Unsupervised Person Re-Identification with Iterative Self-Supervised Domain Adaptation

Haotian Tang, Yiru Zhao, and Hongtao Lu*

Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
Department of Computer Science and Engineering,
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
{kentang, yiru.zhao, htlu}@sjtu.edu.cn

Abstract

In real applications, person re-identification (re-id) is an inherently domain adaptive computer vision task which often requires the model trained on a group of people to perform well on an unlabeled dataset consisting of another group of pedestrians without supervised fine-tuning. Furthermore, there are typically a large number of classes (people) with small number of samples belonging to each class.

Based on the characteristics of person re-id and general assumptions related to domain adaptation, we put forward a novel algorithm for cross-dataset person re-id. Our idea is simple yet effective: first, we preprocess the source dataset with style transfer GAN and train a baseline on it in a supervised learning manner, then we assign pseudo labels to unlabeled samples in target dataset based on the model trained on labeled source dataset; finally, we train on the target dataset with pseudo labels in traditional supervised learning manner. We adopt the idea of co-training in the training process to make the pseudo labels more reliable. We show the superiority of our model over all state-of-the-art methods through extensive experiments.

1. Introduction

Person re-identification (re-id) is a challenging computer vision task. In person re-id, given a query person and a gallery, the task is to find this person in the large gallery. Person re-id, together with face recognition, is widely used in real-world security systems. In such real-world applications, it is necessary for our designed algorithm to be domain adaptive because it is often very expensive, or even impossible to manually label a large gallery of pedestrian images from surveillance videos. As a result, we don't have identity annotations in real-world datasets as we do in academic datasets, so traditional supervised learning methods for person re-id actually may not work well in real-world settings.

Based on such an important characteristic of person re-id, in this paper we investigate the self-supervised approach for domain adaptive person re-identification. Unsupervised domain adaptation (UDA) has already been well studied in computer vision community. Popular UDA topics include the image-image domain adaptation such as CycleGAN [48] and the domain adaptation for classification tasks [2]. The former makes use of **generative adversarial networks** (GANs) [10] and emphasizes visual similarity, while the latter takes use of mathematical assumptions and focuses more on feature-level adaptation.

Currently, the state-of-the-art method in person re-identification is proposed in [30], which outperforms the other UDA person re-id methods by a large margin. The iterative pseudo labeling framework is used in this paper. However, our experiments show that problems exist despite the high performance of this pseudo-labeling paradigm. Respectively, the sensitivity to initialization, and the internal inconsistence of the model.

In our work, we propose a novel method based on the iterative pseudo labeling paradigm, which solves the two problems in the clustering-and-labeling process. We utilize the style transfer GAN as the first stage of our method to obtain more stable and better initialization. Then, the network is trained using classical iterative pseudo labeling algorithm until the labels become stable. Finally, we propose a novel **W**ithin-**M**odel **C**o-**T**raining (WMCT) method to eliminate the internal inconsistency within the model, which influences the pseudo label generation process. Actually, WMCT provides self-supervision to the clustering

*Corresponding author.

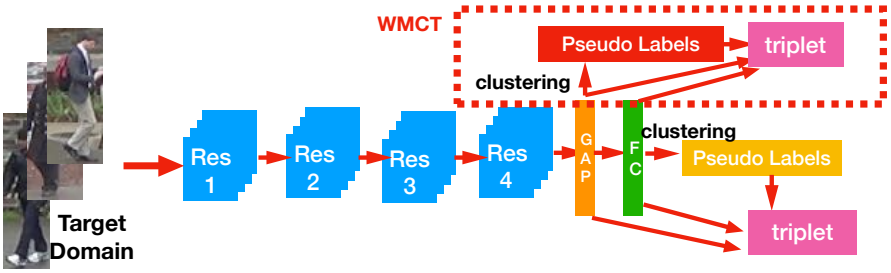

Figure 1. Illustration of the model architecture iterative pseudo labeling algorithm and WMCT. The feature from FC layer is always used to generate pseudo labels, and feature from GAP layer and FC layer are both optimized using triplet loss with the pseudo labels. In WMCT, We generate pseudo labels from the GAP layer additionally as is indicated in the dashed box. The two triplet losses are summed up for within-model co-training.

process using the model itself[1]. Our overall method is called **I**terative **S**elf-**S**upervised **D**omain **A**daptive Person **Re-Id**entification (ISSDA-ReID).

