 Intra-contrast learning

As illustrated in Fig. 2, we employ multiple branches for joint training in the intra-contrast learning stage. According to Eq. 8, we can derive the contrastive loss for the branch c mentioned in Sec. 3.1 as follows:

$$\begin{aligned} L^c_{contrast}(f,m) &= L^c_{contrast}(E(\theta_c,f),m) \\ &= -\log \frac{\exp(E(\theta_c,f)\cdot \phi_m/\tau)}{\sum_{k=0}^{K}\exp(E(\theta_c,f)\cdot \phi_k/\tau)} \end{aligned} \tag{10}$$

where $E(\theta_c,\cdot)$ represents the encoder with parameter θ_c. The loss of the intra-contrast learning stage is equal to the sum of the losses of all branches in this stage and can be formulated as:

$$L_{intra} = \sum_{c=1}^{C} L^c_{intra} \tag{11}$$

Eq. 11 effectively improves the discriminative ability of the encodings extracted by each camera encoder. In addition, the optimization of multiple branches also improves the discriminative ability of the model for images from different cameras.

3.5 Inter-contrast learning

In the inter-contrast learning stage, the encoding distance between samples is determined using Eq. 4. Due to camera variations, the encoding distance between different samples of the same identity tends to increase. Therefore, we subtract d_2 from the encoding distance of samples from distinct cameras during the encoding distance calculation. d_2 can be calculated as follows:

$$d_2(x_i,x_j) = \begin{cases} 0, & c_i = c_j \\ -J(R_i,R_j), & c_i \neq c_j \end{cases} \tag{12}$$

where $J(\cdot)$ represents the Jaccard distance, the Jaccard distance between two samples is smaller when their combination coding is more similar. The corresponding Jaccard distance of the combination coding is calculated as:

$$J(R_i,R_j) = 1 - \frac{R_i \cap R_j}{R_i \cup R_j} \tag{13}$$

where $\cap$ indicates that the combination coding R takes a smaller value at the corresponding location, and $\cup$ indicates that it takes a larger value. In order to prevent samples with different identities from having similar combination codings leading to them being mistakenly recognised as the same identity, we use d_3 to further reduce the effect of camera variations, and the d_3 distance can be denoted as:

$$d_3(x_i,x_j) = \begin{cases} \lambda_c, & c_i = c_j \\ 0, & c_i \neq c_j \end{cases} \tag{14}$$

4 Experiment

4.1 dataset and Evaluation Protocols

We evaluated our method on three widely-used person ReID datasets, including Market-1501 [24], PersonX [16] and DukeMTMC-ReID [15]. The details of these three datasets are summarized in Table 1. During training, we only utilized the images and camera information from the training sets of each dataset, without using any other annotation information. Note that the camera ID is automatically obtained at the moment of capturing and is no need for human labeling. Performance is evaluated by the Cumulative Matching Characteristic (CMC) and meanAverage Precision (mAP).

4.2 Implementation details

To ensure a fair comparison with other methods, we used a pre-trained ResNet50 [7] on ImageNet [2] as the backbone network for feature extraction. After layer 5, we removed all submodule layers and added a batch normalisation layer [8], which will produce 2048 dimensional coding using the combination of these two layers as the encoder. During testing and clustering, we calculated the similarity between samples using the encodings obtained after passing through the backbone and the encoder.

During training, the input images are resized to 256×128. In each round, we perform intra-contrast learning and inter-contrast learning in sequence. The training consists of 50 rounds. We use the Adam optimizer to train both stages of the re-ID model with weight decay of 0.0005. The initial learning rate $lr = 0.00035$ and then decays to $1/10$ of the previous every 20 rounds. The momentum update factor $\lambda = 0.99$. Every mini-batch integrates 256 images of 16 fake person identities (16 images per identity).

Dataset	# train IDs	# train images	# test IDs	# query images	# total images	# cameras
Market-1501	751	12,936	750	3,368	32,668	6
PersonX	410	9,840	856	5,136	45,792	6
DukeMTMC-ReID	702	16,522	702	2,228	36,441	8

Table 1 Statistics of datasets used in the experimental section

Methods	Market1501					DukeMTMC-ReID				
	source	mAP	Rank-1	Rank-5	Rank-10	source	mAP	Rank-1	Rank-5	Rank-10
SPGAN[3]	Duke	26.9	58.1	76.0	82.7	Market	26.4	46.6	62.6	68.5
HHL[26]	Duke	31.4	62.2	78.8	84.0	Market	27.2	46.9	61.0	66.7
DGNet++[28]	Duke	61.7	82.1	90.2	92.7	Market	61.8	78.9	87.8	90.4
PDA-Net[10]	Duke	47.6	75.2	86.3	90.2	Market	45.1	63.2	77.0	82.5
NRMT[23]	Duke	71.7	87.8	94.6	96.5	Market	62.2	77.8	86.9	89.5
MMT[5]	Duke	71.2	87.7	94.9	96.9	Market	63.1	76.8	**88.0**	**92.2**
BUC[11]	None	38.3	66.2	79.6	84.5	None	22.1	40.4	52.5	58.2
HCT[21]	None	56.4	80.0	91.6	95.2	None	50.7	69.6	83.4	87.4
MMCL[17]	None	45.5	80.3	89.4	92.3	None	40.2	65.2	75.9	80.0
IICS[19]	None	72.9	89.5	**95.2**	**97.0**	None	64.4	80.0	89.0	91.6
Ours	None	**85.2**	**90.8**	94.4	95.8	None	**71.1**	**80.2**	85.9	88.7

