An unsupervised person re-identification approach based on cross-view distribution alignment

Xibin Jia\(^1,2\) | Xing Wang\(^1\) | Qing Mi\(^1\)

\(^1\) Faculty of Information Technology, Beijing University of Technology, Beijing, China
\(^2\) Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing, China

Correspondence
Qing Mi, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China. Email: 289106834@qq.com

Abstract
Unsupervised clustering is a kind of popular solution for unsupervised person re-identification (re-ID). However, due to the influence of cross-view differences, the results of clustering labels are not accurate. To solve this problem, an unsupervised re-ID method based on cross-view distributed alignment (CV-DA) to reduce the influence of unsupervised cross-view is proposed. Specifically, based on a popular unsupervised clustering method, density clustering DBSCAN is used to obtain pseudo labels. By calculating the similarity scores of images in the target domain and the source domain, the similarity distribution of different camera views is obtained and is aligned with the distribution with the consistency constraint of pseudo labels. The cross-view distribution alignment constraint is used to guide the clustering process to obtain a more reliable pseudo label. The comprehensive comparative experiments are done in two public datasets, i.e. Market-1501 and DukeMTMC-reID. The comparative results show that the proposed method outperforms several state-of-the-art approaches with mAP reaching 52.6% and rank1 71.1%. In order to prove the effectiveness of the proposed CV-DA, the proposed constraint is added into two advanced re-ID methods. The experimental results demonstrate that the mAP and rank increase by \(\sim 0.5-2\%\) after using the cross-view distribution alignment constraint comparing with that of the associated original methods without using CV-DA.

1 INTRODUCTION

Person re-identification (re-ID) \([1]\) aims to locate the target person in surveillance videos with a given probe image. With the rapid development of deep learning to techniques, most person re-ID methods focus on supervised algorithm and have achieved high accuracy on public datasets. However, they need to have many pairs of label data between each pair of camera views, which limits the scalability of large-scale applications. In large-scale programs, manually marking pairs of re-ID data is time-consuming and laborious, only unlabelled data is available. In order to solve the scalability problem of re-ID large-scale program, unsupervised re-ID has been widely concerned. Unsupervised learning is achieved by migrating the source domain to the target domain. However, due to the data deviation between the source dataset and the target dataset, the target domain’s performance may decrease significantly \([2, 3]\).

The most popular solution to this problem is the unsupervised adaptive domain (UDA).

Currently, UDA has been widely used in image classification, target detection, face recognition \([4-6]\). There are two main types of UDA methods: one is to try to match the source domain’s feature distribution and target domain \([7, 8]\). The other is to utilise adversarial generative models as a style transformer to convert pedestrian images (with identity annotations) of a source domain into a target domain \([2, 9, 10]\). The style-transferred images are then used to train a re-ID model in the target domain. Many UDA methods preserve discriminative information across domains or camera styles, but they largely ignore the unlabelled samples and substantial sample distributions in target domains. Recent approaches \([11, 12]\) alleviate this problem by predicting pseudo label in the target domain and achieve high performance. They directly use clustering (pseudo) label to fine-tune the model. Still, they are usually susceptible...
to noises and difficult samples, which prevents them from maximising model discrimination in the target domain. Most clustering (pseudo) label methods do not consider the differences of camera cross-views, which leads to the same person’s images in different cameras, clustering and get the different pseudo label.

In order to reduce the impact of camera cross-view, existing methods mainly focus on learning robust and discriminative representation or robust similarity matching measures in a supervised way. For instance, Qian et al. [13] used GAN to match person from different perspectives to the same perspective. Yu et al. [14] ignored the view information and tried to find a shared space where view-specific bias is alleviated. Feng et al. [15] proposed a framework based on deep neural network, which uses view information to learn the view specific network of each camera view with cross-view Euclidean constraint and cross-view centre loss in feature extraction phases. Liu et al. [16] using adversarial thoughts’ advantages, the feature is confused as view-invariant by iterative training between feature extractor and view classifier. However, in unsupervised re-ID, there are few methods to solve the difference of unsupervised cross camera images due to the lack of person label information. Due to the large variations in person images, the use of part-level features is a good method. Wang et al. [17] achieved good performance in person re-recognition through a novel multi-scale multi-patch network and a novel exclusively regularised softmax loss function.

