# Domain-Adaptation Person Re-Identification via Style Translation and Clustering

Peiyi Wei[1], Canlong Zhang[1(✉)], Zhixin Li[1], Yanping Tang[2], and Zhiwen Wang[3]

[1] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
zcltyp@163.com, lizx@mailbox.gxnu.edu.cn
[2] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[3] School of Electric, Electronic and Computer Science, Guangxi University of Science and Technology, Liuzhou Guangxi, 545006, China

**Abstract.** To solve the two challenges of the high cost of manual labeling data and significant degradation of cross-domain performance in person re-identification (re-ID), we propose an unsupervised domain adaptation (UDA) person re-ID method combining style translation and unsupervised clustering method. Our model framework is divided into two stages: 1) In the style translation stage, we can get the labeled source image with the style of the target domain; 2) In the UDA person re-ID stage, we use Wasserstein distance as the evaluation index of the distribution difference between domains. In addition, to solve the problem of source domain ID labels information loss in the process of style translation, a feedback mechanism is designed to feedback the results of person re-ID to the style translation network, to improve the quality of image style translation and the accuracy of ID labels and make the style translation and person re-ID converge to the best state through closed-loop training. The test results on Market-1501, DukeMTMC, and MSMT17 show that the proposed method is more efficient and robust.

**Keywords:** Person re-identification · Style translation · Unsupervised clustering · Unsupervised domain adaptation

## 1 Introduction

Recently, the accuracy of person re-ID based on supervised learning has been greatly improved, thanks to labeled data sets for training. However, supervised person re-ID faces the following two challenges: 1) When the model obtained by using a labeled data set for learning is transferred to the real scene, the problem

Supported by Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University.


T. Mantoro et al. (Eds.): ICONIP 2021, LNCS 13108, pp. 464–475, 2021.
https://doi.org/10.1007/978-3-030-92185-9_38

of domain mismatch often occurs due to the discrepancy between domains; 2) Labeling large-scale training data is costly, and manual labeling often brings great subjective discrepancy. UDA person re-ID is expected to overcome these problems. The main principle is to transfer the person re-ID model trained on the labeled source domain to the unlabeled target domain.

Recent research on UDA person re-ID is mainly divided into two categories: style-translation-based methods and pseudo-label-based methods. The former mainly uses a style translation model (such as Cyclegan [19]) to translate the style of tagged source domain image to unlabeled target domain image or translate the style of images captured by multiple cameras. The translated image will have the ID label and target domain style of the source domain to reduce the domain discrepancy. Then, the person re-ID model trained by the image after style translation can be generalized to a certain extent. The latter is to generate pseudo tags by clustering instance features or measuring the similarity with sample features, and then use pseudo tags for person re-ID learning. Compared with the former, the latter can achieve better performance, so it has become the mainstream research direction of UDA person re-ID. SPGAN [3] proposes to use the self-similarity before and after the image style translation to solve the image ID label loss during the translation process. However, it still can not solve this kind of problem very well. PTGAN [13] proposed a method of scene transfer combined with style translation to reduce the inter-domain differences caused by factors such as lighting and background. [17,18] mainly use the style translation between cameras to learn the characteristics of camera invariance, to solve the problem of intra-domain discrepancy caused by style changes in the images collected between different cameras, and then obtain a model of camera invariance and domain connectivity.[5] are through repeatedly mining the global information and local information of pedestrians to improve the accuracy of clustering to obtain a high-confidence training set with pseudo-labels for re-ID training. [14] designed an asymmetric collaborative teaching framework, which reduces the noise of pseudo-labels through cooperation and mutual iteration of two models, thereby improving the clustering effect.