Table 2 Experiments on Market-1501 and DukeMTMC-ReID datasets. The comparison with recent person ReID methods, including domain adaptation methods and fully unsupervised methods, where "None" represents the fully unsupervised method and other values represent the source domain datasets in domain adaptive methods. The black bold font represents the optimal value of each metric

Methods	PersonX				
	source	mAP	Rank-1	Rank-5	Rank-10
MMT[5]	Market	78.9	90.6	96.8	98.2
SPCL[6]	None	72.3	88.1	96.6	98.3
Ours	None	**91.8**	**94.5**	**97.6**	**98.6**

Table 3 Experiments on PersonX datasets. Where "None" represents the fully unsupervised method and other values represent the source domain datasets in domain adaptive methods. The black bold font represents the optimal value of each metric

For every round of training, we train the model for two epochs at both stages. We use the standard hierarchical clustering method [13], as done in [19], we set the number of clusters for each camera to be 600 in the intra-contrast learning stage and 800 in the inter-contrast learning stage.

4.3 Comparison with State-of-the-arts

We compare recent fully unsupervised methods and domain adaptation methods on Market-1501 [24] , PersonX [16], and DukeMTMC-ReID [15]. The results of the comparison are summarised in Table 2 and Table 3. First we compare domain adaptive methods, including methods that perform style transfer via GAN (SPGAN [3],et al.), methods that reduce the effect of domain gap by disentangling features (DGNet++ [28],et al.) and methods that reduce the effect of low-quality pseudo-labelling by mutual training (NRMT [23],et al.).

These domain adaptation techniques depend on manually annotated labels from the source domain, whereas our methodology achieves better results even without such reliance. We also compared our method with some fully unsupervised methods (BUC [11],*et al*), and it is clear that our approach outperformed most of these methods based on various metrics relying on the more reliable calculation of the sample encoding distances used in the clustering process.

4.4 Ablation Studies

The impact of individual components. In this section we evaluate the effectiveness of the two stages of intra-contrast learning and inter-contrast learning in our method. The experimental results are summarised in Table 5. As shown in the table, relying solely on inter-contrast learning for training leads to poor performance, indicating that the distance calculations between samples from different cameras are unreliable. On the other hand, when only intra-contrast learning is used, the rank-1 accuracy on the Market-1501 and PersonX datasets can reach 86.9% and 93.0% respectively. This shows that the distance calculation of the sample coding is more accurate when it is not influenced by camera variations. However, without considering the distribution gap between the cameras, the addition of the inter-contrast learning stage results in a decrease in performance on PersonX. This shows that although the sample coding produced by the model improves after the intra-contrast learning stage, the calculation of distances between samples from different

Dateset	Market-1501		PersonX	
Settings	mAP	Rank-1	mAP	Rank-1
d_1	83.1	87.4	88.0	92.5
$d_1 + \mu d_2$	84.6	90.1	91.0	93.6
$d_1 + d_3$	84.1	89.7	90.0	92.9
$d_1 + \mu d_2 + d_3$	85.2	90.8	91.8	94.5

Table 4 Investigate the effect on the results of using different parts of Eq. 4 in stage 2

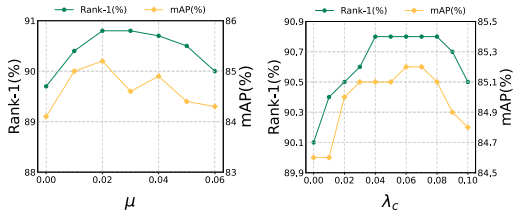

Fig. 3 Parameter analysis on Market-1501

cameras remains unreliable. When we use Eq. 4 to calculate the sample coding distance in the inter-contrast learning stage, there is a significant improvement in accuracy, demonstrating that our proposed distance calculation method successfully mitigates the effects of camera variances on sample distance calculations.