At present, unsupervised re-ID using clustering method to obtain pseudo labels have a strong competitiveness, but this kind of method seldom considers the difference of re-ID cross-views. In order to solve the cross-view difference in the process of unsupervised re-ID clustering, we propose a using cameras cross-view distribution alignment (CV-DA) clustering technique for domain adaptive person re-ID. The images of person re-ID datasets are collected under different cameras, as shown in Figure 1. We can see that there are some differences between different cameras, such as the change of view angle, the change of light and the different shooting time, which leads to people of different identities have more similar appearance characteristics than people of the same identity, making the clustering results of people of the same identity inconsistent. Therefore, the pseudo multilabel should be consistently good across different camera views to be cross-view comparable. Ours proposed CV-DA technology introduces the similarity score by calculating the distance between the image of the target dataset and the marked image of the source dataset and calculates the difference between the similarity score of each person in each camera in the target domain and the overall similarity score as the loss constraint. Through network training to reduce this loss, the network can learn the differences between different camera perspectives, reduce the impact of the camera perspective environment, improve the accuracy of a class of people, these people are similar to the inquired people. In detail, we add the FC layer to calculate the similarity score with the source domain and get the distribution of the image label in the target domain. Therefore, we can get the label distribution of the whole target domain and then calculate the difference between the whole and each camera by calculating each camera’s label distribution in the target domain. The mean value and the distribution distance of standard deviation are used to constrain the difference between the two, so as to achieve the consistency of pseudo label across views for the same kind of people. In addition, we also consider whether the performance of using high-order central moment is more competitive. We carried out a series of ablation studies to verify which of the n-order central distance is more competitive.

The main contributions of this paper can be summarised in three aspects.

- We propose a method to use the cameras cross-view distribution alignment learning to guide the process of clustering to obtain the pseudo label and compare the results of different centre moments to get the most appropriate centre moment.
- We conduct extensive experiments and ablation study on several popular benchmarks including Market1501 [18], DukeMTMC-ReID [19], to demonstrate the effectiveness of the proposed CV-DA, and applied the proposed camera cross-view distribution alignment method to two advanced papers. Compared with the original methods, it has a certain improvement, which verifies our method’s effectiveness.

2 | RELATED WORKS

Although people have done extensive research on person re-ID from different perspectives, we mainly review the domain adaptive person re-ID approaches, primarily driven by unsupervised domain adaptive methods.
2.1 Unsupervised domain adaptation re-ID (UDA re-ID)

2.1.1 Unsupervised domain adaptation

UDA is used to describe that the source domain has a complete label. Then the label of the sample in the target domain is an unknown learning problem. In order to learn the discrimination patterns in the target domain, the early methods focused on the feature/sample mapping between the source domain and the target domain [20, 21]. As a representative method, correlation alignment (CORAL) [21] pursued minimising domain shift by aligning the mean and covariance of source and target distributions. Recent methods [22–24] attempt to reduce domain shift by learning pixel level transformations using generated countermeasure networks (GANS). The most representative cycada [22] transferred samples across domains at both pixel and feature level.

2.1.2 Unsupervised re-ID

Some early unsupervised re-ID methods [25–27] based on handcraft features, they always perform poorly on some large-scale datasets, because they ignore the distribution of samples in the datasets. Benefit from the successful development of deep learning in recent years, some recent work attempt to solve unsupervised domain adaptation based on a deep learning framework. There are two main categories of the existing UDA methods for target recognition. One is the methods including pseudo label [28–30], and the other is domain transformation methods [2, 31, 32]. Deng et al. [33] proposed a similarity preserving generation countermeasure network, which aims to transform the image from the source domain to the target domain. And the transformed image is used to train the recognition model in a supervised way. Wang et al. [7] a transferable joint attribute identity was deep learning (TJ-AIDL) is proposed to learn the attribute semantic and identity distinguishing feature representation space of the target domain without using additional tag data in the target domain.

Due to the better performance of the pseudo label, many methods are based on the pseudo label. The methods based on the pseudo label usually adopt a two-phases training scheme: (1) pre-training of ground truth recognition in the source domain; (2) pseudo label adaptive in the target domain. The pseudo label can be generated by clustering instance features [28, 34, 35] or measuring the similarity with sample features [29, 30, 36], in which the clustering-based pipeline maintains the latest performance. The clustering method's main challenge is how to improve the accuracy of pseudo label and how to reduce the impact of the pseudo-noise label. Jiang et al. [37] adopted local human features to assign multi-scale pseudo labels. Zhang et al. [38] introduced to utilise multiple regularisations alternately. Fabian et al. [35] proposed to generate more robust soft labels via the mutual mean-teaching. Zhai et al. [28] incorporated style-translated images to improve the discriminate of instance features. Although various attempts along this direction have brought great performance improvement, they neglect to make full use of all the valuable information between the two domains, limiting their further progress. When using the pseudo label to fine-tune the target domain model, they simply discard the labelled images in the source domain and the unclustered outliers in the target domain, making the training samples larger The amplitude is reduced. Although these unsupervised domain adaptive methods have made good progress, their performance is still unsatisfactory compared with fully supervised domain adaptive methods.

2.2 Cross camera view difference pseudo label learning

Our method is to learn the similarity between the target domain and the source domain by clustering algorithm, and to learn the invariance of cross camera view by constraints. In unsupervised re-ID, we usually cannot get the identity label information of the person. We can still get the camera's device information, such as location information and camera label information. However, the recognition performance of the camera decreases due to illumination and other environmental impacts. He et al. [39] introduced the heterogeneous learning (HHL) method to improve the re-ID model’s generalisation ability on the target set by achieving camera invariance and domain connectivity time. Yu et al. [36] obtained soft multilabel by referral agents, and used simplified 2-wasserstein distance to learn the consistency of cross-view labels. On this basis, we use it to mine the pseudo label of clustering, and use the high-order centre distance of paper [40] to optimise the Wasserstein distance form to guide the learning of pseudo label in the clustering process, so as to better constrain the alignment loss of pseudo label in the camera cross view distribution. As shown in Figure 2, ID number represents different personal identities. Each photo comes from a different camera and has a similarity score with the query image. We find that the same person’s similarity score may be reduced due to the change of camera view, and the similarity score of different persons may be higher. Therefore, we hope to make the same kind of person similarity score as close as possible by constraining the difference.