We show through extensive experiments that our improvements over the state-of-the-art iterative pseudo labeling person re-id algorithm are significant. Moreover, all of them are very easy to implement and most importantly, completely **label-free**. Thus, our model is very compatible with real world applications which generally require the property of *plug-and-play*.

2. Related Work

In our proposed three-stage model, the unsupervised image-image translation, unsupervised domain adaptation and person re-identification methods are used. We briefly review some recent advances in these fields.

Unsupervised Image-Image Translation. Image-image translation involves creating a mapping between two image domains. With the rise of generative adversarial networks (GANs) [10] and the introduction of cycle-consistency loss [48], it is possible to realize the unpaired image-image translation task. In our paper, we use the unsupervised image-image translation as preprocessing step.

Unsupervised Domain Adaptation. Unsupervised domain adaptation involves learning a model for the target domain when we have a fully labeled source domain and an unlabeled target domain. There are three mainstream ideas for unsupervised domain adaptation currently. First, Huang et al. proposed methods of sample selection aiming at simulating the target distribution using source samples [16]. Similar to distribution simulation, [11] use maximum mean discrepancy (MMD) in Reproductive Kernel Hilbert Space (RKHS) to align source and target distributions. Second, [2] mentioned three important assumptions in UDA (covariate shift, probablistic Lipschitzness and weight ratio) and proposed iterative pseudo labeling algorithmic paradigm. Third, [35, 23] borrow the idea of adversarial learning and try to make features from source and target domain indistinguishable. Earlier, there are some self-supervised (i.e. using the model itself to create supervision) methods that solve the same problem such as [3], [47].

Supervised Person Re-Identification. Traditional supervised person re-identification tasks arouse great attention these years, they are solved either by metric learning methods or classification-based methods [1, 14]. Since 2017, the state-of-the-art supervised person re-id methods generally take use of regional attention. In short, [12] uses a siamese network to extract the features from different stages, [28] proposed a dual-attention module based on pixelwise feature sequencies. [36] is based on the 14-keypoint pose prior, [15, 32, 29] focus on the occlusions and the influence of background using the attention mechanism. Other methods utilize virtual examples generated by GAN. For example, [45, 40] focus on style transfer, [25, 22] focus on pose-normalized image generation to mitigate pose variation. Probabilistic graph models[27, 5] based on CRFs and random walk and special-purpose network architecture design[19, 4] are also widely used in these years.

Domain Adaptive Person Re-Identification. Recently, unsupervised domain adaptive person re-identification has also been studied. SPGAN [6] and PTGAN [34] utilize image-image translation algorithms to preprocess the source dataset and perform supervised learning directly. MMD is employed to align mid-level features [11]. [33] put forward the TJ-AIDL method, which takes into consideration the additional attribute labels. [20] designs a similar structure to CycleGAN to generate features instead of images. [8] and [30] start from the clustering assumption (or more formally, probablistic Lipschitzness) and propose it-

[1]Unlike most other self-supervised learning methods such as [18] that use additional data to create self-supervision.

erative clustering-based methods.

3. Proposed Method

3.1. Iterative Pseudo Labeling Revisit

[30] proposes to use iterative pseudo labeling in UDA person re-id. The pipeline can be epitomized as figure 1 (without including the dashed box).

The intuition of this method is quite straightforward. When we have an arbitrary model trained on a source dataset that performs relatively poor on the target dataset, we can at least obtain the pairwise distance between target samples (we use distance between GAP features in this stage). Such distance can be used to do density-based clustering with DBSCAN [7], which can also be viewed as unsupervised pseudo labeling. We want to point out that though Euclidean distance is widely used in density-based clustering, the K-reciprocal Encoding Distance [41] has shown its superiority in person re-identification tasks recently. So it is very important to substitute the Euclidean distance with K-reciprocal Encoding Distance.