However, whether based on style translation or pseudo-label clustering, the adaptation effect achieved is limited. Therefore, this article believes that combining image style translation and the clustering UDA person re-ID method will have greater potential. This paper proposes a UDA person re-ID method that combines style translation and clustering. Figure 1 shows the overall structure of the model. Specifically, cross-domain image translation (CIT) is on the left, which can transform the source domain image without a pair of samples. The converted image has the same identity label, and with the style of the target domain image, CIT can initially reduce the difference between the two domains. The Unsupervised Domain Adaptation Re-identification (UDAR) is on the right. In this process, the adversarial learning method is used. The data distribution of the two datasets is measured by a more stable Wasserstein distance so that the discrepancy between the domains can be further reduced. Then, we use the unsupervised clustering algorithm to assign pseudo labels to the target domain and

then combine the source domain and the target domain into a unified dataset K for person re-ID training. Finally, to solve the ID labels in the style translation process loss and the quality of the generated image. We design a positive feedback mechanism, that is, the result of person re-ID is feedback to the image style translation module to ensure semantic consistency before and after the image style translation, that is, when an ideal pedestrian is obtained when recognizing the model, even if there is a visual difference between the image after the style translation and the original image, their ID labels are unchanged.

In summary, our key contributions are:

- Our proposed UDA person re-ID model of joint image style translation and clustering can make style translation and unsupervised clustering complementary. And better transfer the knowledge learned from the source domain to the target domain. To improve the generalization ability of the model.
- We design a positive feedback mechanism, which can feedback the person re-ID results of the second stage to the style translation of the first stage so that the image's visual quality after the style translation and the accuracy of the corresponding ID label information can be improved.
- In the unsupervised domain adaptation stage, we design a adversarial learning module and measure the distribution discrepancy between the two datasets by Wasserstein distance. The stability of the model is guaranteed when the domain-invariant representation is learned, the effect of to achieve reinforced domain adaptation.

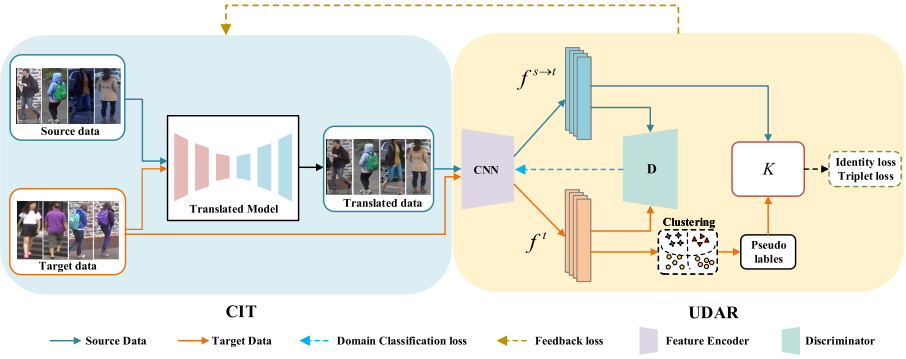


**Fig. 1.** The design of our UDA framework.

## 2 Proposed Method

Given a source domain sample $S$, it is represented as $S = \{(x_i^s, y_i^s)|_{i=1}^{N^s}\}$, where $x_i^s$ and $y_i^s$ are the $i$-th training sample of the source domain and its corresponding

label, respectively. $N^s$ is the number of images in the dataset. Given unlabeled target domain sample $T$, it is represented as $T = \{x_i^t|_{i=1}^{N^t}\}$, where $x_i^t$ is the $i$-th training sample of the target domain. $N^t$ is the number of images in the dataset. Define a feature encoder $F(\cdot|\theta_e)$ with $\theta_e$ as the parameter.

### 2.1 Cross-Domain Image Translation

CIT uses CycleGAN [19] network structure as the main body. The structure is mainly composed of two sets of generators-discriminators $\{G^{s\to t}, D_T\}$ and $\{G^{t\to s}, D_S\}$ to perform image style translation in two directions, forming a cyclic network structure to ensure consistent style translation. The source domain dataset $S'$ after style translation can be obtained through CIT, where $S'$ will keep the same label as $S$.