3 THE PROPOSED METHOD

3.1 Problem and correlated solution description

For unsupervised re-ID task, we have a source dataset S and a target dataset T. One is a labelled source dataset denoted as \( S = \{(x_s^i, y_s^i) | x_s^i \in \{X^S\}, y_s^i \in \{Y^S\}\} \), where \( \{X^S\} \) indicates the sample set of the source domain and \( \{Y^S\} \) represents the associated identity label set. The S dataset contains \( N^S \) person images and each image \( x_s^i \) has a corresponding label \( y_s^i \), where
The model of supervised pre-training

In the pre-training phase, ResNet50 is used as the backbone network, the source dataset is used as the training dataset, and the final FC layer output is the identification number of the source data. The triplet loss and softmax loss are used as the total loss function in the pre-training phase. The pre-training phase is usually according to the parameters in [41–43], the input image is uniformly resized to 256 × 128, and image enhancement is performed by random flipping and random erasing. The pre-trained network can better extract fine-grained features of the person and focus on some key areas of persons, such as clothes, pants, backpacks etc. So that the network can well extract the features of the target domain clustering, and obtain the more reliable pseudo label. Otherwise, if training from scratch, the initial model’s generalisation ability is poor, and the pseudo labels obtained by clustering are unreliable, which will affect the final accuracy of the model.

3.3 Clustering model for pseudo label generation

If we test the pre-trained network directly on different test sets, we will find that the result does not achieve the expected goal, because there is a certain data deviation between the target and source domains. However, its performance is much better than training ResNet50 model directly on the Imagenet dataset because the latter’s performance is almost zero. From this result, we believe that the model trained directly on the source dataset can still learn some useful re-ID task representations. The target dataset results are so low that the similarity between different images cannot be found correctly. To solve this problem,
FIGURE 3  Overview of the proposed CV-DA approach, \(X^S\) is the source dataset, and \(X^T\) is the target dataset. The CNN model is ResNet50 and pre-trained on source dataset, The last fully connected (FC) layer is discarded, and two additional FC layers are added. The first one has 2048 dimensions. The second FC layer’s output is \(P_s\) dimensional, where \(P_s\) is the number of identity in the source dataset, for cross-view pseudo multilabel learning.

Among them, \(M_{ij}\) calculates the similarity between \(i\) and \(j\) through the features behind the pooling layer, \(R^*(i, k)\) is the refined \(k\)-reciprocal set for sample \(i\) which are obtained by adding some specific reliable constraints as mentioned in [44].

We can get the distance expression between \(x_i\) and \(x_j\):

\[
d_j = (x_i, x_j) = 1 - \frac{\sum_{k=1}^{m_t} \min(M_{ik}, M_{jk})}{\sum_{k=1}^{m_t} \max(M_{ik}, M_{jk})} \tag{3}
\]

And for each \(x_i\) in the target dataset, we search for the nearest neighbour in the source dataset. The confidence function of the distance measure for \(x_i\) is expressed as:

\[
d_w(x_i) = 1 - e^{-\|x_i - N_S(x_i)\|^2} \tag{4}
\]

Where \(N_S(X)\) is the nearest neighbour in the source domain for target image \(i\). The smaller the \(d_w\), the higher the confidence. We transform \(d_w\) and \(d_j\) into the same scale. We normalise \(d_w\) and divide it by max(\(d_w(x_i)\)). Combined with the above formula, we can get the final clustering distance is:

\[
d(x_i, x_j) = (1 - \lambda)d_j(x_i, x_j) + \lambda(d_w(x_i) + d_w(x_j)) \tag{5}
\]

where \(\lambda \in [0, 1]\) is the balancing factor, and we set it to 0.1 in this study.

In Figure 3, after the baseline model, we add the FC layer, which is 2048-dims, with the pseudo label obtained in the clustering stage as the label. In particular, this FC layer is also updated during the training phase. This FC layer is used to extract the information of key areas and integrate the relationship between features.
3.4 | Noise sample processing

This paper mainly uses the DBSCAN clustering method to obtain the pseudo label, but as with most clustering methods, some noise samples will be introduced in the process of clustering. DBSCAN is a density-based clustering algorithm, which assumes that the compactness of sample distribution can determine the category, and the samples of the same category are closely related. In other words, the samples of the same category are not far away from the samples of the same category. By dividing closely connected samples into a cluster, a cluster is obtained. By classifying all closely related samples, the final clustering result is obtained. DBSCAN algorithm usually divides data points into three categories: core points: points with radius EPS greater than minpts, boundary points: points with radius EPS less than minpts but falling near the core points, noise points: points that are neither core points nor boundary points. In the traditional DBSCAN method, some unlabelled noise points are discarded directly. These points are some difficult samples. If we discard them directly, it will cause the loss of data sets and reduce the generalisation ability of the model.

