Structured Domain Adaptation With Online Relation Regularization for Unsupervised Person Re-ID

Yixiao Ge, Feng Zhu, Dapeng Chen, Member, IEEE, Rui Zhao, Member, IEEE, Xiaogang Wang, Member, IEEE, and Hongsheng Li, Member, IEEE

Abstract—Unsupervised domain adaptation (UDA) aims at adapting the model trained on a labeled source-domain dataset to an unlabeled target-domain dataset. The task of UDA on open-set person reidentification (re-ID) is even more challenging as the identities (classes) do not have overlap between the two domains. One major research direction was based on domain translation, which, however, has fallen out of favor in recent years due to inferior performance compared with pseudo-label-based methods. We argue that domain translation has great potential on exploiting valuable source-domain data but the existing methods did not provide proper regularization on the translation process. Specifically, previous methods only focus on maintaining the identities of the translated images while ignoring the intersample relations during translation. To tackle the challenges, we propose an end-to-end structured domain adaptation framework with an online relation-consistency regularization term. During training, the person feature encoder is optimized to model intersample relations on-the-fly for supervising relation-consistency domain translation, which in turn improves the encoder with informative translated images. The encoder can be further improved with pseudo labels, where the source-to-target translated images with ground-truth identities and target-domain images with pseudo identities are jointly used for training. In the experiments, our proposed framework is shown to achieve state-of-the-art performance on multiple UDA tasks of person re-ID. With the synthetic→real translated images from our structured domain-translation network, we achieved second place in the Visual Domain Adaptation Challenge (VisDA) in 2020.

Index Terms—Domain translation, person reidentification (re-ID), relation consistency, unsupervised domain adaptation (UDA).

I. INTRODUCTION

PERSON reidentification (re-ID) [11]–[16] aims at identifying images of the same person across multiple cameras. Despite great advances of deep-learning-based re-ID methods in recent years, large domain gaps still pose great challenges on generalizing the trained models from a labeled source domain to an unlabeled target domain. There are two main research directions toward solving the problem of domain adaptive person re-ID, i.e., domain-translation-based methods [1]–[5], [17] and pseudo-label-based methods [7]–[10], where the latter ones dominate the current literature with state-of-the-art performance.

Although domain-translation-based methods have fallen out of favor in recent years due to their uncompetitive performance, we argue that they have great potential to make full use of valuable source-domain data with accurate identity labels if they are imposed with proper regularization strategy. Translating source-domain images into the target domain to create new training samples with identity labels is at the core of domain-translation-based methods. Previous works used identity-based regularization (e.g., classification loss [3], [4], [18], contrastive loss [1], or triplet loss [5]) to preserve ID-related appearance during image translation, i.e., forcing different identities’ images to be well-separated from each other after translation. However, we observe that their domain-translated images cannot well maintain intersample relations even with such ID-based constraints. The intersample relations in our article are represented by semantic affinities/similarities between samples’ encoded features in the latent space, rather than simple positive/negative relations given by ground-truth identity labels. As demonstrated in Fig. 1(a), in the existing image-to-image translation frameworks, although images of the same class can still be well-classified into the same class after translation, the intersample affinities in the latent space might not be well-maintained to effectively regularize the translation process. For example, in the first case of Fig. 1(c), persons from two distinct identities are in yellow and orange, respectively. After domain translation by CycleGAN [6] and SPGAN [1], the images of different persons show similar color and appearances. The intersample relations are not well-maintained. In the second case, a male and a female are in the same color—blue. After translation by CycleGAN [6], the two images of the male change to green and blue, respectively. After translation by SPGAN [1], the two images for the male are both in green while the girl is still in blue. The intersample relations are not well-maintained by the existing translation-based methods. We argue that maintaining intersample relations during image-to-image translation is critical for generating informative training samples, which cannot
be effectively regularized by identity-based regularization in [1], [3]–[5], [18].

To tackle this challenge, we propose an end-to-end structured domain adaptation (SDA) framework with a novel online relation-consistency regularization term. It consists of a structured domain-translation (SDT) network, a source-domain person image encoder, and a target-domain encoder. The SDT network generates the source-to-target translated images, which can benefit the training of the target-domain encoder. At the same time, two domains’ encoders are coupledly trained to model intersample relations on-the-fly, which regularize the training of the domain-translation network. We alternatively optimize the target-domain encoder and SDT network in each iteration, to exploit their mutual benefits.

More specifically, the SDT network adopts the CycleGAN [6] architecture for translating source- and target-domain images. A novel relation-consistency loss is proposed to regularize the training of source-to-target domain translation for maintaining the original intersample relations, which are generated by the source-domain encoder on-the-fly [Fig. 1(b) and (c)]. Note that our relation-consistency loss is different from conventional ID-based constraints in [1]–[5], which only apply “one-hot” regularization that requires translated images of different classes to be well-separated after translation but ignores intersample relations.

Meanwhile, the source-domain encoder is trained with source-domain images and ground-truth identities. An improved target-domain encoder is trained with both the source-to-target translated images and target-domain images via a joint cross-domain label system, which is constructed with their associated ground-truth and pseudo labels. In this way, both the target-domain encoder and its generated pseudo labels can be improved with the optimization of the SDT network. As shown in the first case of Fig. 1(c), the triplet translated by our method shows distinguishable appearance for images of different persons. In the second case, the clothes of the male and the female are both translated to the same color (green). Our proposed method is able to effectively maintain their original intersample relations during translation via the proposed SDT network and relation-consistency loss.

In summary, the contributions of this article are threefold.

1) To properly exploit the valuable source-domain data in domain-translation-based unsupervised domain adaptation (UDA) methods, we propose a novel online relation-consistency regularization term to better supervise the domain translation process.

2) The domain-translated images can serve as informative training samples to improve the target-domain encoder and help generate more accurate pseudo labels. The domain-translation network and target-domain encoder alternately promote each other to achieve optimal re-ID performance.

3) Our framework achieves state-of-the-art performance on multiple domain adaptive person re-ID benchmarks. Moreover, the proposed SDT method contributes to our solution [19] in the Visual Domain Adaptation Challenge (VisDA) in 2020, which ranks second out of 153 teams.
Table VII “Baseline” versus “Baseline + raw source-domain data”). Our proposed SDT method can be well-compatible with pseudo-label-based methods, including a clustering-based baseline [22] or state-of-the-art MMT [10] (see Section IV-C).

B. Generic Methods for UDA

Feature-level and pixel-level adaptations were commonly adopted by UDA methods for tackling more general tasks. The feature-level adaptation methods [25]–[30] aimed at aligning feature distributions between the source and target domains by learning domain-invariant features with a domain adversarial discriminator [27], [31] or reducing the maximum mean discrepancy (MMD) [32] distance between domains. However, such methods are unable to handle the open-set re-ID problem with disjoint label systems in two domains [24], [33], [34].