In CIT, GAN loss $L_{GAN}$, cycle reconstruction loss $L_{cyc}$, and appearance consistency loss $L_{ac}$ [12] are mainly used. Due to the lack of paired training data, there may be infinitely many mapping functions to choose from. Therefore, CycleGan [19] introduced $L_{cyc}$ to try to restore the original image after a cycle of translation and reverse translation, so as to reduce the possible mapping functions. $L_{ac}$ can make the color composition of the image after the style translation similar to that before the translation. They are defined as follows.

$$\begin{aligned} L_{GAN}\left(G^{s\to t}, G^{t\to s}, D_S, D_T\right) &= \mathbb{E}_{x_i^s\sim S}\left[D_S\left(x_i^s\right)^2\right] \\ &+ \mathbb{E}_{x_i^t\sim T}\left[\left(D_S\left(G^{t\to s}\left(x_i^t\right)\right)-1\right)^2\right] \\ &+ \mathbb{E}_{x_i^t\sim T}\left[D_T\left(x_i^t\right)^2\right] \\ &+ \mathbb{E}_{x_i^s\sim S}\left[\left(D_T\left(G^{s\to t}\left(x_i^s\right)\right)-1\right)^2\right], \end{aligned} \tag{1}$$

and

$$\begin{aligned} L_{cyc}\left(G^{s\to t}, G^{t\to s}\right) &= \mathbb{E}_{x_i^s\sim S}\left[\left\|G^{t\to s}\left(G^{s\to t}\left(x_i^s\right)\right)-x_i^s\right\|_1\right] \\ &+ \mathbb{E}_{x_i^t\sim T}\left[\left\|G^{s\to t}\left(G^{t\to s}\left(x_i^t\right)\right)-x_i^t\right\|_1\right], \end{aligned} \tag{2}$$

and

$$\begin{aligned} L_{ac}\left(G^{s\to t}, G^{t\to s}\right) &= \mathbb{E}_{x_i^s\sim S}\left[\left\|G^{t\to s}\left(x_i^s\right)-x_i^s\right\|_1\right] \\ &+ \mathbb{E}_{x_i^t\sim T}\left[\left\|G^{s\to t}\left(x_i^t\right)-x_i^t\right\|_1\right], \end{aligned} \tag{3}$$

In CIT, the source domain ID labels information is often lost. To solve this problem and further improve the quality of style translation, we design a positive feedback mechanism, which uses the person re-ID results of the second stage to feedback a loss $L_{fb}$ to the first stage style transfer module. $L_{fb}$ can make the two modules form a closed-loop, and the entire model framework is more closely connected. Here we use UDAR to calculate the relevant feature operation of the feedback loss, which is defined as follows.

$$
\begin{aligned}
L_{fb}\left(x_i^{s'}, x_i^t\right) &= \lambda_{fb}\mathbb{E}_{x_i^s\sim S, x_i^{s'}\sim S'}\left[\left\|R\left(x_i^s\right) - R\left(x_i^{s'}\right)\right\|_1\right] \\
&+\lambda_{fb\_recon}\mathbb{E}_{x_i^s\sim S, x_i^{s'}\sim S'}\left[\left\|R\left(G^{t\to s}\left(x_i^{s'}\right)\right) - R\left(x_i^s\right)\right\|_1\right] \\
&+\lambda_{fb}\mathbb{E}_{x_i^t\sim T, x_i^{t'}\sim T'}\left[\left\|R\left(x_i^t\right) - R\left(x_i^{t'}\right)\right\|_1\right] \\
&+\lambda_{fb\_recon}\mathbb{E}_{x_i^t\sim T, x_i^{t'}\sim T'}\left[\left\|R\left(G^{s\to t}\left(x_i^{t'}\right)\right) - R\left(x_i^t\right)\right\|_1\right],
\end{aligned} \tag{4}
$$

where $R(\cdot)$ denotes the person re-ID calculation of UDAR. $\lambda_{fb}$ and $\lambda_{fb\_recon}$ are equilibrium coefficients. By using the positive feedback mechanism, when an ideal person re-ID model is trained, even though $S'$ and $S$ or $T'$ and $T$ have visually existed differently, the translation image and the original image should have the same ID label information.

The overall loss for training the CIT is defined as

$$
L_{CIT} = \lambda_{GAN}L_{GAN} + \lambda_{cyc}L_{cyc} + \lambda_{ac}L_{ac} + L_{fb}. \tag{5}
$$

where $\lambda_{GAN}$, $\lambda_{cyc}$ and $\lambda_{ac}$ are weighting parameters.

### 2.2 Unsupervised Domain Adaptation for Person Re-ID

**Domain-Invariant Representation Learning.** In domain-invariant representation learning, when the marginal distributions of two domains are very different or even not overlapped, the gradient may disappear. Therefore, to prevent this problem, this paper uses Wasserstein distance [1] to measure the distribution discrepancy between the source domain and the target domain so that the model can better learn the invariant-feature representation of the domain while can keep the stability.

