MULTI-DOMAIN SYNCHRONOUS REFINEMENT NETWORK FOR UNSUPERVISED CROSS-DOMAIN PERSON RE-IDENTIFICATION

Sikai Bai, Junyu Gao*, Qi Wang*, Xuelong Li

School of Computer Science and School of Artificial Intelligence, Optics and Electronics Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China {whitesk1973, gjy3035, crabwq}@gmail.com, xuelong_li@nwpu.edu.cn.

ABSTRACT

Unsupervised cross-domain person re-identification (re-ID) is a challenging task, because it is an open-set problem with completely unknown person identities in the target domain. Existing methods attempt to tackle the challenge by transferring image style across domains or generating pseudo labels in the target domain, whereas the valuable information in multiple domains (i.e., source domain, style-transferred data, and target domain) is not taken fully into consideration. To this end, we propose a novel multi-domain synchronous refinement (MDSR) network, where valuable knowledge from multiple domains is sufficiently exploited and refined to enforce the discriminative ability of the model. MDSR network contains two complementary modules dedicated to source-to-target domain adaptation and style-transferred data to the target domain adaptation, respectively. The domain adaptive knowledge from two modules is aggregated in the final stage. Extensive experiments verify our method achieves significant improvements over the state-of-the-art approaches on multiple unsupervised domain adaptive person re-ID tasks.

Index Terms—Person re-identification, unsupervised domain adaptation, synchronous refinement

1. INTRODUCTION

Person re-identification (re-ID) aims at matching the same person from images captured by multiple cameras. It has received more and more attention due to its wide applications, such as video surveillance and public safety. Encouraged by the success of deep learning and the availability of large-scale labeled datasets, supervised re-ID methods [1], [2], [3] have made impressive progress. However, as shown in Fig. 1(a), re-ID models pre-trained on label-rich datasets (source domain) are directly applied to other unlabeled datasets (target domain), and the performance is decreased by a big margin due to domain gaps between different datasets. Moreover, it is impractical to continuously annotate person identities in various new target domains.

To mitigate the domain gap, a large number of methods have been proposed to focus on unsupervised domain adaptation (UDA) for person re-ID. They can be mainly categorized as two directions: 1) Domain translation-based methods [4], [5], [6], [7]. These methods focus on the translation from source domain to the target domain to create style-transferred samples by using generative adversarial networks (GAN). However, as shown in Fig. 1(b), source domain data is simply discarded after translation, and there are no pseudo-lables for target domain data in these approaches. 2) Pseudo-label-based methods [8], [9], [10], [11]. These methods first train re-ID models on the labeled source domain, then finetune the pre-trained models with pseudo-labeled data on the target domain, where the pseudo labels can be predicted by some cluster algorithms. Although this pipeline dominates current state-of-the-art performance, noisy pseudo-labels and large domain gaps between source and target domains have

Fig. 1. Cross-domain person re-ID methods considered in this paper. The orange arrow represents the domain translation operation for generating style-transferred samples. The red cross indicates that the input data is simply discarded, and the blue cross denotes that the corresponding input data is only used during pre-training.
negative effects on the performance. Besides, as shown in Fig. 1(c), the useful knowledge from style-transferred samples is not considered in these methods.

In this paper, we propose a multi-domain synchronous refinement (MDSR) network for unsupervised domain adaptive person re-ID. MDSR network learns the valuable knowledge from two designed modules, and aims to reduce the negative impacts of domain gap, preserve original inter-sample relations in the source domain and make full use of the useful data in multiple domains. Specifically, we first propose a unified style-similarity hybrid learning (SSHIL) module to reduce the negative impacts of the domain gap. SSHIL consists of a domain translator, a feature encoder and a cluster generator. The domain translator aims at generating style-transferred samples with a similar style to the target domain. The feature encoder is used to perform joint training for the style-transferred samples and target-domain data. The cluster generator aims to predict reliable pseudo labels on the target domain by using the extracted features from above encoder. Meanwhile, a relation-invariant hybrid learning (RHIIL) module is devised for preserving inter-sample relations in the source domain. RHIIL contains a feature encoder and a cluster generator. The feature encoder in RHIIL aims at joint training for the labeled source domain data and the pseudo-labeled target data. The cluster generator in RHIIL utilizes the extracted features from the feature encoder in RHIIL to predict reliable pseudo labels. Furthermore, to construct a multi-domain synchronous refinement network, the two above modules are synchronously trained to sufficiently exploit the valuable knowledge from multiple domains. Finally, the discriminative ability of the model is further refined by aggregating the learned knowledge from two modules.