In this paper, we adopt a mechanism of recycling to reduce the loss of datasets and increase the generalisation ability of the model. Specifically, we divide the samples after clustering into two categories, one is the set of samples with pseudo labels, the other is the noise samples without pseudo labels. For noise samples, we calculate the distance between them and the set of labelled samples. By sorting the distance, we select the nearest sample and give it the same pseudo label, as shown in Figure 5. By recycling the noise label, the loss of its data set is reduced and the generalisation ability of the model is increased.

3.5 | Cross-view consistent pseudo multilabel learning

In the unsupervised re-ID, we often cannot get the identity information of the picture. Still, in practice, we can get the image collection's camera information, such as the camera label information, which camera is photographed under, and the camera head's location information. Most of re-ID is based on the cross-view pairs of two images captured by different camera views. Therefore, the same person pseudo label should always be consistent between different cameras. According to the feature distribution of clustering, unlabelled images in the target domain should depend on people's appearance features to reduce the influence of the camera. But the actual clustering results will be affected by the illumination of different cameras and some other factors. For example, in summer, most people wear light-coloured clothes, so no matter which camera the pictures are taken from, people's pseudo labels in light clothes are as similar as possible. In contrast, people in dark clothes have different pseudo labels. MAR [36] is to add a soft multilabel to the unlabelled target dataset by comparing the target person and the reference person who owns the label information (The label consistent with the reference person is different from the pseudo label obtained by clustering), and to learn the discriminative depth feature embedding, that is, to use the soft multilabel to distinguish the visually similar target pairs, so as to reduce the differences of soft multilabel between different views. However, soft label relies on the reference dataset, and the final performance will be affected. The method of obtaining a pseudo label by clustering can obtain better competitiveness, so we use this method to guide the process of learning a pseudo label. We add the FC layer to the original network structure. The FC layer outputs the P2 identity number category, and P3 is the identification number in the source dataset. In this way, the similarity score between the target dataset and the source dataset can be obtained. The process is shown in Figure 6. Where, p1 represents the similarity distribution of images under camera 1, and p2 represents the similarity distribution under camera 2. Based on the analysis, we propose a pseudo multilabel cross-view distribution alignment (CV-DA) loss:

\[ \text{Loss}_c = \sum_i d(p_1(y), p_2(y))^2 \]  

Among them, \( p(y) \) is the similarity score of images in the target dataset, and \( p_v(y) \) is the similarity score of images in the V camera. \( d(.,.) \) is the distance between two scores. We can use some distribution distance measures to calculate this loss—for example, KL divergence [45], JS divergence and Wasserstein distance [46]. Since \( p(y) \) is a two-dimensional variable distributed in \( \sim N \) and its distribution is partial to the normal distribution, we use Wasserstein distance in this paper. Therefore, we can adopt the simplified Wasserstein distance [47] and update the above equation to its Equation (7):

\[ \text{Loss}_c = \sum_i (||\mu_v - \mu||_2^2 + ||\sigma_v - \sigma||_2^2) \]
FIGURE 6 Lc loss calculation flow chart, the image of the target dataset through the FC layer, can obtain person similarity score with the source dataset. The same person gets different scores under different cameras due to different views. We calculate the distribution distance to constrain the difference, so as to reduce the impact of different views under different cameras.

Where $\mu_v$ is the mean value of the image’s similarity score under the $v$-th camera, and $\mu$ is the mean value of the similarity score of the whole dataset. $\sigma_v$ is the standard deviation of the $v$-th camera, and $\sigma$ is the standard deviation of the whole dataset’s image similarity score.

Since the network is in a dynamic learning process, pseudo label are not immutable. Therefore, we have an updated strategy. We store a matrix of similar scores $H = [h_1, ..., h_{N_p}] \in \mathbb{R}^{N_p \times N_u}$ in the CPU. $N_p$ is the number of identities in the source dataset, and $N_u$ is the number of unlabelled images in the target dataset. In the process of continuous training, this similar score matrix is also updated in the following ways:

$$h_i \leftarrow \theta h_i + (1-\theta)y_i$$  (8)

From Formula (7), it is found that the cross-view distribution alignment proposed in this paper depends on the similarity density assumption between different cameras. If there is a large amount of identity overlap between two cameras, the network can learn the differences between different views, constrain its loss. If only a few identities overlap between the two cameras, it is found that the formula is only related to the number of pictures in the camera through Formula (7). By increasing the score of a certain kind of people in the camera, because not knowing the identity information of pedestrians without supervision may lead to higher scores of some difficult samples.