Then we finetune the re-id model trained on the source dataset in an attempt to minimize only the triplet loss, in which the criterion for positive/negative samples is actually based on the generated pseudo labels. Please notice that the pseudo labels are generated using the feature from the **FC** layer only. This will help produce stable labels in the first few iterations of training, but lead to problems when the training proceeds. Detailed analysis lead to our proposed WMCT algorithm, which will be described later.

After finetuning the model for given epochs, we again update the pairwise distances between the target samples using the finetuned model and use DBSCAN to conduct clustering (or labeling) again. Such procedure will be repeated for a given number of iterations. Finally, the model will converge to a much higher accuracy.

3.2. Pre-Training with Style-Transferred Person Images

Iterative pseudo labeling process is proved to be very effective. However, it strongly relies on the initialization. Generally, the network used in the iterative pseudo labeling process is initialized by training on the source dataset. It is proved in previous works [2] that the success of the pseudo labeling process is largely related to the **weight-ratio** between source and target datasets.

For details of the definition of weight-ratio, we refer the readers to the original paper. Intuitively, this assumption means that the distribution between source and target domain data cannot be too large. To be specific, the source distribution should not be too sparse at region where the target distribution are densely distributed.

To reinforce such assumption, instead of doing it explicitly in the pseudo-labeling process, as [30] did, we propose to do it implicitly through style transfer-based data preprocessing. In detail, we will first use style-transfer GAN (which will be detailed later) to transfer all the person images in the source domain to the style of the target domain, and then supervisedly train a baseline for person re-identification using source labels on the style-transfered dataset. We will compare the performance gain of our model from enhancing the weight-ratio assumption to the one proposed by [30] in next section. The empirical results defend our designs.

Then we detail our design of the style-transfer GAN. Basically, we experimented with multiple GAN settings starting from the original CycleGAN [48] by iteratively adding identity constraint and foreground mask. To be detailed, for the original CycleGAN we have the following standard loss, where F, G represents forward and backward generators, D is the discriminator:

Adversarial Loss:

$$L_{\mathcal{T}adv}(G, D_{\mathcal{T}}, p_x, p_y) = \mathbb{E}_{y\sim p_y}[(D_{\mathcal{T}}(y)-1)^2]+ \mathbb{E}_{x\sim p_x}[D_{\mathcal{T}}(G(x))]^2, \quad (1)$$

$$L_{\mathcal{S}adv}(F, D_{\mathcal{S}}, p_x, p_y) = \mathbb{E}_{y\sim p_y}[(D_{\mathcal{S}}(y)-1)^2]+ \mathbb{E}_{x\sim p_x}[D_{\mathcal{S}}(F(x))]^2. \quad (2)$$

Cycle Consistency Loss:

$$L_{cyc}(F, G) = \mathbb{E}_{x\sim p_x}||F(G(x))-x||_1+ E_{y\sim p_y}||G(F(y))-y||_1. \quad (3)$$

For the identity constraint we have the following equation to ensure that the same person will have similar features, while different people will have far apart features. Θ is a light-weighted neural network to extract the features from the person images. This design is similar to [6]:

Identity Constraint:

$$L_{id}(F, G) = [||x-F(x)||_2 - ||x-y||_2 + \alpha]_+ \quad (4)$$

$$+[||y-G(y)||_2 - ||x-y||_2 + \alpha]_+. \quad (5)$$

Note that x, y belong to different domains, so they can never be the same person. However, though x and $F(x)$, y and $G(y)$ have different styles, they must be the same person. In that way, such loss constrains the feature distribution of the same person to be closely clustered.

Also, we have foreground mask. This is inspired by [34] by replace the cyclic loss with a masked cyclic loss which separates the background from the foreground.