The major contribution of our work can be summarized as three-fold. (1) In response to the negative impacts of domain gaps and the loss of original inter-sample relations during translation, this work develops a style-similarity hybrid learning (SSHIL) module and a relation-invariant hybrid learning (RHIIL) module, respectively. (2) In order to sufficiently exploit the knowledge from multiple domains, this work proposes a multi-domain synchronous refinement (MDSR) network by synchronously training the two above modules, and the ability of the model is further refined by aggregating the learned knowledge from two modules. (3) The proposed MDSR network achieves consistent performance gain over the state-of-the-arts on multiple unsupervised domain adaptation tasks of person re-ID.

2. RELATED WORK

2.1. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) attempts to generalize the learned knowledge from the labeled source domain to the target domain with unknown classes well. Some approaches aim at aligning the feature distributions between the source and target domains [12], [13]. Correlation Alignment (CORAL) [12] mapped the covariance and mean of two distributions, and CYCADA [13] transferred samples across domains at both pixel- and feature-level. Moreover, some other methods try to find domain-invariant feature spaces [14], [15], [16], [17]. Long et al. [14] and Tzeng et al. [15] employed the Maximum Mean Discrepancy (MMD). Ganin et al. [16] and Ajakan et al. [17] built domain confusion loss for this purpose. However, the above methods only focus on conventional domain adaptation, where the source and target domains share the same set of class labels.

2.2. Cross-domain Person Re-Identification

Cross-domain person re-identification is an open-set problem with disjoint person identities between different domains. There are many works focusing on unsupervised methods include domain translation and pseudo-label generation. Specifically, the domain translation-based methods [4], [6], [7] used the re-ID model pre-trained in labeled style-transferred samples to perform unsupervised finetuning in the target domain. SPGAN [4] and PTGAN [6] maintained the ID-related features invariant by identity-based regularization. SDA [7] used online relationship consistency regularization term to maintain inter-sample relations. The pseudo-label-based approaches [18], [19], [20], [21] enforce feature learning in the target domain by utilizing the pseudo labels, where the labels are commonly generated by cluster algorithms (e.g. Kmeans or DBSCAN). SSG [20] assigned multi-scale pseudo labels by introducing human local features. Besides, in order to further optimize the predicted pseudo labels, MMT [21] and SPCL [18] introduced mutual mean-teaching and self-paced learning mechanism, respectively. Nevertheless, the above methods are unable to consider the valuable knowledge from style-transferred samples. And most of these methods suffer from the adverse impacts of domain gap between source and target domain and noisy pseudo labels.

3. PROPOSED METHOD

The overview architecture of the proposed multi-domain synchronous refinement (MDSR) network is illustrated in Fig. 2. In this section, we first give some essential symbolic definitions, then introduce the details of the proposed SSHIL module and RHIIL module. Finally, the proposed MDSR network is presented by integrating the above two modules to sufficiently exploit the valuable knowledge from multiple domains.

3.1. Notation

To facilitate understanding, some symbolic definitions are given here. Given the labeled source domain data $X^s$ and un-
labeled target domain data $X^t$, where $X^s = \{(x^s_i, y^s_i)\}_{i=1}^{N_s}$ contains $N_s$ person images with corresponding ground-truth identity labels, and $X^t = \{(x^t_i, y^t_i)\}_{i=1}^{N_t}$ involves $N_t$ target images with pseudo labels predicted by the clustering algorithm. Furthermore, we denote style-transferred samples as $X^{st} = \{(x^{st}_i, y^{st}_i)\}_{i=1}^{N_{st}}$, in which $x^{st}_i$ and $y^{st}_i$ are generated from the source data domain.

3.2. Style-Similarity Hybrid Learning

Previous UDA methods for person re-ID commonly attempt to reduce domain gap by two ways, i.e., generating style-transferred samples for source domain data or generating pseudo labels in the target domain. Different from these works, the designed style-similarity hybrid learning (SSHIL) module considers domain adaptation in two aspects, where it integrates the processes of domain translation and pseudo label prediction, while performing joint training with the generated style-transferred samples and target domain data. Specifically, as shown in Fig. 2, there are three sub-modules in SSHIL, namely, the domain translator $T$, the feature encoder $E_1$, and the cluster generator $G_1$. They are parameterized with $\theta_t$, $\theta_{e1}$ and $\theta_{c1}$, respectively.