The other category of pixel-level adaptation methods [35]–[37] minimized the domain shifts by translating images to the same domain, which has been widely studied in semantic segmentation. These pixel-level adaptation methods focused on cross-domain class/prediction consistency, which are related to our method. However, they still ignored the consistency of intersample relations and data distributions during translation, facing the same challenge as translated-based UDA methods for person re-ID [1]–[4].

C. Relation Preserving Embedding

Relation preserving embedding [38], [39]–[41] aims at encoding features that preserve certain prior structure or relational information among samples. There exist a large number of embedding techniques, which mainly include conventional matrix-factorization-based methods [38], [39] and deep-learning-based methods [40], [41]. The matrix-factorization-based methods try to represent intersample relations/affinities in the form of a matrix, such as adjacency matrix, Laplacian matrix, node transition probability matrix, and Katz similarity matrix. However, those algorithms cannot be used as regularizations in our framework, as their objective functions are optimized via matrix factorization and do not support gradient-based optimization. The deep-learning-based methods were also studied to capture nonlinear structures among data points. However, these algorithms are different from our method in two aspects. First, those methods have different purposes from our method, i.e., they target at directly encoding low-dimensional features to maintain certain intersample relations while we maintain intersample relations to regularize translation network training. More importantly, they adopt different types of manually specified intersample relations as learning targets while our intersample relational supervisions are online generated and are also alternatively optimized with the domain-translation network.

III. STRUCTURED DOMAIN ADAPTATION

We propose an SDA framework with a novel online relation-consistency regularization term to tackle UDA for person re-ID. The overall framework, as illustrated in Fig. 2, consists of an SDT network and two domain-specific person image encoders. The translation network and target-domain encoder are jointly optimized and promote each other to learn more discriminative person features.

The key innovation of our framework lies in the generation of informative training samples by translating source-domain images into the target domain under relation-consistency regularization generated by the image encoders on-the-fly.

A. Source-Domain Encoder Pretraining

We pretrain the source-domain person image encoder \( F_s \) for 1) providing “ground-truth” intersample relations between source-domain images to regularize the proposed structured domain translation and 2) providing weight initialization for the target-domain person image encoder \( F_t \).

Given source-domain samples \( X_s \), the encoder \( F_s \) is trained to transform each sample \( x \in X_s \) into a feature vector \( f_s \). If the feature vector \( f_s \) is properly embedded, it could be used to correctly predict its ground-truth identity \( y \). With a learnable classifier \( C : f_s \rightarrow \{1, \ldots, p_s\} \), where \( p_s \) is the number of identities in the source domain. A cross-entropy classification loss \( L_{ce} \) and a triplet loss [42] are adopted jointly for training

\[
L_{ce}(F_s, C_s) = \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| C_s(F_s(x_s), y_s) \right| \right| \right]
+ \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| f_s(x_s) - f_s(x_s') \right| \right| + m - \left| \left| f_s(x_s) - f_s(x_s') \right| \right| \right]^+ 
\tag{1}
\]

where \((\cdot)^+ = \max(0, \cdot)\) with a margin \(m\), and the subscripts \(ps\) and \( pt\) denote the mini-batch’s hardest positive and negative feature indexes for the anchor \( f_s \), respectively.

Once trained, \( F_s \) is frozen to provide stable regularizations for intersample relations.

B. Structured Domain Translation

We propose an SDT network to generate informative training samples by translating source-domain images \( X_s \) to the target domain, which focuses not only on image-style transfer but more on how to maintain their original intersample relations and distributions. We adopt the widely used CycleGAN [6] architecture for our translation network, which is trained to translate images along two directions with corresponding generators \( G^{s\rightarrow t} \) and \( G^{t\rightarrow s} \).

1) Conventional Cycle Generation Losses: The general training objective of a CycleGAN [6] for image-to-image translation consists of adversarial losses \( L_{adv} \) and \( L'_{adv} \), the cyclic reconstruction loss \( L_{cyc} \), and the appearance consistency loss \( L_{app} \). We adopt the loss function of LSGAN [43] with two domain discriminators \( D^s \) and \( D^t \) as

\[
L_{adv}^s(G^{t\rightarrow s}, D^s) = \mathbb{E}_{x_t \sim X_t} \left[ \left| \left| D^s(G^{t\rightarrow s}(x_t)) - 1 \right| \right| \right]
+ \mathbb{E}_{x_t \sim X_t} \left[ \left| \left| D^s(G^{t\rightarrow s}(x_t)) - 1 \right| \right| \right]
\tag{2}
\]

\[
L_{adv}^t(G^{s\rightarrow t}, D^t) = \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| D^t(G^{s\rightarrow t}(x_s)) - 1 \right| \right| \right]
+ \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| D^t(G^{s\rightarrow t}(x_s)) - 1 \right| \right| \right]
\]

The cyclic reconstruction loss supervises pixel-level generation by translating the images along two directions twice

\[
L_{cyc}(G^{s\rightarrow t}, G^{t\rightarrow s}) = \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| G^{t\rightarrow s}(G^{s\rightarrow t}(x_s)) - x_s \right| \right| \right]
+ \mathbb{E}_{x_s \sim X_s} \left[ \left| \left| G^{t\rightarrow s}(G^{s\rightarrow t}(x_s)) - x_s \right| \right| \right]
\tag{3}
\]
The appearance consistency loss \[\mathcal{L}_{\text{apr}}(G^{s \rightarrow t}, G^{t \rightarrow s}) = \mathbb{E}_{x^s \sim \mathcal{X}^s} \left[ \|G^{s \rightarrow t}(x^s) - x^t\|_1 \right] + \mathbb{E}_{x^t \sim \mathcal{X}^t} \left[ \|G^{t \rightarrow s}(x^t) - x^s\|_1 \right]. \] (4)

Despite the fact that the above loss terms guide the source-domain images to have target-domain image style and the appearance consistency loss \(\mathcal{L}_{\text{apr}}\) can preserve person appearance to some extent, the generated images generally fail to maintain their original identities, where interclass images cannot be well-separated after translation. To tackle the problem, the existing domain-translation-based methods adopt ID-based constraints to regularize the translation process, e.g., contrastive loss [1], classification loss [3], [4], [18] or triplet loss [5].

However, we observe that “one-hot” ID-based regularizations adopted by the existing works [1]–[5], [18] only require the translated images to be well-separated according to their identities, but are not able to preserve accurate intersample structures and the original source-domain data distributions [Fig. 1(c)]. As a result, the translated images are not accurate enough to optimize the target-domain image encoder. We argue that properly maintaining the original intersample relations is crucial to domain translation.