Masked Cycle Consistency Loss:

$$L_{mask}(F, G) = \mathbb{E}_{x\sim p_x}||M_x \odot (F(G(x))-x)||_1+ E_{y\sim p_y}||M_y \odot (G(F(y))-y)||_1. \quad (6)$$

Figure 2. Some images generated by the style transfer GAN. The first row includes pedestrian images from original DukeMTMC-reID dataset, and the second row corresponds to style-transferred DukeMTMC-reID images in Market1501 style.

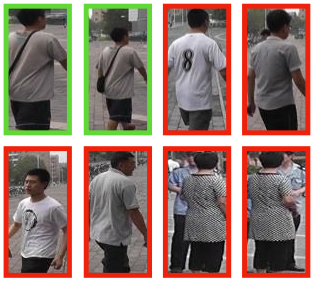

Figure 3. A sample cluster generated in the iterative labeling process. This cluster remains stable since iteration 5. The task is transfer learning from DukeMTMC-reID to Market1501. The four people in the second row no longer belong to this cluster after WMCT.

Later, we will show in experiments that **only the identity-constraint loss**, combined with CycleGAN can boost the performance of the iteratively trained baseline with pseudo labels. CycleGAN itself or CycleGAN trained with foreground mask will not help much. This demonstrates the necessity of feature space alignment despite not very impressive visual quality (see Figure 2 for details about the generated images).

After we transfer all the images in the source domain to the target-styled images, we train a baseline model in a traditional supervised learning manner with the given labels in the source domain. We show the details of the baseline model we used in Section 4.2. We follow common practices to use both cross entropy loss and triplet loss to boost the performance of the baseline model on the source dataset. After the baseline model converges on the style-transferred source dataset, we use it to initialize the iterative pseudo labeling procedure directly.

3.3. Iterative Self-Supervised Post-Processing

Standing on the shoulder of the first two stages which performs iterative pseudo labeling starting from a very good initialization, the last stage of our model aims at further improving the performance of domain adaptive person re-id. The model architecture in this stage is indicated in figure 1. Besides the dashed box in figure 1, the WMCT also adds up two triplet losses, instead of using just one in the iterative pseudo labeling process.

This stage is actually based on an observation of a small flaw in the iterative pseudo labeling process. True, the pseudo labels are refined through iterative clustering and are becoming closer and closer to the ground truth. However, there's absolutely no supervision on the clustering or labeling process. In this case, if the pseudo labels begin to deviate from ground truth seriously from one iteration, it is possible that even poorer pseudo labels will be generated in the coming iterations. Another problem is that the labeling result will become relatively stable after a long enough period of time, and of course this labeling result will not be the ground truth labeling.

As an example, a certain cluster in the Market1501 dataset is illustrated in figure 3 after we run the image-to-image translation based dataset preprocessing and iterative pseudo labeling. The first two images belong to one person clearly, and the last two images also belong to another person. The other four images belong to different people. This cluster remain unchanged since the fifth iteration of the iterative pseudo labeling process. Obviously, this cluster is wrong.[2]

In this case, further finetuning just means supervised training using wrong labels. This may lead our model to poor performance. So we actually need some **self-supervision** (since we cannot directly tell our model which cluster is wrong) to help us break the **erroneous** stability.

To handle this problem, we think of using the model itself to create supervision on the labeling process. We are greatly inspired by the classical co-training method [3] in which two classifiers influence the performance of each other. Typically, co-training method requires the two classifiers to be from two different models and the two models are from different views. Such requirement is too strong for person re-identification tasks. Actually multi-view information is not available in most academic datasets and training two models is too computationally expensive.

Because of the limitations of the original co-training method, we propose our novel Within-Model Co-Training (WMCT) method (see algorithm 1) in this stage which ignores the original restrictions of co-training methods.

The WMCT method is simple (with only minor modifications to iterative pseudo labeling) yet effective. The key difference from iterative pseudo labeling lies in line 5 and line 10 of the proposed algorithm. Note that there are two features from our model (GAP and FC). In the iterative pseudo labeling stage, we use only GAP feature to generate pseudo labels and use such labels as the criterion for the triplet loss. In WMCT, we simply generate two clustering

[2]Just for illustration. The knowledge that *this cluster is wrong* and any labeled information related to the target dataset will **never** be used in training.