**Domain Translator.** The domain translator focuses on translating the annotated source domain data to style-transferred samples, which is implemented by a generative adversarial network, where the samples enjoy ground-truth identity labels derived from the source domain data and have a similar style with the target domain data. The constructed domain translator only need to maintain identity-related information invariant during translation. Given a source data pair $(x^s, y^s)$ as input, the corresponding style-transferred sample pair $(x^{st}, y^{st})$ generated by the domain translator $T$ can be formally expressed as:

$$x^{st} = T(x^s; \theta_t), y^{st} = y^s.$$  \hfill (1)

**Feature Encoder.** The feature encoder in SSHIL is designed to provide more discriminative target features by performing joint training with the style-transferred samples and target domain data. Initially, the style-transferred samples enjoy the ground-truth identity labels while the target domain data has no identity labels. It is worth noting that our feature encoder does not require additional pre-training in the source domain or style-transferred samples. Formally, given a style-transferred image $x^{st}$ and target domain image $x^t$, the corresponding output features $f^{st}$ and $f^t$ are obtained by the following mapping:

$$f^{st} = E_1(x^{st}; \theta_{e1}), f^t = E_1(x^t; \theta_{e1}).$$ \hfill (2)

**Cluster Generator.** To refine joint training for the above feature encoder, we first predict the pseudo labels for the target domain data by using DBSCAN cluster algorithm. Then a hybrid-label system HLS1 is constructed by concatenating the predicted pseudo labels and the source-ground truth identity labels. Furthermore, we refine the clustering algorithm by a novel clustering criterion and take it as our cluster generator, which encourages to preserve more reliable data in clusters by measuring the reliability of each feature point during clustering. Concretely, given the feature $f^t$ in the target domain, our clustering criterion can be represented as:

$$r_1(f^t) = \max(0, \frac{\|f^t\|_2 \cdot \gamma_1(f^t)}{\|f^t\|_1 \cdot \gamma_1(f^t)} - \alpha),$$

$$r_2(f^t) = \max(0, \frac{\|f^t\|_2 \cdot \gamma_2(f^t)}{\|f^t\|_1 \cdot \gamma_2(f^t)} - \beta),$$ \hfill (3)
where $\gamma(f')$ denotes the cluster set containing the feature $f'$; $\gamma_1(f')$ and $\gamma_2(f')$ are the clusters including $f'$ when shrinking and enlarging the distance threshold in the DBSCAN algorithm, respectively. Two dynamic thresholds (i.e., $\alpha$ and $\beta$) and two obtained clustering rules (i.e., $r_1(f')$ and $r_2(f')$) focus on identifying reliable feature points in the clusters on-the-fly. Overall, we compare the initial clustering results of feature points with the results when shrinking and enlarging the threshold distance respectively, to analyze the clustering reliability of feature points from target domain data. Finally, given the output feature $f'$ of the above feature encoder, the corresponding pseudo-label $\hat{y}_k'$ can be obtained by the following mapping:

$$\hat{y}_k' = \begin{cases} G_1(f'; \theta_1), & c_1(f') > 0, c_2(f') > 0, \\ -1, & \text{otherwise}, \end{cases}$$

when $\hat{y}_k' = -1$, the corresponding samples are divided into discrete instances with individual person class labels. Therefore, the number of pseudo labels equals $n_t^l + n_t^r$, where $n_t^l$ and $n_t^r$ represent the number of reliable clusters and discrete instances, respectively.

### 3.3. Relation-Invariant Hybrid Learning

Although the style-transferred samples have similar styles to the target domain and reduce the negative impact of the domain gap, it is found in [7] that these samples only retain identity-related information while losing the original inter-sample relations from the source domain during translation.