2) Online Relation-Consistency Regularization: Since there are no “ground-truth” intersample relations, we propose to use the pretrained \(\mathcal{F}\) to provide relation supervision on-the-fly for effectively regularizing translation. Intuitionally, intersample relations can be modeled by the ratio of feature similarities between images. In person re-ID, we found that triplets with intra-/inter-identity samples are generally representative and can be used for modeling intersample relations.

Given a source-domain image \(x^s\), its positive sample \(x^s_p\) with the same identity, and its negative sample \(x^s_n\) with a different identity, we can measure their similarity-based intersample relations on-the-fly among the triplet with a softmax-like function

\[
\mathcal{R}(x^s; \mathcal{F}) = \frac{\exp(f^s_s, f^s_p)}{\exp(f^s_s, f^s_p) + \exp(f^s_s, f^s_n)} \in [0, 1] \quad (5)
\]

where \(f^s_s\), \(f^s_p\), and \(f^s_n\) are the features encoded by the pretrained source-domain encoder \(\mathcal{F}\) on the image samples \(x^s\), \(x^s_p\), and \(x^s_n\), respectively, and \(\langle \ldots, \cdot \rangle\) is the inner product between two feature vectors to measure their similarity. Similar to [42], we use only the most difficult triplet of each sample \(x^s\) within a batch, i.e., the hardest positive (\(f^s_p\)) and negative (\(f^s_n\)) samples for each \(f^s\). Note that \(\mathcal{R}(x^s; \mathcal{F})\) is a continuous value in \([0, 1]\) to measure the ratio of pairwise similarities.

After translating source-domain images to the target domain by \(G^{s \rightarrow t}\), we obtain the features of the source-to-target translated triplet \((f^{s \rightarrow t}_s, f^{s \rightarrow t}_p, f^{s \rightarrow t}_n)\), which are encoded by the target-domain encoder \(\mathcal{F}\) (to be discussed in Section III-C). Similarly, the continuous similarity ratio in \([0, 1]\) between the translated images can also be measured by a softmax-like function as

\[
\mathcal{R}(x^t; G^{s \rightarrow t}, \mathcal{F}) = \frac{\exp(f^{s \rightarrow t}_s, f^{s \rightarrow t}_p)}{\exp(f^{s \rightarrow t}_s, f^{s \rightarrow t}_p) + \exp(f^{s \rightarrow t}_s, f^{s \rightarrow t}_n)} \in [0, 1]. \quad (6)
\]

We claim that if the domain-translation network well preserves the source-domain images’ intersample relations, their softmax-triplet responses before and after translation should be as close as possible. Based on this assumption, a novel relation-consistency loss is introduced to regularize intersample relations after source-to-target translation. The loss is modeled as a “soft” binary cross-entropy loss

\[
\mathcal{L}_{\text{rc}} (G^{s \rightarrow t}) = \mathbb{E}_{x^s \sim \mathcal{X}^s} \left[ \ell_{\text{bce}} (\mathcal{R}(x^s; G^{s \rightarrow t}, \mathcal{F}), \mathcal{R}(x^s; \mathcal{F})) \right] \quad (7)
\]

where \(\ell_{\text{bce}}(p, q) = -q \log p - (1 - q) \log(1 - p)\) with the soft label \(q \in [1, 0]\). In other words, we use source-domain intersample relations \(\mathcal{R}(x^s; \mathcal{F})\) as soft learning targets for supervising translated intersample relations \(\mathcal{R}(x^t; G^{s \rightarrow t}, \mathcal{F})\).

We have also investigated alternative designs of our relation-consistency loss for measuring the intersample
relations and regularizing the translation process, which will be further discussed in Section IV-E2.

3) Differences From Existing ID-Based Regularizations: There are two key differences between the proposed online relation-consistency loss (7) and ID-based regularizations in the existing works, e.g., classification loss [3], [4], [18], contrastive loss [1], or triplet loss.

1) The existing ID-based regularizations only require the samples from different classes to be well-separated after translation, which is too weak to maintain intersample relations and their original data distributions. As long as the translated samples are classified correctly, even if they did not maintain intersample relations well, they receive little penalty. In contrast, our proposed online regularization aims at maintaining continuous and more sensitive relation measurements (5) during domain translation.

2) The existing regularizations use static learning targets (identity labels), while our regularization term generates relation measurements with the current image encoders on-the-fly to provide adaptive supervisions. In other words, previous ones only acquire knowledge from ground-truth labels, while ours exploits valuable knowledge from both ground-truth labels and pretrained source-domain encoder.

Besides the domain-translation-based UDA methods as mentioned above, there exist some works [45]–[47] which leveraged generative models on fully supervised person re-ID tasks. They also focused on preserving person identities with ID-based regularizations, which are too weak to maintain intersample relations as discussed above.

4) Joint Training Objective: During training, we fix \( F^t \) and alternately update \( F^s \) and SDT in each iteration to avoid bias amplification, where the SDT network is optimized with

\[
L_{\text{std}}(G^{s\rightarrow t}, G^{t\rightarrow s}, D^s, D^t) = \lambda_{rc} L_{rc}(G^{s\rightarrow t}) + \lambda_{cyc} L_{cyc}(G^{s\rightarrow t}, G^{t\rightarrow s}) + \lambda_{apr} L_{apr}(G^{s\rightarrow t}, G^{t\rightarrow s}) + \lambda_{adv}(L_{adv}(G^{s\rightarrow t}, D^s) + L_{adv}^t(G^{t\rightarrow s}, D^t)). \tag{8}
\]

Here, \( \lambda_{rc}, \lambda_{cyc}, \lambda_{adv}, \) and \( \lambda_{apr} \) are the weighting factors for different loss terms.

C. Target-Domain Encoder With Translated Images

For training the target-domain encoder \( F^t \) in our framework, arbitrary pseudo-label-based methods (e.g., UDAP [22], MMT [10]) can be adopted as a baseline, and we can improve them by jointly training two domains’ images, i.e., the source-to-target images translated by our SDT and raw target-domain images. Specifically, we can create a unified training image set \( \mathcal{X} = \mathcal{X}^{s\rightarrow t} \cup \mathcal{X}^{t\rightarrow s} \) with a unified label set to supervise a cross-domain identity classifier \( C^t : f \rightarrow \{1, \ldots, p^t + p^s\} \), where both the labeled source-to-target translated images \( \mathcal{X}^{s\rightarrow t} \) and the pseudo-labeled target-domain images \( \mathcal{X}^{t\rightarrow s} \) serve as informative training samples with nonoverlapping real or pseudo identity labels.