In the previous CV-DA method, we only use first-order and second-order central moments for distance measurement. We consider whether our method is more competitive under higher-order central moments. Referring to the CMD [40] distance formula, we use higher-order central moments such as skewness of third-order central moments and kurtosis of fourth-order central moments:

$$\text{CMD}(p, q) = \sum_{k=2}^{\infty} \frac{1}{|b-a|^k} \| \| c_k(x) - c_k(y) \|_2 \|$$  (9)

Where $c_k(x)$ is:

$$c_k(x) = (E(\prod_{i=1}^{N} (X_i - E(X_i))^r)) \cdot k + \ldots + r_n = k \quad r_1, \ldots, r_n \geq 0$$  (10)

Therefore, we can update the $L_c$ loss to the following formula:

$$L_c = \frac{1}{|b-a|} \sum_{k=1}^{\infty} \| c_k(X) - c_k(Y) \|_2$$  (11)

Among them, $c_{k,v}$ is the similarity scores of pictures under $V$ cameras, and $c_k$ is the similarity scores of all pictures.

3.6 Design of loss function

In the pre-training phases of Section 3.2, softmax loss is used to classify the loss, and the triplet loss for metric learning as the loss function in the pre-training phases. These two methods are also common methods for supervised person re-ID.

For the pre-training multi-classification problem, for the learning feature $f_i$ of the $i$ phases, softmax loss is defined as:

$$L_{\text{softmax}} = - \sum_{i=1}^{N} \log \frac{e^{w_k^T f_i}}{\sum_{k=1}^{C} e^{w_k^T f_i}}$$  (12)

Where, $w_k$ corresponds to a weight vector for class $k$, with the size of mini-batch in training process N and the number of classes in the training dataset C.

Another loss function is the triplet loss function, which calculates the triplet loss function of each batch through the output multidimensional vector. Typically, the loss function is
formulated as follows:

\[
L_{\text{tri}} = \sum_{j=1}^{N_b} \left[ \| x_p - x_a \|_2^2 - \| x_n - x_a \|_2^2 + m \right]_+. \tag{13}
\]

Among them, \( N_b \) is the training batch size, \( p \) and \( n \) are the most dissimilar positive samples and the most similar negative samples to image \( a \), and \( x_p, x_n, x_a \) are the characteristics of the corresponding positive, negative and query images respectively.

Our pre-training phase loss function for optimisation is the combination of softmax loss and batch-hard triplet loss as follows:

\[
L_{\text{Pre-training}} = L_{\text{softmax}} + L_{\text{tri}} \tag{14}
\]

In our clustering phases, because we give a pseudo label to the target dataset, and there is a certain gap with the real label, we only use the triplet loss function in the clustering phases. For camera cross-view differences, we use the \( L_c \) loss described in Section 3.5 to constraint. To summarise, the loss objective of pseudo label learning guided by camera difference is formulated by:

\[
\text{Loss}_{\text{total}} = L_c + L_{\text{tri}} \tag{15}
\]

4 | EXPERIMENT

4.1 | Data set and evaluation metrics

In our experiment we use two large-scale person re-identification (re-ID) benchmark datasets: Market1501 [18] and DukeMTMC-reID [19].

4.1.1 | Market1501

This dataset contains 32668 images, which are collected from six cameras. A total of 1501 persons are marked. 12936 images of 751 identities of persons detected by DPM(target detection algorithm) are used as training set. In order to test, a total of 19732 images with 750 identities and some interference are included in the image set, and 3368 manually cropped character regions form the query set from 750 identities.

4.1.2 | DukeMTMC-ReID

This dataset is a subset of the DukeMTMC. It contains 1812 identities collected by eight cameras. There are 16522 images in training set, 2228 images in query set, 17661 images in gallery. persons with 1404 identities appear in more than two cameras. In addition, similar to Market1501, the remaining 408 identities are disturbed.

We adopt performance evaluation metrics that are commonly used ones in re-ID problem, mean average precision (mAP) and rank1, rank5 and rank10 in cumulative matching characteristic (CMC) curve to evaluate the performance of re-ID.

4.2 | Implementation details

When training with data from the target domain, we maintain the size of the input image, each image is resized to 256 \( \times \) 128 pixels, and the data is expanded. We use random cutting, flipping and random erasing [48]. In order to meet the requirements of hard batch triplet loss, each mini-batch randomly selects \( P = 16 \) identities for sampling, and each identity randomly extracts \( k = 8 \) images from the training set, so the mini-batch size is 128. In our experiment, we set the edge parameter to 0.5. We use Adam [49] and SGD with weight attenuation of 0.0005 to optimise the parameters in the training process. The training times are generally 30 epoch, and the initial learning rate is set to \( 2 \times 10^{-4} \). When the number of training is more than 7 epoch, it becomes 0.3 times of the initial when it is less than 14 epoch, and it becomes 0.1 times when it is more than 14 epoch. Our network model is implemented on the platform of PyTorch [50], using two NVIDIA 1080ti GPU for training. All our experiments on different datasets follow the same setup.

4.3 | Ablation study

We chose to use DBSCAN as the clustering method rather than the commonly used K-means in our method. To evaluate the effect of DBSCAN, we conduct an ablation study. In section, we compare the clustering performance of DBSCAN and K-means. The results are shown in Table 1.