To maintain original inter-sample relations invariant and implement joint learning for source domain and target domain data, we devise a relation-invariant hybrid learning (RIILH) module. As illustrated in Fig. 2, RIILH contains a feature encoder $E_2$ and a cluster generator $G_2$, which have the same architecture details as sub-modules in SSIL, but the parameters are not shared with them. We parameterize two sub-modules in RIILH with $\theta_{E_2}$ and $\theta_{G_2}$ respectively, and build a new hybrid-label system HILS2 by concatenating the source ground-truth and the target pseudo label set. These parameterized sub-modules are iteratively trained to strengthen the discriminative power of features, which is implemented by simultaneously using the source domain data with original inter-sample relations and the pseudo labeled target domain data as the input of the module. Therefore, the output feature of the feature encoder $E_2$ in RIILH can be denoted as:

$$\begin{cases} f_k^l = E_2(x'; \theta_{E_2}), & x' \in X^l, \\ f_k^r = E_2(x'; \theta_{E_2}), & x' \in X^r. \end{cases}$$

### 3.4. Multi-domain synchronous refinement network

In order to perform multiple domains synchronous training and refinement, we construct a novel network by integrating the proposed SSIL module and RIILH module, referred as MDSR network. We use the abundant data from the source domain and target domain as the MDRS network input for synchronous training. Specifically, for source domain data, the MDSR network reduces the negative impact of the domain gap by generating style-transferred samples in SSIL, and the original inter-sample relations invariant is considered by directly utilizing the rich source information in RIILH. For target domain data, the valuable complementary knowledge is extracted from the two designed modules, where we refine the predicted pseudo labels based on the designed clustering criterion. At the last stage, we perform aggregation for the learned knowledge from two modules by a weighted factor $\lambda$ to further refine the discriminative ability of the model, and the aggregated results are regarded as the final outputs of the MDSR network. Therefore, given target domain data $x^t$, the final output $y_o$ of the MDSR network can be given by:

$$y_o = \lambda C_1(f_1^t) + (1 - \lambda) C_2(f_2^t) = \lambda C_1(F_1(x'; \theta_{E_1})) + (1 - \lambda) C_2(F_2(x'; \theta_{E_2})),$$

where $C_1$ and $C_2$ are two learnable classifiers from SSIL and RIILH to identify person class.

### 4. EXPERIMENTS

#### 4.1. Datasets and Implementation Details

**Datasets.** The proposed approach is evaluated on three wide-used person re-ID datasets, including Market-1501 [22], DukeMTMC-reID [23] and MSMT17 [6]. Data statistics are shown in Table 1, DukeMTMC-reID considers additional 408 identities as distractors, and MSMT17 is the most challenging and large-scale dataset.

**Implementation Details.** We adopt the ImageNet-pretrained IBM-ResNet50 as the backbone of our feature encoders (i.e., $E_1$ and $E_2$), and utilize eSAPGAN [5] as our domain translator $T$ in SSIL. The training parameters include 50 epochs, batch size 64, momentum 0.2, weight decay 0.0005. We initialize the learning rate to be 0.00035 and decrease it by 10 every 20 epochs. For two dynamic thresholds (i.e., $\alpha$ and $\beta$) in clustering criterion, inspired by [18], they are set to the sum and difference of the top-90% $\gamma(f'; \gamma_1(f'))$ in the first epoch and the maximum $\gamma(f'; \gamma_1(f'))$ in the cluster of each epoch. And the weight factor $\lambda$ in our MDSR network

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Test</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market-1501 [22]</td>
<td>751/12936</td>
<td>1501/19732</td>
<td>1501/32688</td>
</tr>
<tr>
<td>DukeMTMC-reID [23]</td>
<td>702/16522</td>
<td>702/19899</td>
<td>1812/36411</td>
</tr>
</tbody>
</table>
Table 2. Comparison of our MSDR network with state-of-the-art methods. The best results are highlighted with bold fonts.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td>Rank-1</td>
<td>Rank-5</td>
</tr>
<tr>
<td>BCN [24]</td>
<td>CVP’19</td>
<td>43.0</td>
<td>75.1</td>
</tr>
<tr>
<td>SSG [20]</td>
<td>ICCV’19</td>
<td>58.3</td>
<td>80.0</td>
</tr>
<tr>
<td>MMCL [19]</td>
<td>CVPR’20</td>
<td>60.4</td>
<td>84.4</td>
</tr>
<tr>
<td>BCN++ [25]</td>
<td>TPAMI’20</td>
<td>63.8</td>
<td>84.1</td>
</tr>
<tr>
<td>DG-Net++ [26]</td>
<td>ECCV’20</td>
<td>63.8</td>
<td>78.9</td>
</tr>
<tr>
<td>JVTCC++ [27]</td>
<td>ECCV’20</td>
<td>67.2</td>
<td>86.8</td>
</tr>
<tr>
<td>GPR [28]</td>
<td>ECCV’20</td>
<td>71.5</td>
<td>88.1</td>
</tr>
<tr>
<td>SDA [7]</td>
<td>αXiv’20</td>
<td>74.3</td>
<td>89.7</td>
</tr>
<tr>
<td>MMT [21]</td>
<td>ICLR’20</td>
<td>76.5</td>
<td>90.9</td>
</tr>
<tr>
<td>SPCL [18]</td>
<td>NeurIPS’20</td>
<td>79.2</td>
<td>91.5</td>
</tr>
</tbody>
</table>