Note that pseudo label creation is a general pipeline in UDA tasks and is not the focus of our method. Our framework can also work without pseudo labels (to be discussed in Section IV-D1), i.e., only the labeled source-to-target translated images are adopted for training the target-domain encoder.

Here, we take the modified version of a clustering-based baseline [22] as an example. Note that the original version of [22] only adopted the triplet loss for training, while our modified version adopts both the cross-entropy classification loss and triplet loss to achieve better performance. Target-domain data \( \mathcal{X}^{t\rightarrow s} \) ’s encoded features \( \{f^t\} \) are clustered into \( p^s \) classes and images within the same cluster are assigned the same pseudo label. The target-domain encoder \( F^t \) can then be trained in a fully supervised manner. Specifically, each sample \( x \in \mathcal{X}^{t\rightarrow s} \) is assigned a corresponding label \( y \in \{1, \ldots, p^t + p^s\} \), and \( F^t \) is optimized with the objective function similar to source-domain encoder learning in (1) but with a joint-domain label set

\[
L_{\text{enc}}(F^t, C^t) = \mathbb{E}_{x \sim \mathcal{X}^{t\rightarrow s}}[\ell_{ce}(C^t(f), y)] + \mathbb{E}_{x \sim \mathcal{X}^{t\rightarrow s}}[(\|f - f_p\| + m - \|f - f_y\|)^+]. \tag{9}
\]

The target-domain encoder \( F^t \) can therefore take full advantages of 1) the source-to-target images translated by our \( G^{s\rightarrow t} \), which better maintains their intersample relations and 2) the unified cross-domain label set that consists of both valuable ground-truth source-domain identity labels and target-domain pseudo labels. \( F^t \) trained by this strategy is shown to encode more discriminative features for distinguishing target-domain identities.

In our overall framework, the source-domain encoder \( F^s \) is fixed after pretraining, and the SDT network and the target-domain encoder \( F^t \) alternately promote each other via joint training to achieve optimal re-ID performance. When fixing \( F^t \), it measures translated intersample relations (6) for regularizing SDT via \( L_{\text{enc}} \) (7). When fixing SDT, it generates training samples \( \mathcal{X}^{s\rightarrow t} \) to optimize \( F^t \) (9). Once \( F^t \) is further trained to achieve better re-ID performance on the target domain, it could in turn generate more accurate pseudo labels and measure more accurate intersample relations for further improving SDT.

After training, only \( F^s \) is used to encode target-domain samples into features for pairwise similarity estimation without extra parameters and computational costs. The overall algorithm is summarized in Algorithm 1.
Algorithm 1 SDA for Person Re-ID

Require: Labeled source-domain data \( X_l \), unlabeled target-domain data \( X_t \);
Require: Weighting factors \( \lambda_c, \lambda_{cyc}, \lambda_{adv}, \lambda_{apr} \) for (8);
1: Pretrain source-domain encoder \( F_s \) by minimizing (1) on \( X_l \);
2: Initialize target-domain encoder \( F_t \) by loading the weights of \( F_s \);
3: for \( n \) in \([1, \text{num\_epochs}]\) do
4: Create pseudo labels by clustering \( F_s(X_l) \);
5: for each mini-batch \( B_s \subset X_l, B_t \subset X_t \) do
6: Translate \( B_t \) into the target domain as \( B_t^{s\rightarrow t} \) by \( G_{s\rightarrow t} \);
7: Update \( G_{s\rightarrow t}, G_{t\rightarrow s} \) by minimizing the objective function (8) with \( D^s, D^t \) fixed, where the intersample relations are measured by \( F_s \) and \( F_t \) on-the-fly;
8: Update \( F_t \) by minimizing the objective function (9) with \( B_t^{s\rightarrow t} \bigcup B_s \);
9: Update \( D^s, D^t \) by maximizing the objective function (8) with \( G_{t\rightarrow s}, G_{s\rightarrow t} \) fixed.
10: end for
11: end for

IV. EXPERIMENTS

A. Datasets and Evaluation Metric

We evaluate our framework on four real → real adaptation tasks for person re-ID, including DukeMTMC-reID → Market-1501, Market-1501 → DukeMTMC-reID, Market-1501 → MSMT17, and DukeMTMC-reID → MSMT17 following the experimental setup in state-of-the-arts [8]–[10].

1) DukeMTMC-reID [48] contains 36,411 images of 702 identities for training and another 702 identities for testing, with all the images captured from eight cameras.
2) Market-1501 [49] consists of 12,936 images of 751 identities for training and 19,281 images of 750 identities for testing, which are shot by six cameras.
3) MSMT17 [2] is the most challenging dataset with 126,441 images of 4101 identities from 15 cameras, where 1041 identities are used for training.

Besides the widely used real → real benchmarks above, we also adopt the proposed SDA method in the synthetic → real adaptation task of the Virtual Domain Adaptation Challenge (VisDA-2020).

1) The synthetic source-domain dataset PersonX [50] is generated based on Unity [51] engine, containing 20,280 images out of 700 identities captured by six cameras.
2) The target-domain dataset consists of real-world images captured from five cameras, i.e., 13,198 images for training, 377 images for the query of target_val, 3600 images for the gallery of target_val, 1578 images for the query of target_test, and 24,006 images for the gallery of target_test. Only the train set for the real-world images is used for training. The target_val is used for evaluation offline and the target_test is used for evaluation online.

Note that both the train and test sets for the source and target domains are required in our experiments, where only train sets are used for model optimization and the test sets are used for model evaluation. Specifically, for source-domain pretraining, the source-domain train set is used for training and the source-domain test set is used for evaluation. For our SDA training, the train sets of both the source and target domains are used for training, and the target-domain test set is used for evaluation. Mean average precision (mAP) and cumulative matching characteristics (CMC) accuracies are used to test the methods’ performance.

B. Implementation Details

1) Network Architecture: We adopt ResNet-50 [52] as the backbone for the source-domain and target-domain person image encoders, which are initialized with ImageNet-pretrained [53] weights. We adopt the CycleGAN [6] \( G_{s\rightarrow t} \) and \( G_{t\rightarrow s} \) and the PatchGAN [54] architecture for our discriminators \( D^s \) and \( D^t \). Specifically, each generator consists of three convolution-IN-ReLU blocks, nine residual blocks \( [52] \), two deconvolution-IN-ReLU blocks, and the last one convolution mapping feature maps to RGB images. The target-domain encoder \( F_t \) and SDT network are alternately updated in each iteration to avoid unstable training. Furthermore, we adopt a momentum encoder \( [55] \) (denoted as \( \theta^t \)) to replace \( F_t \) in (6) for measuring more stable triplet relations after domain translation. In particular, we denote the parameters of \( F_t \) and \( F_t^\theta \) as \( \theta^T \) and \( \theta^T_0 \) at iteration \( T \). \( \theta^T_0 \) can be calculated as \( \theta^T_0 = a\theta^T_{T-1} + (1-a)\theta^T_0 \), where \( \theta^T_0 = \theta^T_0 \) and \( a = 0.999 \) is the momentum coefficient. Intuitively, the momentum encoder could provide more reliable intersample relations for regularizing structured domain translation (SDT) since it eases the training bias caused by unstable translation results.