Algorithm 1 WMCT

Require:

Pretrained feature encoders for target domain $\mathbf{x}_{GAP}^{(0)}, \mathbf{x}_{FC}^{(0)}$. Target dataset $\mathcal{T}$ with size N. Percentage p, minimal cluster size M, iteration number T.

Ensure:

Finetuned feature encoder $\mathbf{x}_{GAP}^{(final)}, \mathbf{x}_{FC}^{(final)}$.

Obtain pairwise distance matrix $T_{GAP}^{(0)} = \mathbf{x}_{GAP}^{(0)}(\mathcal{T}), T_{FC}^{(0)} = \mathbf{x}_{FC}^{(0)}(\mathcal{T})$.
Process $T^{(0)}$ with k-reciprocal encoding and obtain $D_{GAP}^{(0)}, D_{FC}^{(0)}$.
Get clustering threshold τ_{GAP} and τ_{FC} using the average of top pN elements of $D_{GAP}^{(0)}, D_{FC}^{(0)}$.
Get pseudo labels $L_{GAP}^{(0)} = DBSCAN(D_{GAP}^{(0)}, \tau_{GAP}, M), L_{FC}^{(0)} = DBSCAN(D_{FC}^{(0)}, \tau_{FC}, M)$.
Train $\mathbf{x}_{GAP}^{(1)}$ and $\mathbf{x}_{FC}^{(1)}$ with $L_{GAP}^{(0)}, L_{FC}^{(0)}$ using equation 7.
for $i = 1; i \leq T; i++$ **do**
    Obtain pairwise distance matrix $T_{GAP}^{(i)} = \mathbf{x}_{GAP}^{(i)}(\mathcal{T}), T_{FC}^{(i)} = \mathbf{x}_{FC}^{(i)}(\mathcal{T})$.
    Process $T^{(i)}$ with k-reciprocal encoding and obtain $D_{GAP}^{(i)}, D_{FC}^{(i)}$.
    Get pseudo labels $L_{GAP}^{(i)} = DBSCAN(D_{GAP}^{(i)}, \tau_{GAP}, M), L_{FC}^{(i)} = DBSCAN(D_{FC}^{(i)}, \tau_{FC}, M)$.
    Train $\mathbf{x}_{GAP}^{(i+1)}, \mathbf{x}_{FC}^{(i+1)}$ using $L_{GAP}^{(i)}, L_{FC}^{(i)}$ and equation 7.
end for

results using both features at a time. Then, the triplet loss is calculated with respect to each of the labeling systems. Formally, the loss function is defined as:

$$L_{triplet} = L_{triplet-GAP}^{(GAP)} + L_{triplet-FC}^{(GAP)} + L_{triplet-GAP}^{(FC)} + L_{triplet-FC}^{(FC)}, \quad (7)$$

where the superscript denotes the clustering criterion, and the subscript indicates the loss is calculated from feature of which layer.

At the start of WMCT, GAP and FC layers produce different labeling results. This is actually desired, both layers can utilize the prediction of the other layer to make their own predictions more reliable, because if a negative pair is classified as positive pair in one labeling system and as negative pair in the other system, the loss tends to lower the confidence of the system predicting the pair to be positive. This is fundamentally different from the original iterative pseudo labeling framework, in which such pairs must be more firmly clustered as a positive pair because their distance will continuously narrow down as the training process goes on. We will show the effectiveness of our proposed WMCT method in next section.

4. Experiments

For simplicity, we denote the style-transfer based preprocessing as stage 1, iterative pseudo labeling as stage 2, WMCT as stage 3 in our experiments according to the order of their appearance in the implementation.