Specifically, we define a domain discriminator $F(\cdot|\theta_w)$ with $\theta_w$ as the parameter. $F(\cdot|\theta_w)$ can map the features $f^{s\to t} = F\left(x_i^{s'}|\theta_e\right)$ and $f^t = F\left(x_i^t|\theta_e\right)$ into the shared feature space. The Wasserstein distance between two different marginal data distributions $\mathbb{P}_{f^{s\to t}}$ , $\mathbb{P}_{f^t}$ is

$$
\begin{aligned}
W_1\left(\mathbb{P}_{f^{s\to t}}, \mathbb{P}_{f^t}\right) &= \sup_{\|F(\cdot|\theta_w)\|_L\le 1} \mathbb{E}_{\mathbb{P}_{f^{s\to t}}}\left[F\left(f^{s\to t}|\theta_w\right)\right] \\
&- \mathbb{E}_{\mathbb{P}_{f^t}}\left[F\left(f^t|\theta_w\right)\right] \\
&= \sup_{\|F(\cdot|\theta_w)\|_L\le 1} \mathbb{E}_{\mathbb{P}_{x_i^{s'}}}\left[F\left(F\left(x_i^{s'}|\theta_e\right)|\theta_w\right)\right] \\
&- \mathbb{E}_{\mathbb{P}_{x_i^t}}\left[F\left(F\left(x_i^t|\theta_e\right)|\theta_w\right)\right],
\end{aligned} \tag{6}
$$

where $\|\cdot\|_L$ denotes Lipschitz continuous [1,7], the Lipschitz constant of $F(\cdot|\theta_w)$ is set to 1 (i.e., 1-Lipschitz) for the convenience of calculation.

When the parameter of the function $F(\cdot|\theta_w)$ satisfies 1-Lipschitz, the critical loss $L_{wd}$ of the parameter $\theta_w$ can be approximately estimated by maximizing

the Wasserstein distance, as follows

$$L_{wd}\left(x_i^{s'}, x_i^t\right) = \frac{1}{N^s}\sum_{i=1}^{N^s} F\left(F\left(x_i^{s'}|\theta_e\right)|\theta_w\right) - \frac{1}{N^t}\sum_{i=1}^{N^t} F\left(F\left(x_i^t|\theta_e\right)|\theta_w\right), \tag{7}$$

After each gradient update, the $\theta_w$ will be limited to the range of [-c, c] for clipping [1], which may cause the gradient to explode or disappear. Therefore, a gradient penalty $L_{grad}$ [7] should be imposed on $\theta_w$

$$L_{grad}\left(\hat{f}\right) = \mathbb{E}_{\hat{f}\sim\mathbb{P}_{\hat{f}}}\left[\left(\left\|\nabla_{\hat{f}} F\left(\hat{f}|\theta_w\right)\right\|_2 - 1\right)^2\right], \tag{8}$$

where $\hat{f}$ is the random point on the line between the feature distributions of two domains. Note that the gradient penalty here is for each individual input constraint, not for the whole batch. Therefore, layer normalization is used to replace batch normalization in the structure of domain discriminator.

Since the Wasserstein distance is continuous, we can first train the domain discriminator to the optimal result, then keep the parameters unchanged, minimize the Wasserstein distance, and finally make the feature encoder extract the feature representation with domain-invariance

$$\min_{\theta_e}\max_{\theta_w}\left\{L_{wd} + \lambda_{grad}L_{grad}\right\}. \tag{9}$$

where $\lambda_{grad}$ is the balancing coefficient, which should be set to 0 in the process of minimization.