2) Training Data Organization: We adopt a \( PK \) sampler for training, i.e., mini-batch = \( P \) identities \( \times K \) samples. For source-domain pretraining, each mini-batch contains 56 source-domain images of \( P = 8 \) ground-truth identities (\( K = 7 \) for each identity). When jointly training the SDT network and target-domain encoder, each mini-batch contains 56 source-domain images of eight ground-truth identities and 56 target-domain images of eight pseudo identities. The pseudo identities are assigned by the clustering algorithm and updated before each epoch. With such a \( PK \) sampler, the training samples for two domains can be well-balanced even the two domains’ datasets differ a lot regarding their scales. All the images are resized to 256 × 128. Randomly erasing [56], cropping and flipping are applied to each image.

3) Network Optimization: ADAM optimizer is adopted to optimize the networks with weighting factors \( \lambda_c = 1, \lambda_{adv} = 1, \lambda_{cyc} = 10, \) and \( \lambda_{apr} = 0.5 \) and triplet margin \( m = 0.3 \). The initial learning rates (lr) are set to 0.00035 for person image encoders and 0.0002 for the SDT network. The source-domain pretraining iterates for 30 epochs, and the learning rate decreases to 1/10 of its previous value every 10 epochs. The proposed joint training scheme (Algorithm 1) iterates for 50 epochs, where the learning rate is constant for the first 25 epochs and then gradually decreases to 0 for another 25 epochs following the formula \( lr = lr \times (1.0 - \max(0, epoch - 25)/25) \).

C. Comparison With State-of-the-Arts

We compare our proposed SDA framework with the state-of-the-art methods on four domain adaptive re-ID tasks in
Tables I and II. Our method is plug-and-play with the existing pseudo-label-based target domain encoders. Note that the focus of our SDA is to generate informative training samples rather than pseudo label refinery as the previous methods. The translated images by our SDA are used as additional training samples to further improve the pseudo-label-based encoder.

1) Modified UDAP [22] as a Baseline for the Target-Domain Encoder: Following the label generation pipeline introduced by a clustering-based baseline method UDAP [22], we adopted DBSCAN [75] to generate target-domain pseudo labels for our target-domain encoder. Rather than using the sole triplet loss in the original UDAP [22], we modified it using both classification loss and triplet loss as the training objective to achieve better baseline performance, dubbed as “UDAP [22] w/ DBSCAN (modified)” in Table I. The detailed introduction of our modified training objective can be found at (9). Also note that we denote our SDA model as “Our SDA w/ DBSCAN” instead of “Our SDA + UDAP w/ DBSCAN”, since the SDA model is trained based on our modified version of UDAP instead of the original one. Our SDA outperforms the baseline UDAP [22] by large margins, which indicates the effectiveness of our generated source-to-target training samples.

We also tested k-means to generate target-domain pseudo labels for our target-domain encoder on SDA with the optimal k value following the state-of-the-art [10], i.e., 500 for Duke→Market, 700 for Market→Duke, 1500 for Duke→MSMT, and Market→MSMT. The results are denoted as “Our SDA w/ k-means”. The training pipeline is the same as “Our SDA w/ DBSCAN”, except for the clustering algorithm. Our SDA is consistently effective without the need of setting k to be close to the actual identity numbers. As shown in Table III, even with different k’s, our SDA stably improves the already strong baselines, which are trained with only the target-domain samples and clustering pseudo labels [22]. Note that the value of k can be considered as a hyperparameter and can be determined by searching for the optimal performance on the validation set. Selecting proper hyperparameters for clustering-based methods would not limit their usage in practical use.

2) MMT [10] as a Baseline for the Target-Domain Encoder: Although MMT [10] shows superior performances over “Ours SDA w/ k-means” by adopting dual networks with two times more parameters and computations for mutual training, our SDA is well-compatible with it and can be combined to achieve further improvements. Specifically, the target-domain encoder is trained with source-to-target translated images and target-domain raw images under the training pipeline of MMT, where the source-to-target images are translated by SDA. Both soft losses and hard losses in MMT are adopted for training.
The combination “Our SDA w/ k-means+MMT [10]” shows further 4.3% and 5.3% mAP gains on Duke→Market and Market→Duke. Note that adopting the existing domaintranslation-based methods (e.g., SPGAN [1]) on top of MMT [10] even degrades the performance, since MMT [10] itself is already very strong and can only be boosted by informative enough translated images.

D. Comparison With Domain Translation-Based Methods

The previous translation-based methods [1]–[4] did not use pseudo labels and therefore cannot be directly compared. We carefully design comparative experiments to verify the importance of the proposed online relation-consistency regularization term.

1) Our Framework Without Pseudo Labels: We evaluate our framework using a target-domain encoder without pseudo labels, which is a common strategy in previous translation-based methods [1], [2], [4], [76], i.e., the target-domain encoder is trained with only source-to-target translated images and their source-domain identities. As shown in Table IV, our method stably outperforms the existing domain translation-based methods in terms of mAP on both Duke→Market and Market→Duke adaptation tasks without generating pseudo labels in the target domain. Specifically, we outperform state-of-the-art method CGAN-TM [5] by 3.7% and 1.4% mAP on Duke→Market and Market→Duke, respectively. Note that CGAN-TM [5] achieved slightly better top-1/5/10 performance than our method on Market→Duke since they adopted a deeper DenseNet-121 as the backbone, compared with our plain ResNet-50 backbone.

To further verify the effectiveness of our introduced relation regularization term, we conduct an experiment by removing $L_{rc}$ from “Our SDA w/o pseudo labels”, dubbed “Our SDA w/o pseudo labels (w/o $L_{rc}$)” in Table IV. We observe significant performance drops when discarding such a relation regularization term from training.

2) Existing ID-Based Regularizations in Our Framework: The existing translation-based UDA methods [1]–[4] adopted identity-based losses with static targets to regularize domain translation, including contrastive loss in SPGAN [1], classification loss in eSPGAN [3] and CR-GAN [4], and triplet loss in CGAN-TM [5]. Generally, previous losses only require the source-to-target translated images to be correctly classified after translation, but ignore intersample relations and similarities in their original domain. Since our pseudo-label-based target-domain encoder shows much better baseline performance than theirs, for fair comparison, we replace the online relation regularization $L_{rc}$ (7) in our framework with the previous methods’ ID-based regularizations to demonstrate the effectiveness of our regularization.