From Table 1, we can see that compared with the backbone method using K-means, the backbone method using DBSCAN improves the mAP by 2.9%, rank1 by 3%, rank5 by 0.9%, and rank10 by 1%. We can conclude that the DBSCAN method is much better than the K-means method.

We have conducted several ablation studies to demonstrate the effectiveness of cross-view pseudo multilabel consistency loss, as shown in Table 2. By observing the table results, we can see that the second-order central moment has the best effect. The third-order central moment has also improved compared with the original network, and the fourth-order central moment’s performance has decreased. Specifically, when we use the second-order central moment from Market1501 \( \rightarrow \) DukeMTMC-ReID, we improve the mAP and rank1 by 0.3%

| Table 1 | Taking Market1501 as the source dataset and DukeMTMC-ReID as the target dataset, the clustering performance of K-means and DBSCAN is compared in backbone

| Method                  | Market1501 \( \rightarrow \) DukeMTMC-ReID |
|-------------------------|--------------------------------------------|
|                          | mAP | rank1 | rank5 | rank10 |
| Backbone with K-means   | 49.60% | 67.90% | 80.10% | 83.30% |
| Backbone with DBSCAN    | 52.50% | 70.90% | 81.00% | 84.30% |
TABLE 2  Compare various central moments above the target domain. When testing on DukeMTMC-ReID as the target dataset, use Market1501 as the source dataset and vice versa. The backbone represents the underlying network, as shown in Figure 3, with no additional FC layer added. CV-DA is our network structure in Figure 3, and the value of $k$ represents different central moments.

| Method     | Market1501 $\rightarrow$ DukeMTMC-ReID | DukeMTMC-ReID $\rightarrow$ Market1501 |
|------------|----------------------------------------|--------------------------------------|
|            | mAP | rank1 | rank5 | rank10 | mAP | rank1 | rank5 | rank10 |
| Backbone   | 52.50% | 70.90% | 81.00% | 84.30% | 57.90% | 78.70% | 87.60% | 90.40% |
| CV-DA ($k=2$) | 52.80% | 71.60% | 80.90% | 84.50% | 59.80% | 79.70% | 89.00% | 91.40% |
| CV-DA ($k=3$) | 52.60% | 71.10% | 81.20% | 84.20% | 58.20% | 79.00% | 87.90% | 91.00% |
| CV-DA ($k=4$) | 49.40% | 69.60% | 79.10% | 82.50% | 58.70% | 78.00% | 88.80% | 91.60% |

FIGURE 7  Dynamic change of cluster number during training

and 0.7%. Similarly, when the centre moment of $k=2$ is used in the model from DukeMTMC-ReID $\rightarrow$ Market1501, mAP and rank1 are increased by 1.9% and 1%, respectively. We can conclude that our CV-DA method is better than the original benchmark method.

We visualise the pseudo label change graph of training time, and the result is shown in Figure 7(a,b). The blue one is the basic network without using CV-DA, the orange one is our method, and the grey one is the real person identity number of the dataset. Through observation, we find that our method is closer to the real label distribution no matter in that dataset.

4.4  Comparison with the state-of-arts

In this section, we compare CV-DA with six state-of-arts methods on Market1501 and DukeMTMC-ReID. The comparative results are listed in Table 3.

4.4.1  Results on DukeMTMC-ReID

On DukeMTMC-ReID, we compare our results with SPGAN and ATNet of style conversion, and domain migration methods TJ-AIDL, MMFA, UDAP, CDS, SSL. SPGAN and ATNet are two style conversion methods trained by converting the image style of the source domain to the target domain. These two methods cannot obtain competitive results. In the training of the target set, the other four domain transfer methods can achieve better results than the method of style conversion, and the CV-DA method proposed by us is better. For pseudo multilabel view alignment constraints, we achieve, rank1 accuracy = 71.6%, rank5 accuracy = 80.9%, and rank10 accuracy = 84.50%, mAP = 52.8%.

4.4.2  Results on Market1501

When we test on the Market1501 dataset, we can also observe a similar situation. Our method is better than other methods. Specifically, we achieve mAP = 59.8%, rank1 accuracy = 79.7%, rank5 accuracy = 89.0%, rank10 accuracy = 91.4%. The effectiveness of the proposed method is further verified.

4.5  Discussion

In order to verify the effectiveness of our proposed method, we selected two advanced clustering unsupervised re-ID papers, ACT [53] and UDAP [34]. The original paper results are compared with those of the CV-DA method proposed by us and the performance of each order central moment. The results are shown in Table 4.
TABLE 3  Comparison of the proposed CV-DA approach with state-of-the-art unsupervised domain adaptive person re-ID methods: For the transfers Market1501 → DukeMTMC-ReID and DukeMTMC-ReID → Market1501