**Unsupervised Person Re-ID.** In this section, we use the unsupervised clustering algorithm (*e.g.*, $k$-means, DBSCAN [4]) to cluster the target domain dataset to obtain the target domain dataset $\hat{T} = \left\{(x_i^t, y_i^t)|_{i=1}^{N^t}\right\}$ with pseudo labels, where $y_i^t$ denotes the pseudo label. And then merge the source domain and the target domain to create a unified dataset $K = S' \cup \hat{T}$, which is used for person re-ID training, where $K = \left\{(x_i, y_i)|_{i=1}^{N}\right\}$, $x_i = x_i^s \bigcup x_i^t$, $y_i = y_i^s \cup y_i^t$, $N = N^s + N^t$. Then, use an identity classifier $C$ to predict the identity of the feature $f = f^{s\to t} \bigcup f^t$. Here, the classification loss and triple loss are adopted to training.

$$L_{id}\left(\theta_e\right) = \frac{1}{N}\sum_{i=1}^{N} L_{ce}\left(C\left(F\left(x_i|\theta_e\right)\right), y_i\right), \tag{10}$$

and

$$L_{tri}\left(\theta_e\right) = \frac{1}{N}\sum_{i=1}^{N}\max(0, m + \left\|F\left(x_i|\theta_e\right) - F\left(x_{i,p}|\theta_e\right)\right\|_2 - \left\|F\left(x_i|\theta_e\right) - F\left(x_{i,n}|\theta_e\right)\right\|_2), \tag{11}$$

where $L_{ce}$ is the the cross-entropy loss, $\|\cdot\|$ denotes the $L^2$-norm distance, m is the margin of the triple distance. $x_{i,p}$ and $x_{i,n}$ is the positive and negative of the same sample.

Then the loss function of person re-ID is

$$L_R^1 = \lambda_{id} L_{id} + L_{tri}, \tag{12}$$

where $\lambda_{id}$ is the weight coefficient of the two losses.

The overall loss for training the UDAR is defined as

$$L_R = \min_{\theta_e} \left\{ L_R^1 + \lambda_w \max_{\theta_w} \left( L_{wd} + \lambda_{\text{grad}} \, L_{\text{grad}} \right) \right\}. \tag{13}$$

where $\lambda_w$ is the balance coefficient.

## 3 Experiments

### 3.1 Datasets and Evaluation Metrics

This section shows the experimental results, mainly including performance comparison with other network models and ablation studies. We used Market-1501 [15], DukeMTMC-reID [10] and MSMT17 [13] for experiments. Evaluation indicators mainly use average accuracy (mAP) and cumulative matching curve (CMC, rank-1) [15] to evaluate the performance of the model algorithm.

### 3.2 Implementation Details

The entire model framework is trained on two NVIDIA Tesla V100 GPUs (64g). The backbone uses ResNet-50 [8] trained on ImageNet [2].

**Stage 1:** The stage 1 training is set to 120 epochs, and there are 200 iterations in each epoch. The learning rate $lr$ for the first 50 epochs is set to 0.0002, and the next 70 epochs will be gradually reduced to 0 according to the lr $=$ lr $\times(1.0 - \max(0, \text{epoch} - 50)/50)$. Use Adam optimizer to optimize the network. The weighting coefficient are respectively set as $\lambda_{GAN} = 1$, $\lambda_{cyc} = 10$, $\lambda_{ac} = 1$, $\lambda_{fb} = 0.1$, $\lambda_{fb\_recon} = 10$.

**Stage 2:** The structure of the domain discriminator $F(\cdot|\theta_w)$ is similar to [9]. Except for the convolutional layer of the last layer, the other three layers all adopt the structure design of FC+LN+LeakyReLU. Stage 2 training defines 50 epochs. The learning rate is set to $10^{-4}$, the Weighting coefficient $\lambda_w = 1$, $\lambda_{grad} = 10$.

**Stage 3:** In this stage, the unsupervised clustering algorithm $k$-means and DBSCAN [4] assign pseudo labels to the target domain. The stage 3 training is set to 120 epochs, and the initial learning rate is set to $3.5 \times 10^{-4}$, which is reduced to $3.5 \times 10^{-5}$ after the 40th epoch and $3.5 \times 10^{-6}$ after the 70th epoch, where each epoch has 200 iterations. Weighting coefficient $\lambda_{id} = 10$.