The results are reported in Table V. We observe that replacing our proposed $L_{rc}$ with previous regularization (denoted as “$L_{rc}$ → ID-based contrastive loss [1]”, “$L_{rc}$ → ID-based classification loss [3], [4], [18]”, “$L_{rc}$ → ID-based triplet loss [5]”, and “$L_{rc}$ → ID-based classification loss & triplet loss”), all lead to worse performances than our method, demonstrating the superiority of our stronger relation regularization term over weaker regularizations in previous domain-translation-based UDA methods.

3) Domain Translation Examples: Besides the example triplets shown in Fig. 1 (c), we also illustrate several translated examples of CycleGAN [6], SPGAN [1], and our method in Fig. 3. SPGAN adopts ID-based regularizations (i.e., contrastive loss), showing inferior generation results than our method. ID-based regularizations are too weak to preserve intersample relations during translation. For instance, the man in the first row appears to be in different colors (e.g., orange, yellow and green) within a tuple after translation by CycleGAN and SPGAN.

Besides the illustration of the generated images, we also evaluate the quality of translated samples by calculating the
TABLE IV

Comparison With Domain Translation-Based UDA Methods Using Target-Domain Encoder without Pseudo Labels.

| Methods w/o Pseudo Labels | DukeMTMC-reID→Market-1501 | Market-1501→DukeMTMC-reID |
|---------------------------|---------------------------|---------------------------|
|                           | mAP | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 |
| PTGAN [2] (CVPR’18)       | -   | 38.6  | -     | -      | -   | 27.4  | -     | -      |
| SPGAN [1] (CVPR’18)       | 22.8| 51.5  | 70.1  | 76.8   | 22.3| 41.1  | 56.6  | 63.0   |
| CR-GAN [4] (ICCV’19)      | 29.6| 59.6  | -     | -      | 30.0| 52.2  | -     | -      |
| SBSGAN [17] (ICCV’19)     | 27.3| 58.5  | -     | -      | 30.8| 53.5  | -     | -      |
| CGAN-TM [5] (TIP’20) (DenseNet-121) | 31.3| 61.4  | 78.4  | 84.9   | 32.9| 54.9  | 68.8  | 74.3   |
| Our SDA w/o pseudo labels | 35.0| 64.5  | 79.5  | 84.6   | 34.3| 53.1  | 67.1  | 72.4   |
| Our SDA w/o pseudo labels (w/o \(\mathcal{L}_{rc}\)) | 31.0| 59.9  | 75.2  | 82.0   | 30.2| 51.2  | 66.0  | 70.9   |

TABLE V

Comparison Between Our Online Relation-Consistency Regularization \(\mathcal{L}_{rc}\) and ID-Based Regularizations in Previous Domain-Translation-Based Methods. \(k\)-Means Algorithm Is Adopted Here to Generate Pseudo Labels.

| Methods               | DukeMTMC-reID→Market-1501 | Market-1501→DukeMTMC-reID |
|-----------------------|---------------------------|---------------------------|
|                       | mAP | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 |
| \(\mathcal{L}_{rc}\) \(\to\) ID-based contrastive loss [1] | 56.5| 78.7  | 91.5  | 94.0   | 51.7| 69.9  | 80.5  | 85.3   |
| \(\mathcal{L}_{rc}\) \(\to\) ID-based classification loss [3], [4], [18] | 63.4| 84.9  | 92.7  | 95.1   | 53.8| 70.9  | 81.9  | 85.6   |
| \(\mathcal{L}_{rc}\) \(\to\) ID-based triplet loss [5] | 64.1| 85.2  | 92.9  | 95.4   | 54.5| 72.1  | 82.3  | 85.9   |
| \(\mathcal{L}_{rc}\) \(\to\) ID-based classification loss & triplet loss | 64.5| 85.0  | 93.1  | 95.6   | 54.7| 72.6  | 82.4  | 85.9   |
| Our SDA w/ \(\mathcal{L}_{rc}\) | 66.4| 86.4  | 93.1  | 95.6   | 56.7| 74.0  | 84.1  | 87.7   |

Fig. 3. Domain-translated examples of CycleGAN [6], SPGAN [1], and our method. Note persons on each row are of the same identity. Other translation-based methods [2]–[4] did not provide trained models or translated images, and thus not illustrated here. However, we have carefully discussed and compared them with their ID-based regularizations in Sections IV-D1 and IV-D2. Best viewed in color.

TABLE VI

Evaluation of Generated Images by Domain-Translation-Based Methods for UDA Person Re-ID in Terms of the FID Score [77].

|                           | SPGAN [1] | Ours |
|---------------------------|-----------|------|
| Market-1501→DukeMTMC-reID | 47.6      | 77.6 |
| DukeMTMC-reID→Market-1501 | 46.1      | 46.0 |

FID score [77]. Specifically, we use FID metric to compute the similarity between the source-to-target translated dataset and the raw target-domain dataset. As demonstrated in Table VI, our method considerably outperforms SPGAN [1] on the task of Market→Duke. Note that the FID score can only evaluate the quality of generated images, but cannot evaluate the preserved intersample relations and affinities. One possible way to evaluate the preserved intersample relations of translated images is to use them as additional training samples for re-ID training and then evaluate the trained model by the re-ID metric, just as what we did in the above sections.

E. Ablation Studies

We conduct ablation studies on Duke→Market and Market→Duke tasks to analyze the effectiveness of each component in our framework. Detailed ablation experiments can be found in Table VII.

1) Effectiveness of Source-to-Target Translated Images: We treat the target-domain encoder \(\mathcal{F}^t\) trained with only target-domain images and clustering-based pseudo labels as our baseline model, which can be treated as a reproduction of UDAP [22]. Our framework significantly outperforms the baseline model by properly exploiting valuable source-domain data (see “Ours (full)” versus “Baseline” in Table VII).

A naïve way to use source-domain images is to directly train on both domains’ raw images, denoted as “Baseline+raw source-domain data”. The performance is even worse than the baseline on Duke→Market due to the large domain gaps, which indicates the necessity of properly leveraging different domains’ images.
Our SDA can also be integrated and benefit the state-of-the-art pseudo-label-based method [10] (see “Our SDA+MMD” versus “MMD” in Table I). The improvements demonstrate the effectiveness of source-to-target images translated by our structured domain translation method.

2) Alternative Designs of Online Relation-Consistency Regularization: Our SDT applies regularizations on the softmax-triplet relations (7). We further explore two alternative forms, prediction-consistency regularization \( L_{pc} \) and batch-all relation-consistency regularization \( L_{brc} \) (Table VIII), to verify the effectiveness of our proposed intersample relation constraint.