| Method     | Market1501 → DukeMTMC-ReID | DukeMTMC-ReID → Market1501 |
|------------|-----------------------------|----------------------------|
|            | mAP  rank1  rank5  rank10 | mAP  rank1  rank5  rank10 |
| TJ-AIDL [7] | 23.00%  44.30%  59.60%  65.00% | 26.50%  58.20%  74.80%  81.10% |
| MMFA [8]   | 24.70%  45.30%  59.80%  66.30% | 27.40%  56.70%  75.00%  81.80% |
| SPGAN [2]  | 26.40%  46.90%  62.60%  68.50% | 26.90%  58.10%  76.00%  82.70% |
| ATNet [51] | 24.90%  45.10%  59.50%  64.20% | 25.60%  57.00%  73.20%  79.40% |
| CDS [12]   | 42.70%  67.20%  75.90%  79.40% | 39.90%  71.60%  81.20%  84.70% |
| UDAP[49]   | 49.00%  68.40%  80.10%  83.50% | 53.70%  75.80%  89.30%  93.20% |
| MAR [36]   | 48.00%  67.70%  79.80%  81.20% | 40.00%  67.70%  81.90%  - |
| SSL [52]   | 28.60%  52.50%  63.50%  68.90% | 37.80%  71.70%  83.80%  87.40% |
| Ours(CV-DA)| 52.60%  71.10%  81.20%  84.20% | 59.80%  79.70%  89.00%  91.40% |

TABLE 4  Market1501 is the source domain, and DukeMTMC-ReID is the target domain. k represents the results of different central moments, and “+” represents its submethods

| Method     | Market1501 → DukeMTMC-ReID |
|------------|-----------------------------|
|            | mAP  rank1  rank5  rank10 |
| ACT+ [53]  | 51.60%  70.30%  80.80%  83.90% |
| ACT+(k=2) | 52.40%  70.60%  81.30%  84.20% |
| ACT+(k=3) | 51.50%  70.80%  79.90%  83.10% |
| ACT+(k=4) | 49.40%  69.60%  79.10%  82.50% |
| UDAP [54]  | 49.00%  68.40%  80.10%  83.50% |
| UDAP(k=2) | 50.70%  70.40%  80.90%  83.80% |
| UDAP(k=3) | 51.10%  71.40%  80.10%  83.80% |
| UDAP(k=4) | 51.10%  69.80%  80.10%  83.30% |

From Table 4, we can see that the ACT method with CV-DA is better than the original ACT method. When k = 2, the performance of the mAP is the best, which is equivalent to 0.8% improvement of the original method. When k = 3, rank1 is the best, which is 0.5% improvement. In UDAP, the same is true. The performance of UDAP method with CV-DA is better. When k = 3, compared with the original method, mAP increases by 2.1%, rank1 increases by 2%. We can conclude that using CV-DA method can obtain the more reliable pseudo label.

In order to verify that our method is still effective in other datasets, we continue a comparative experiment, the source dataset and the target dataset are transformed to further prove the effectiveness of our method. In the last comparative experiment, the source dataset is Market1501, and the target dataset is DukeMTMC-ReID. In this experiment, we take DukeMTMC-ReID as the source dataset and Market1501 as the target dataset, and the rest of the experimental conditions do not change. The results are shown in Table 5.

From Table 5, we can see that after transforming the dataset. Our method is still effective and has not been affected by the data set. The ACT method using CV-DA is still better than the original method. When k = 2, the mAP has the best performance, increased by 0.8%. When k = 3, rank1 has the best performance, increased by 0.5%. In UDAP, when k = 3 using the CV-DA method has the best performance, mAP and rank1 promote 2.1% and 3.0% respectively.

By observing the results in Tables 4 and 5, we can see that the CV-DA method we proposed above is still effective in other clustering technology papers, and the effect is different with different central moments. Compared with the original paper, the performance will be improved by using 2–3 order centre moment, while the performance will be decreased when using 4-order centre distance. After the analysis of the experimental results, the possible reason is that after using the cross-view consistency constraint of pseudo multilabel, the similarity score of images with the same identity as the query image increases, but the scores of some difficult samples also increase, which leads to some wrong classification results. And when using higher-order, the loss function will become very large, and it is difficult to optimise, resulting in the result is not very good in high-order.

Through the analysis of all our previous experiments, we find that the performance of unsupervised re-ID is also affected by the dataset. The rank1 using DukeMTMC-ReID dataset as
source domain is 5–10% higher than that using Market1501 dataset as source domain. This is because the DukeMTMC-ReID dataset is collected from eight cameras. Compared with the Market1501 dataset, the content of images is more abundant, and the scene transformation is more complicated. Moreover, the number of person identities in DukeMTMC-ReID dataset is 300 more than that in Market1501 dataset, and the generalisation ability of the pre-trained model is stronger. Therefore, choosing a good dataset as the pre-training dataset can better improve the performance of our model.