4.1. Datasets and Evaluation Metrics

We conduct our experiments on two popular academic datasets, **Market1501** [38] and **DukeMTMC-reID** [40, 26]. For Market1501 dataset, there are 12,936 pedestrian images from 1,501 identities in the training set and 19,732 images from 1,500 people in the test set and 3,368 images in the query set. Six cameras are utilized to capture the pedestrian video, and deformable part model (DPM) [9] detector is used to extract the bounding boxes. All the bounding boxes have the same size of 128 $\times$ 64. For DukeMTMC-reID dataset, there are 1,404 identities in all. Half of them are used for training and another half is used for testing. There are totally 2,228 images in the query probe and 17,661 images in the test gallery. The bounding boxes in this dataset have varied sizes.

As to the evaluation metric, we follow [40] to use cumulative matching characteristic (CMC) and mean average precision (mAP) as our evaluation metrics.

4.2. Implementation Details

We use PyTorch framework to train our ISSDA-ReID model. For the person re-id baseline model, we use ResNet50 [13] as the backbone network. The layers before the final global average pooling layer remain unchanged. After this layer, we add a batch normalization layer [17] followed by a leaky linear rectified unit (Leaky-ReLU) layer. Then, two fully connected layers with size 2048, 751 (or 702 for DukeMTMC-reID, i.e. number of classes) follow. The first fully connected layer is equipped with batch normalization and leaky relu (see figure 1). The baseline model is trained with an initial learning rate of 1×10^{-4} with warmup and stepwise learning rate decay. The mini-batch is composed of 64 images, with 16 people and 4 images for each person.

In the second stage, we define $p = 0.016$, minimal cluster size $M = 4$, iteration number $T = 20$ (definitions of all parameters can be found in algorithm 1). The learning rate is set to 6×10^{-5}. We use the stochastic gradient descent

Methods	Rank1	Rank5	Rank10	mAP
PUL [8]	44.7	59.1	65.6	20.1
SPGAN [6]	51.5	70.1	76.8	22.8
MMFA [21]	56.7	-	-	27.4
TJ-AIDL [33]	58.2	74.8	81.1	26.5
CamStyle [46]	58.8	78.2	84.3	27.4
HHL[43]	62.2	78.8	84.0	31.4
ARN [20]	70.2	80.4	86.3	39.4
ECN [44]	75.1	87.6	91.6	43.0
IPL[30]	75.8	89.5	93.2	53.7
PartAligned[37](s)	81.0	92.0	94.7	63.4
SVDNet[31](s)	82.3	-	-	62.1
Ours	**81.3**	**92.4**	**95.2**	**63.1**

Table 1. Comparison with state-of-the-art methods on the Market1501 dataset. All the methods are evaluated in the single-query mode and the source dataset is DukeMTMC-reID. The best result in the **domain adaptation** setting is in bold font. Note that *(s)* means the model is trained supervisedly on labeled Market1501 dataset. We call [30] **I**terative **P**seudo **L**abeling based on our understanding of the paper.

Methods	Rank1	Rank5	Rank10	mAP
PUL [8]	30.4	44.5	50.7	16.4
SPGAN [6]	41.1	56.6	63.0	22.3
TJ-AIDL [33]	44.3	59.6	65.0	23.0
MMFA [21]	45.3	-	-	24.7
HHL[43]	46.9	61.0	66.7	27.2
CamStyle [46]	48.4	62.5	68.9	25.1
ARN [20]	60.2	73.9	79.5	33.4
ECN [44]	63.3	75.8	80.4	40.4
IPL[30]	68.4	80.1	83.5	49.0
LSRO[40](s)	67.7	-	-	47.1
PAN[39](s)	71.6	-	-	51.5
Ours	**72.8**	**82.9**	**85.9**	**54.1**

Table 2. Comparison with state-of-the-art methods on the DukeMTMC-reID dataset. All the methods are evaluated in the single-query mode and the source dataset is Market1501. The best result in the **domain adaptation** setting is in bold font. Note that *(s)* means the model is trained supervisedly on labeled DukeMTMC-reID dataset.

(SGD) method with Nesterov gradient acceleration in our experiments. The L_2 weight decay is set to 5×10^{-4}. We also use random crop, random erasing [42] for train-time augmentation. In the third stage, all the parameters are the same as what is used in stage 2. In both stage two and stage three, the batch organization is 32 people per batch and 4 images per person.