Specifically, prediction-consistency regularization ensures that each individual image in the source domain should maintain the same “soft” class prediction after source-to-target translation. The loss function is formulated as

\[
L_{pc}(G; x^t) = \mathbb{E}_{x^s \sim X^s} \left[ -C_t(f^s) \cdot \log(C_t(f_{x^s \rightarrow x^t})) \right].
\]

As shown in Table VIII, 3.1% and 2.4% mAP drops are observed on the two tasks. Here, prediction-consistency regularization \( L_{pc} \) is different from conventional ID-based classification regularization [3], [4], [18], which can be formulated as \( \mathbb{E}_{x^s \sim X^s} \left[ -y^s \cdot \log(C_t(f_{x^s \rightarrow x^t})) \right] \) with “one-hot” ground-truth label \( y^s \). Briefly, \( L_{pc} \) is supervised by “soft” \( C_t(f^s) \), while ID-based classification regularization is supervised by “hard” \( y^s \). \( C_t(f^s) \) is predicted by the source-domain encoder, which has captured the original intersample relations in the latent space. Note in this work, we denote the detailed data distributions and structured relations among data points, rather than simple positive and negative identity relations, as intersample relations. \( L_{pc} \) can better preserve the intersample relations, while ID-based classification regularization can only ensure that images belonging to different identities are well-separated.

Our \( L_{brc} \) (7) aims at preserving relations within hardest triplets, while the alternative batch-all relation-consistency loss \( L_{brc} \) tries to preserve all available relations within batches, no matter whether they are easy or hard. The batch-all regularization term follows the similarity-preserving loss in [40], which preserves intersample affinities in knowledge distillation. We model the batch-all intersample relations by measuring the similarities

\[
R(x^i; F) = \text{softmax}\left[ (f^s, f^t), \ldots, (f^s, f^t) \right].
\]

The term consists of pairwise dot products between each sample \( x^s \) and all other ones in the same batch. The similarity vector is normalized by a softmax function, and a soft cross-entropy loss is adopted to regularize all the relations after translation

\[
L_{brc}(G_{x^i \rightarrow x^t}) = \mathbb{E}_{x^t \sim X^t} \left[ -R(x^s; F^t) \cdot \log R(x^s; G_{x^i \rightarrow x^t}, F^t) \right].
\]

1.8% mAP and 2.1% mAP drops can be observed on the two tasks. The reason might be that batch-all relations contain many easy cases that cannot provide effective supervisions for training.

To show the necessity of adopting relation regularization during translation, we also tested totally removing \( L_{pc} \) from our framework, dubbed “Ours w/o relation-consistency loss \( L_{pc} \)” in Table VII. Significant mAP decreases of 3.4% and 3.8% are observed on Duke→Market and Market→Duke tasks. We observe that replacing our relation-consistency loss \( L_{pc} \) with either conventional ID-based constraints (Table V) or alternative designs of relation-consistency regularization (Table VIII) would achieve only slight improvements over “Ours w/o relation-consistency loss \( L_{pc} \)”. This is because pseudo-label-based baseline is already strong enough to achieve satisfactory performance on the target domain. Further improvements are quite challenging and require informative enough translated images as additional training samples. The comparison results well validate the effectiveness of our proposed relation-consistency regularization.

3) Effectiveness of Training With the Unified Label Set:

We observe that the target-domain encoder also benefits from the unified label set by training the classifier on all the \( p^t + \hat{p}^t \) classes across the two domains. To show it, we design an experiment with separate classifiers for source-to-target translated images and target-domain images, i.e., \( C^t : f \rightarrow \{1, \ldots, p^t + \hat{p}^t\} \) is split into \( C^t_{x^i \rightarrow x^t} : f_{x^i \rightarrow x^t} \rightarrow \{1, \ldots, p^t\} \) and \( C^t : f^t \rightarrow \{1, \ldots, \hat{p}^t\} \). We report the performance in Table VII as “Ours w/o unified label system”. 1.6% and 1.9% mAP drops are observed on the two tasks, which indicate the effectiveness of our proposed relation-consistency regularization.

4) Further Benefits From the Momentum Encoder \( \mathcal{F}^t_s \):

As described in Section IV-B, we use a momentum encoder [55] for more stable training and better performance. To verify that the main contribution is not from the momentum encoder, we perform an experiment by removing \( \mathcal{F}^t_s \) while keeping...
all other components unchanged. The experimental results are denoted as “Ours w/o momentum encoder [55]” in Table VII. We observe slight drops of 1.1% and 1.4% mAP on two tasks.

5) Performance w/o Joint Training of Domain-Translation Network and Target-Domain Encoder: The domain-translation network and target-domain encoder in our framework promote each other via joint training. However, a simpler training scheme would be to first train a source-to-target translation network with the proposed regularization and to translate all the source-domain images into the target domain. A pseudo-label-based target-domain encoder is then trained with such fixed source-to-target translated images and target-domain images. Note that in this scheme, a target-domain encoder pretrained with only unlabeled data and clustering labels is used for regularizing SDT training.

We evaluate both our framework and existing translation-based methods [1], [6] when adopting such a separate training strategy and k-means clustering for pseudo label generation. The results in Table IX show that informative training samples generated by our proposed SDT network could effectively improve the already strong baseline even without our joint training scheme, while the source-to-target images generated by CycleGAN and SPGAN might even worsen the performance because their generated images might not well capture the distributions of target-domain data and maintain their original intersample relations.

6) Hyperparameter Analysis: There are five hyperparameters in our training scheme (Section IV-B), including the loss weights $\lambda_{rc}$, $\lambda_{adv}$, $\lambda_{cyk}$, $\lambda_{apr}$, and the triplet margin $m$. Note that the values of $\lambda_{adv}$, $\lambda_{cyk}$, and $\lambda_{apr}$ are directly inherited from CycleGAN [6], and the setting of $m = 0.3$ is widely acknowledged in re-ID-related tasks [10], [78]. We therefore conduct the experiments of hyperparameter analysis on $\lambda_{rc}$, which is also the most important factor in our introduced SDT network. As shown in Fig. 4, our proposed model is robust when the value of $\lambda_{rc}$ varies from 0.1 to 2.0.

F. Application of SDA in VisDA-2020 Challenge

VisDA-2020 Challenge introduces a synthetic→real adaptation task, where labeled synthetic images and unlabeled real-world images are provided. The final performances on the test set of the real-world domain are used for ranking the teams in the challenge. Note that the model’s performance on the test set is evaluated on online server and the ground-truth labels cannot be accessed. We therefore tune our models on the validation set during the challenge and also report the performance on the validation set for comparison in this article.