5 Conclusion

In this paper, we discuss the unsupervised re-ID methods. The commonly used unsupervised re-ID methods give a pseudo label for classification learning through clustering method. Because the images of person dataset are collected in different cameras, there exists diversity among different views. In order to solve the cross-view differences in unsupervised re-ID, we propose cross-view distributed alignment (CV-DA) learning for the pseudo label, which constrains the inconsistency of pseudo label of the same kind of people in different camera views and accordingly reduces the cross-view differences. In addition, we choose the most suitable central moment by comparing the results of different central moments on the proposed CV-DA. Compared with other technologies in two public datasets, our method is proved to be competitive. And further verification of our proposed CV-DA constraint method’s effectiveness is done with merging this constraint in several state-of-art re-ID methods. The experiment results demonstrate that the performance improves with CV-DA and shows our method’s feasibility from another aspect. In the future, we will combine the source dataset with the target dataset to make the network better learn the gap of camera cross-view and obtain a more competitive network.

Acknowledgement

This work is supported by the Beijing Natural Science Foundation under No.4202004, the National Natural Science Foundation of China under Nos.61871276 and U19B2039.

ORCID

Xing Wang https://orcid.org/0000-0002-7176-7036

References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv:1610.02984, (2016)
2. Deng, W. et al.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 994–1003. IEEE, Piscataway, NJ (2017)
3. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Trans. Multimedia Comput. Commun. Appl. 14(4), 83 (2017)
4. Zhong, Z. et al.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European Conference on Computer Vision, pp. 176–192. Springer, Berlin, Heidelberg (2018)
5. Fan, H. et al.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Trans. Multimedia Comput. Commun. Appl. 14(4), (2018)
6. Yu, M. et al.: Dynamic label graph matching for unsupervised video re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5152–5160. IEEE, Piscataway, NJ (2017)
7. Wang, J. et al.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2284. IEEE, Piscataway, NJ 2018.
8. Lin, S. et al.: Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv:1807.01440, (2018)
9. Liu, J. et al.: Adaptive transfer network for cross-domain person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7195–7204. IEEE, Piscataway, NJ (2019)
10. Wei, L. et al.: 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. IEEE, Piscataway, NJ (2018)
11. Lv, J., Wang, X.: Cross-dataset person re-identification using similarity preserved generative adversarial networks. In: Liu, W., Giunchiglia, F., Yang, B., (eds). Knowledge Science, Engineering and Management, pp. 171–183. Springer, Cham (2018)
12. Wu, J. et al.: Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In: 2019 IEEE International Conference on Multimedia and Expo, pp. 886–891. IEEE, Piscataway, NJ (2019)
13. Qian, X. et al.: Pose-normalized image generation for person re-identification. In: European Conference on Computer Vision, pp. 1–18. Springer, Berlin, Heidelberg (2017)
14. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. arXiv:1708.08062, (2017)
15. Feng, Z., Lai, J., Xie, X.: Learning view-specific deep networks for person re-identification. IEEE Trans. Image Process. 27(7), 3472–3483 (2018)
16. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. 2019 IEEE/CVF International Conference on Computer Vision, pp. 6638–6647. IEEE, Piscataway, NJ (2019)
17. Cheng, W.A., et al.: Multi-scale multi-part person re-identification with exclusivity regularized softmax. Neurocomputing 382, 64–70 (2020)
18. Zheng, L. et al.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124. IEEE, Piscataway, NJ (2015)
19. Ristani, E. et al.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jegou, H., (eds). Computer Vision – ECCV 2016 Workshops, pp. 17–35. Springer, Cham (2016)
20. Saenko, K. et al.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N., (eds). In: Computer Vision – ECCV 2010, pp. 213–226. Springer, Berlin, Heidelberg (2010)
21. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. arXiv:1511.05347v2, (2015)
22. Hoffman, J. et al. GyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of Machine Learning Research, pp. 1989–1998. MIT Press, Cambridge, MA (2018)
23. Bouwmans, K. et al.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 95–104. IEEE, Piscataway, NJ (2017)
24. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–18. IEEE, Piscataway, NJ (2017)
25. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision Image Understanding 117(2), 130–144 (2013)
26. Gray, D., Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, pp. 262–275. Springer, Berlin, Heidelberg (2008)
27. Liao, S. et al.: Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206. IEEE, Piscataway, NJ (2015)
28. Zhai, Y. et al.: 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. 9018–9027. IEEE, Piscataway, NJ (2020)

29. Zhong, Z. et al.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 598–607. IEEE, Piscataway, NJ (2019)

30. Liu, H. et al.: Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2167–2175. IEEE, Piscataway, NJ (2016)

31. Deng, W. et al.: 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. IEEE, Piscataway, NJ (2018)

32. Song, L., et al.: Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognition 102, 107173 (2020)

33. Zhong, Z. et al.: Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European Conference on Computer Vision, pp. 172–188. Springer, Berlin, Heidelberg (2018)

34. Zhong, Z. et al.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327. IEEE, Piscataway, NJ (2017)

35. Goodfellow, I. et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680. Curran Associates, Red Hook, NY (2014)

36. Paszke, A. et al.: Automatic differentiation in PyTorch. In: NIPS 2017 Workshop Autodiff Program Chairs, Curran Associates, Red Hook, NY (2017)

37. He, K. et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 770–778. IEEE, Piscataway, NJ (2016)

38. Zellinger, W. et al.: Central moment discrepancy (CMD) for domain-invariant representation learning. International Conference on Learning Representations, ICLR, La Jolla, CA (2017)