4.3. Comparison with State-of-the-art Methods

We compare our algorithm with multiple state-of-the-art methods including PUL [8], SPGAN [6], CamStyle [46], TJ-AIDL [33], MMFA [21], HHL [43], ARN [20], ECN [44], the state-of-the-art method proposed by [30] and the method proposed by [24]. It can be seen clearly from table 1 and 2 that our method have superior performance to all previous methods by a large margin. Actually, we also cite results of some supervised learning methods from some papers published in 2017. The comparison with supervised learning results further shows that our method is very powerful.

Methods	Rank1	Rank5	Rank10	mAP
baseline	25.0	39.8	47.2	12.6
+ST	35.7	52.1	58.8	19.4
+IPL	67.5	80.1	82.9	48.3
+ST+IPL	70.6	82.3	85.2	52.3
+ST+IPL+WMCT	**72.8**	**82.9**	**85.9**	**54.1**

Table 3. Effectiveness of Three Stages. Our model setting is unsupervised domain adaptation from Market1501 dataset to DukeMTMC-reID dataset. ST: style transfer, IPL: iterative pseudo labeling, WMCT: within-model co-training.

Methods	Rank1	Rank5	Rank10	mAP
baseline	40.1	57.2	64.4	16.0
+ST	52.0	69.9	76.8	24.5
+IPL	75.0	88.5	92.6	52.8
+ST+IPL	79.2	90.5	93.8	59.0
+ST+IPL+WMCT	**81.3**	**92.4**	**95.2**	**63.1**

Table 4. Effectiveness of Three Stages. Our model setting is unsupervised domain adaptation from DukeMTMC-reID dataset to Market1501 dataset.

4.4. Ablation Study

In all the tables of this section, *baseline* just means the direct transfer result of a model trained with full supervision on the original source dataset (with no style transfer).

4.4.1 The Effectiveness of ISSDA-ReID

We show the effectiveness of some combinations of all three stages on both *Market to Duke* and the opposite task. The results can be seen in table 3 and 4. For the effectiveness of stage 3, we'd like to mention in addition that directly train stage 2 for more iterations actually also improve the result of *baseline+stage1,2*. However, the improvement is far smaller (only around 1%, compared with ours 2.2%). What's more, our WMCT method can still lift the performance of a more fully-trained model obtained by iterative pseudo labeling. So it's definitely better than just training for more iterations.

4.4.2 The Detailed Effectiveness of Style Transfer Preprocessing

It is mentioned previously that the choice of style-transfer GAN may influence the effect of Stage 1. Here, we want to

Methods	Rank1(D)	mAP(D)	Rank1(A)	mAP(A)
baseline	25.0	12.6	67.5	48.3
CycleGAN	33.1	17.0	67.4	49.0
+identity	35.7	19.4	**70.6**	**52.3**
+mask	**38.3**	**22.1**	68.5	49.2

Table 5. Detailed ablation study for stage 1. In the evaluation metrics, *D* means direct transfer and *A* stands for after iterative pseudo labeling.

show in this subsection that only the identity constraint loss combined with the original CycleGAN helps improve the performance **when combined with stage 2**, even if the direct transfer performance under such setting is **not** the highest.

As is shown in the table 5, despite the fact that adding a mask provides a better direct transfer performance after stage 1, it gives only modest performance gain after iterative pseudo labeling. The original CycleGAN doesn't help at all if not combined with the identity constraint. We therefore conclude that the identity constraint loss is the most important building block for the preprocessing stage, and thus we discard the foreground mask in preprocessing.

We also note that our proposed enhancement of the weight-ratio assumption in the first stage gives an average of 2.15% performance gain, which is significant compared with [30], who achieved less than 0.5% with directly optimizing weight-ratio.

4.4.3 The Stability of Our Results

We mentioned previously that [30] is sensitive to initialization and ours are less sensitive. Concretely, we ran our model with only the iterative pseudo labeling process proposed by [30] and get an average result of 67.5% Rank-1 accuracy on the $Market2Duke$ task, with a highest accuracy of 68.5% and lowest accuracy of 67.1%, which differs by 1.4%. We also run our whole pipeline for 10 times on the same task, getting 72.5% to 73.1% and averaged 72.8%, which is a 0.6% maximal difference, which demonstrates that our results are more stable.