The impact of different partial distances. In this section, we investigate the effectiveness of the d_2 and d_3 distances in Eq. 4. The experimental results are summarised in Table 4. Taking the experimental results on the Market1501 dataset as an example, when we use d_1 directly to calculate the sample coding distance, the rank-1 accuracy is only 87.4%. However, when we use d_2 or d_3 for distance calculation while using d for sample encoding distance calculation, the rank-1 accuracy improves to 90.1% and 89.7%, respectively, indicating that both can reduce the effect of camera variations on the distance calculation. Furthermore, when we calculate the sample encoding distance using d_1, d_2, and d_3 simultaneously, the rank-1 accuracy further improves to 90.8%. This suggests that d_2 and d_3 can improve accuracy individually, and their advantages complement each other. d_2 compensates for d_3's shortcoming of treating all inter-camera variations as equal, while d_3 can explicitly reduce inter-sample coding distances, compensating for d_2's shortcoming of discriminating pedestrians with different identities whose combination codings are close to each other as the same identity.

Influence of hyper-parameters. In this section, we investigate the effect of two important hyperparameters μ and λ_c, as shown in Fig. 3. The parameter μ is used to regulate the importance of d_2. By increasing μ from 0 to 0.02, we observe an increase in both mAP and Rank1. However, further raising μ leads to a decline in mAP and rank-1 to varying extents. Therefore, we select μ as 0.02.

Dateset	Market-1501		PersonX	
Settings	mAP	Rank-1	mAP	Rank-1
Stage1	83.0	86.9	88.3	93.0
Stage2*	81.8	85.7	76.9	88.1
Stage1 + Stage2*	83.1	87.4	88.0	92.5
Stage1 + Stage2* + Eq. 4	85.2	90.8	91.8	94.5

Table 5 Ablation study on individual components. Stage 1 denotes intra-contrast learning stage. Stage 2 denotes inter-contrast learning stage.* denotes only d_1 in Eq. 4 is used in stage 2

For the parameter λ_c, it is used to explicitly decrease the encoding distance between samples from different cameras. It can be observed that when λ_c increased to 0.04, both mAP and Rank1 reached their optimal values, and further increasing λ_c produces a negative effect.

5 Conclusion

This paper introduces two-stage contrastive learning approach for unsupervised person ReID, which aims to mitigate the impact of camera variations by improving the encoding distance calculation across cameras. First, In the intra-contrast learning stage, multi-branching is utilized to train individual encoders for each camera separately. Subsequently, in the inter-contrast learning stage, the encoding results of all encoders are combined to generate a more robust combination coding that is more robust to camera variations. The sample encoding distance is calculated by considering both d_1 (the original distance) and d_2 (the complementary combination coding distance) and d_3 (the Cross-Camera Encouragement distance). Extensive experiments have demonstrated the effectiveness of our proposed method in unsupervised person ReID tasks.

Declarations

Ethical approval Not applicable.

Funding This work was supported by Guangxi Natural Science Foundation (No. 2020GXNSFAA297186), Jiangsu Province Agricultural Science and Technology Innovation and Promotion Special Project (No. NJ2021-21), Guilin Key Research and Development Program (No. 20210206-1), Guangxi Key Laboratory of Precision Navigation Technology and Application (No. DH202227), Guangxi Key Laboratory of Image and Graphic Intelligent Processing (No. GIIP2301). There are no financial conflicts of interest to disclose.

Availability of data and materials The datasets are available at https://virutalbuy-public.oss-cn-hangzhou.aliyuncs.-com/share/data.zip. Code is available at https://github.com/-yjwyuanwu/SET.

References

1. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE transactions on pattern analysis and machine intelligence **40**(2), 392–408 (2017)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee (2009)
3. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 994–1003 (2018)
4. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. Advances in neural information processing systems **32** (2019)
5. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526 (2020)
6. Ge, Y., Zhu, F., Chen, D., Zhao, R., et al.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Advances in neural information processing systems **33**, 11309–11321 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
8. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp. 448–456. pmlr (2015)
9. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding **109**(2), 146–162 (2008)
10. Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C.F.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7919–7929 (2019)
11. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 8738–8745 (2019)
12. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3390–3399 (2020)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
14. Porikli, F.: Inter-camera color calibration by correlation model function. In: Proceedings 2003 international conference on image processing (cat. No. 03CH37429), vol. 2, pp. II–133. IEEE (2003)
15. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, pp. 17–35. Springer (2016)
16. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 608–617 (2019)
17. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10981–10990 (2020)
18. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 79–88 (2018)
19. Xuan, S., Zhang, S.: 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 (2021)
20. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
21. Zeng, K., Ning, M., Wang, Y., Guo, Y.: 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 (2020)
22. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4320–4328 (2018)
23. Zhao, F., Liao, S., Xie, G.S., Zhao, J., Zhang, K., Shao, L.: Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pp. 526–544. Springer (2020)
24. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, pp. 1116–1124 (2015)
25. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2138–2147 (2019)
26. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp. 172–188 (2018)
27. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5157–5166 (2018)
28. Zou, Y., Yang, X., Yu, Z., Kumar, B.V., Kautz, J.: Joint disentangling and adaptation for cross-domain person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 87–104. Springer (2020)