**Table 1.** Comparison with state-of-the-arts on DukeMTMC-reID, Market-1501 and MSMT17. Our proposed method has good performance for UDA person re-ID.

| Method | D→M | | M→D | | M→MSMT | | D→MSMT | |
|---|---|---|---|---|---|---|---|---|
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| PTGAN [13] | – | 33.5 | – | 16.9 | 2.9 | 10.2 | 3.3 | 11.8 |
| SPGAN [3] | 22.8 | 51.5 | 22.3 | 41.1 | – | – | – | – |
| CamStyle [18] | 27.4 | 58.8 | 25.1 | 48.4 | – | – | – | – |
| HHL [17] | 31.4 | 62.2 | 27.2 | 46.9 | – | – | – | – |
| ECN [16] | 43.0 | 75.1 | 40.4 | 63.3 | 8.5 | 25.3 | 10.3 | 30.2 |
| UDAP [11] | 53.7 | 75.8 | 49.0 | 68.4 | – | – | – | – |
| SSG [5] | 58.3 | 80.0 | 53.4 | 73.0 | 13.2 | 31.6 | 13.3 | 32.2 |
| ACT [14] | 60.6 | 80.5 | 54.5 | 72.4 | – | – | – | – |
| Ours($k$-means) | 61.3 | 82.7 | 53.9 | 73.0 | 15.9 | 41.7 | 18.7 | 45.0 |
| **Ours(DBSCAN)** | **64.5** | **85.0** | **58.4** | **75.1** | **18.9** | **45.0** | **21.8** | **48.1** |

### 3.3 Comparison with the State-of-the-art

We compare our proposed method with state-of-the-art methods on four domain adaptation person re-ID tasks. The experimental results are shown in Table 1, where D→M represents that the source domain is DukeMTMC-reID [10], the target domain is Market-1501 [15], and M→D is the opposite. It is not difficult to see that in D→M, M→D, M→MSMT, and D→MSMT, our method achieves 64.5%, 58.4%, 18.9%, and 21.8% accuracy, respectively. Experimental results show that the proposed method has better performance on UDA person re-ID tasks.

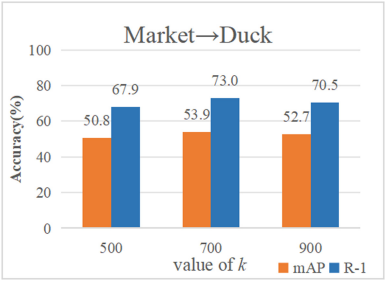


**Fig. 2.** The effect of different $k$ values on model performance (w/$k$-means).

In the stage of unsupervised clustering to obtain pseudo-labels, we used $k$-means to unsupervised clustering of the target domain. Figure 2 shows the impact of different $k$ values on the clustering effect during the clustering process. In the

experiment of M→D, the performance is the best when $k$ is set to 700, which shows that the performance of the model has a certain relationship with the number of cluster clusters set in the target domain. When the set value of $k$ is closer to the actual number of identities in the target domain, the better the model performance.

### 3.4 Ablation Study

In this section, we mainly verify the effectiveness of the proposed method by comparing the results of ablation studies on D→M and M→D (see Table 2).

**Fig. 3.** Image style translation example. The first line represents the original image; The second line is the style translation image when $L_{fb}$ is not used; The third line is to add the style translation image generated by $L_{fb}$.

For the convenience of presentation, $C^{(i)}$ denotes CIT ($i = 0, 1, 2$. Denotes the number of CIT training), $R^{(j)}$ denotes UDAR ($j = 0, 1, 2$. Denotes the number of UDAR training). $C^{(0)}$ denotes that only source domain data is used for training. $C^{(1)}$ denotes that the translated source domain dataset $S'$ for training. $C^{(0)}R^{(1)}$ denotes that $S$ and $T$ are used to train the backbone network. $C^{(1)}R^{(1)}$ denotes that the backbone network is trained using the $S'$ and $T$, only $L_{GAN}$, $L_{cyc}$ and $L_{ac}$ are applied to CIT. $C^{(2)}R^{(2)} + L_{fb}$ denotes that adding $L_{fb}$ to the style translation module for training. Figure 3 shows the effect of the translated image before and after adding feedback loss.