5. Conclusion

In this paper, we proposed a three stage model ISSDA Re-ID for domain adaptive person re-identification. Through style-transfer GAN guided preprocessing, iterative training and labeling and the WMCT stage, we improve the state-of-the-art performance in domain adaptive person re-id by a large margin.

However, there are still some problems with our model which are open for future study. First, we fail to find a positive relationship between the performance after stage 1 and the final performance. We have found ways to improve the stage 1 performance even more compared with what is mentioned in the ablation studies but the final performance can be significantly worse. We leave the possibility of finding better initialization for the iterative pseudo labeling process (for example, the newly proposed CamStyle [46]) and further exploration on the relationship between the performance in stage 1 and the final performance for future research. Second, whether the clustering framework can be substituted with other methods **without loss of accuracy**. The clustering algorithms involve heavy CPU calculation (mainly because of the calculation of K-reciprocal encoding) and long training time. Our model is actually very time-consuming in the training stage. If it is possible to find some alternative algorithm paradigm to clustering, the training efficiency may be improved.

Acknowledgement This paper is supported by NSFC (No. 61772330, 61533012, 61876109), the Basic Research Project of Innovation Action Plan (16JC1402800), the advanced research project (No.61403120201), Shanghai authentication key Lab. (2017XCWZK01), and Technology Committee the interdisciplinary Program of Shanghai Jiao Tong University (YG2015MS43).

References

[1] J. Almazán, B. Gajic, N. Murray, and D. Larlus. Re-id done right: towards good practices for person re-identification. *arXiv*, abs/1801.05339, 2018.

[2] S. Ben-David and R. Urner. Domain adaptation-can quantity compensate for quality? *Ann. Math. Artif. Intell.*, 2014.

[3] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[4] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, June 2018.

[5] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, June 2018.

[6] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, June 2018.

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[8] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv*, abs/1705.10444, 2017.

[9] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *NIPS*, 2006.

[12] Y. Guo and N.-M. Cheung. Efficient and deep person re-identification using multi-level similarity. In *CVPR*, June 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv*, abs/1703.07737, 2017.

[15] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *CVPR*, June 2018.

[16] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2006.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[18] H. Jiang, E. G. Learned-Miller, G. Larsson, M. Maire, and G. Shakhnarovich. Self-supervised depth learning for urban scene understanding. *CVPR*, 2018.

[19] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, June 2018.

[20] Y. Li, F. Yang, Y. Liu, Y. Yeh, X. Du, and Y. F. Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. *CVPR Workshop*, 2018.

[21] S. Lin, H. Li, C. Li, and A. C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *BMVC*, 2018.

[22] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, June 2018.

[23] S. Liu, Y. Sun, D. Zhu, G. Ren, Y. Chen, J. Feng, and J. Han. Cross-domain human parsing via adversarial feature and label adaptation. In *AAAI*, 2018.

[24] J. Lv, W. Chen, Q. Li, and C. Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *CVPR*, June 2018.

[25] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *ECCV*, September 2018.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016.

[27] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, June 2018.

[28] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, June 2018.

[29] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, June 2018.

[30] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv*, abs/1807.11334, 2018.

[31] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, Oct 2017.

[32] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, June 2018.

[33] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, June 2018.

[34] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, June 2018.

[35] M. Wulfmeier, A. Bewley, and I. Posner. Incremental adversarial domain adaptation. *arXiv*, abs/1712.07436, 2017.

[36] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, June 2018.

[37] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, Oct 2017.

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[39] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv*, 2017.

[40] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[41] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661, 2017.

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv*, abs/1708.04896, 2017.

[43] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, September 2018.

[44] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identication. In *CVPR*, 2019.

[45] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, June 2018.

[46] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2019.

[47] Z. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.*, 2005.

[48] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.