MSDR (Ours) 81.2 92.3 97.1 98.1 72.7 84.8 92.0 93.4 34.2 60.3 72.0 76.3

Table 3. Components performance of MSDR network.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Market-1501 → DukeMTMC-reID</th>
<th>mAP</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>32.9</td>
<td>61.7</td>
<td>77.1</td>
<td>82.0</td>
<td></td>
</tr>
<tr>
<td>SSSIL</td>
<td>69.3</td>
<td>83.5</td>
<td>90.5</td>
<td>92.6</td>
<td></td>
</tr>
<tr>
<td>RIIH</td>
<td>70.6</td>
<td>83.8</td>
<td>91.5</td>
<td>93.1</td>
<td></td>
</tr>
<tr>
<td>MSDR</td>
<td>72.7</td>
<td>84.8</td>
<td>92.0</td>
<td>93.4</td>
<td></td>
</tr>
</tbody>
</table>

is 0.5. Besides, we adopt mean average precision (mAP) and CMC Rank-1, Rank-5, Rank-10 accuracies to evaluate the performance of models. All experiments are conducted on 4 GTX-1080TI GPUs, and it is worth noting that no post-processing technology (e.g., multi-query fusion [22] or re-ranking [29]) is used in our experiments.

4.2. Comparison with State-of-the-arts.

We compare the proposed approach with state-of-the-art methods on three unsupervised domain adaptive re-ID tasks. As shown in Table 2, our MSDR network significantly outperforms all existing methods in terms of mAP and CMC Rank-1, Rank-5, Rank-10 accuracies. Specifically, compared with domain translation-based approaches, the proposed MSDR network surpass the state-of-the-art method SDA by a big margin, obtains relative improvements of 6.9% (81.2% vs. 74.3%), 6% (72.7% vs. 66.7%) and 3.9% (34.2% vs. 30.3%) mAP on Market-1501 → DukeMTMC-reID, Market-1501 → DukeMTMC-reID and Market-1501 → DukeMTMC-reID → MSMT17 tasks, respectively. And comparing with pseudo-label-based methods, although these methods have previously achieved the optimal performance, MSDR still outperforms the state-of-the-art approach SPCL by 2% (81.2% vs. 79.2%), 2.8% (72.7% vs. 69.9%) and 2.4% (34.2% vs. 31.8%) mAP on three domain adaptation tasks. These results show that the effectiveness of our MSDR network by synchronously refining the knowledge from multiple domains is impressive, and it yields a new record on the cross-domain person re-id problem.

4.3. Ablation Study

Components Effectiveness. To evaluate the effectiveness of each component in our MSDR network (i.e., SSSIL and RIIH), we create a baseline method that directly apply the pretrained IBN-ResNet50 in the labeled source domain to conduct fine-tuning on the unlabeled target domain, while comparing it with the single modules and the combination of both modules (i.e., MSDR) on Market-1501 → DukeMTMC-reID task. As shown in Table 3, SSSIL utilizes style-transferred samples and the pseudo labeled target data to jointly train the model, and it gains 36.4% mAP improvement. RIIH considers original inter-sample relations on source domain data and predicts pseudo labels on the target domain, and it obtains 37.7% mAP improvements. Furthermore, better performance can be provided by the weighted aggregation for SSSIL and RIIH together.

Parameter Analysis. Finally, we analyze the impacts of the different values of the hyperparameter $\lambda$ on performance, when SSSIL and RIIH are weighted aggregated in the final stage of the model. As shown in Fig. 3, our MSDR network achieves the optimal performance when setting the weighted factor $\lambda$ as 0.5 in Eq. 6.
5. CONCLUSION

In this paper, we propose a multi-domain synchronous refinement network for unsupervised cross-domain person re-ID, which sufficiently exploits the valuable knowledge from multiple domains by two designed modules (i.e., style-similarity hybrid learning module and relation-invariant hybrid learning module). Extensive experiments demonstrate the superiority of our method over recent state-of-the-art approaches.

6. REFERENCES