As shown in Fig. 3, we can observe that before and after using the feedback loss, the quality of the generated image has been significantly improved in terms of color composition and clarity. Figure 4 shows some translated images. It can be seen intuitively that the image translation model can make the lighting and texture of the source domain image close to the style of the target domain image, so as to reduce the discrepancy between domains to a certain extent. The translation between Duke and Market datasets is not very obvious because visually, The image illumination and texture of the two datasets are not very

**Table 2.** Ablation studies for our proposed framework (w/$k$-means).

| Method | D→M | | M→D | |
|---|---|---|---|---|
| | mAP | R-1 | mAP | R-1 |
| $C^{(0)}$ | 25.2 | 56.3 | 16.3 | 31.1 |
| $C^{(1)}$ | 30.5 | 60.5 | 23.6 | 45.1 |
| $C^{(0)}R^{(1)}$ | 53.3 | 77.6 | 41.3 | 62.2 |
| $C^{(1)}R^{(1)}$ | 57.1 | 79.3 | 47.8 | 68.5 |
| $C^{(2)}R^{(2)} + L_{fb}$**(full)** | **61.3** | **82.7** | **53.9** | **73.0** |

different. This can also explain that the accuracy of domain adaptive experiment between the two datasets is relatively high. Compared with the two datasets, the image discrepancy of the MSMT17 dataset is very obvious. Therefore, the visual discrepancy from the original image can be obviously seen after image translation. The translated image with better quality can be obtained through the method in this paper. At the same time, the semantic consistency of images before and after translation is maintained. The experimental data in Table 1 also shows that it can effectively improve the experimental accuracy.

**Fig. 4.** Part of the style translation images. The left column is the original images, and the right column is the translated images. The first line is the Market images, the second line is the Duke images, and the third line is the MSMT images.

Table 3 shows the experimental effects of domain adaptation using Wasserstein distance and MMD [6]. Experimental results show that the effect of using Wasserstein distance to measure the discrepancy in feature distribution between two domains is better than MMD [6] on complex tasks such as UDA person re-ID.

**Table 3.** Comparison of Wasserstein distance and MMD domain adaptation experiment effect (w/$k$-means).

| Method | D→M | | M→D | |
|---|---|---|---|---|
| | mAP | R-1 | mAP | R-1 |
| Ours $+L_{MMD}$ | 58.5 | 80.5 | 51.2 | 70.2 |
| Ours $+L_{wd}$ | 61.3 | 82.7 | 53.9 | 73.0 |

## 4 Conclusion

In this paper, we propose a UDA person re-ID method based on image style translation and unsupervised clustering, which combines the labeled dataset with the unlabeled dataset. Among them, CIT can make the source domain image have the visual style of the target domain image and also can initially reduce the discrepancy between the domains. UDAR can further reduce the discrepancy between domains on the basis of CIT through adversarial training so that the backbone can better learn the domain-invariant feature representation. A positive feedback mechanism designed in this paper can combine style translation and unsupervised clustering person re-ID to form a closed-loop. Finally, a more generalized and robust model is obtained by closed-loop training of the two parts.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (Nos. 61866004, 61966004, 61962007), the Guangxi Natural Science Foundation (Nos. 2018GXNSFDA281009, 2019GXNSFDA245018, 2018GXNSFDA294001), Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (No.20-A-03-01), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, and Innovation Project of Guangxi Graduate Education(JXXYYJSCXXM-2021-007).

## References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (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 (2018)
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
5. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

6. Geng, B., Tao, D., Xu, C.: DAML: domain adaptation metric learning. IEEE Trans. Image Process. **20**(10), 2980–2989 (2011). https://doi.org/10.1109/TIP.2011.2134107
7. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of Wasserstein GANs (2017)
8. 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)
9. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
10. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
11. Song, L., et al.: Unsupervised domain adaptive re-identification: theory and practice. Pattern Recognit. **102**, 107173 (2020)
12. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation (2016)
13. 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 (CVPR) (2018)
14. Yang, F., et al.: Asymmetric co-teaching for unsupervised cross domain person re-identification (2019)
15. 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 (ICCV) (2015)
16. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: exemplar memory for domain adaptive person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
17. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 176–192. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_11
18. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: a novel data augmentation method for person re-identification. IEEE Trans. Image Process. **28**(3), 1176–1190 (2019). https://doi.org/10.1109/TIP.2018.2874313
19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)