We propose to adopt the SDT network to translate the synthetic images to have real-world style. Specifically, we adopt the “w/o joint training” version of SDA as mentioned in Section IV-E5 with three main training steps: 1) Pretraining the source-domain encoder with only labeled data; 2) Pretraining the target-domain encoder with only unlabeled data and clustering labels; 3) Training the SDT network with the online relation-consistency regularization provided by the two encoders on-the-fly via (8). The main reasons why we did not use the “joint training” version of SDA for the competition is that increasing model sizes and numbers of parameters can result in more gains on the dataset, which is important in the competition. “w/o joint training” can save GPU memory and allow training much larger deep models.

After being trained, the SDT network can be adopted to generate informative training samples with ground-truth identities by translating all the source-domain images into the target domain. To verify the effectiveness of our SDT, we adopt a target-domain encoder with a ResNet50-IBN backbone. The encoder is trained with only synthetic-to-real translated images and their ground-truth identities, and then tested on the validation set of the target domain. As shown in Table X, training with “synthetic-to-real images by SDT” could achieve much more performance gain than training with “synthetic-to-real images by SPGAN”. Besides quantitative comparison, we also visualize the translation results in Fig. 5, where our SDT could better preserve the consistent appearance for person images of the same identity in both the cases.

In our solution to VisDA-2020 Challenge, multimodel ensemble and an improved MMT [10], [19] are further introduced to fine-tune the network with both synthetic→real...
V. DISCUSSION AND CONCLUSION

In this work, we propose an end-to-end SDA framework with a novel online relation-consistency regularization term to tackle the UDA problem for person re-ID. The structured translated images in our method are shown to be informative samples for improving the training of pseudo-label-based encoder. The joint optimization scheme of the domain-translation network and re-ID encoder is effective; however, it still has difficulty in handling industrial-scale datasets. Further improvements are called for. Beyond the person re-ID, our proposed intersample relation-consistency regularization may benefit other related UDA tasks.

TABLE IX

| Methods w/o Joint Training | DukeMTMC-reID→Market-1501 | Market-1501→DukeMTMC-reID |
|---------------------------|---------------------------|---------------------------|
|                           | mAP | top-1 | mAP | top-1 | mAP | top-1 |
| Baseline (target-domain data + pseudo labels) | 50.1 | 68.2 | 79.2 | 82.7 |
| Base. + source-to-target data by CycleGAN | 56.0 | 79.6 | 90.6 | 93.9 | 51.2 | 69.5 | 80.4 | 83.5 |
| Base. + source-to-target data by SPGAN [1] | 53.4 | 78.6 | 90.3 | 93.1 | 48.8 | 66.2 | 78.5 | 83.0 |
| Base. + source-to-target data by our SDF | 61.3 | 83.3 | 91.8 | 94.9 | 54.3 | 71.6 | 82.0 | 85.6 |

images translated by SDF and raw unlabeled real-world images, achieving superior performance, as illustrated in Fig. 6. The gain of our proposed SDF over the new baseline model and new adaption task demonstrates its generalization capability. Our final solution ranks second out of 153 teams.

TABLE X

| Images for training | mAP | top-1 |
|--------------------|-----|------|
| Raw synthetic images | 61.0 | 71.6 |
| Synthetic-to-real images by SPGAN [1] | 68.2 | 75.1 |
| Synthetic-to-real images by SDF | 71.2 | 79.3 |

REFERENCES

[1] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 994–1003.

[2] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer GAN to bridge domain gap for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 79–88.

[3] W. Deng, L. Zheng, Q. Ye, Y. Yang, and J. Jiao, “Similarity-preserving image-image domain adaptation for person re-identification,” 2018, arXiv:1811.10551.

[4] Y. Chen, X. Zhu, and S. Gong, “Instance-guided context rendering for cross-domain person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 232–242.

[5] Y. Tang, X. Yang, N. Wang, B. Song, and X. Gao, “CGAN-TM: A novel domain-to-domain transferring method for person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 5641–5651, 2020.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2223–2232.

[7] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 14, no. 4, pp. 1–18, 2018.

[8] X. Zhang, J. Cao, C. Shen, and M. You, “Self-training with progressive augmentation for unsupervised cross-domain person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8222–8231.

[9] Y. Fu et al., “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6112–6121.

[10] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refine for unsupervised domain adaptation on person re-identification,” in Proc. Int. Conf. Learn. Represent., 2020, pp. 1–15.

[11] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, “VehicleNet: Learning robust visual representation for vehicle re-identification,” IEEE Trans. Multimedia, vol. 23, pp. 2683–2703, 2021.

[12] H. Luo et al., “A strong baseline and batch normalization neck for deep person re-identification,” IEEE Trans. Multimedia, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[13] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “GLAD: Global–local-alignment descriptor for scalable person re-identification,” IEEE Trans. Multimedia, vol. 21, no. 4, pp. 986–999, Apr. 2019.

[14] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, “Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5177–5186.

[15] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, “Unsupervised person re-identification via cross-camera similarity exploration,” IEEE Trans. Image Process., vol. 29, pp. 5481–5490, 2020.

[16] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, “Progressive learning for person re-identification with one example,” IEEE Trans. Image Process., vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

[17] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, “SBSSGAN: Suppression of inter-domain background shift for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9527–9536.

[18] Y. Zou, X. Yang, Z. Yu, B. Kumar, and J. Kautz, “Joint disentangling and adaptation for cross-domain person re-identification,” in Proc. Eur. Comput. Vis. Conf., 2020, pp. 1–14.

[19] Y. Ge, S. Yu, and D. Chen, “Improved mutual mean-teaching for unsupervised domain adaptive re-ID,” 2020, arXiv:2008.10313.

[20] F. Yang, Z. Zhong, Z. Luo, S. Lian, and S. Li, “Leveraging virtual and real person for unsupervised person re-identification,” IEEE Trans. Multimedia, vol. 22, no. 9, pp. 2444–2453, Sep. 2020.

[21] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, “Feature affinity-based pseudo labeling for semi-supervised person re-identification,” IEEE Trans. Multimedia, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.

[22] L. Song et al., “Unsupervised domain adaptive re-identification: Theory and practice,” Pattern Recognit., vol. 102, Jun. 2020, Art. no. 107173.

[23] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, “Unsupervised person re-identification by soft multilabel learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2148–2157.

[24] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, “Invariance matters: Exemplar memory for domain adaptive person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 598–607.

[25] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable feature with deep adaptation networks,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 97–105.

[26] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands: Springer, 2016, pp. 443–450.
270

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 35, NO. 1, JANUARY 2024

[27] E. Zheng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7167–7176.

[28] D. Das and C. S. George Lee, “Unsupervised domain adaptation using regularized hyper-graph matching,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 3758–3762.

[29] Q. Xie, W. Zhou, G.-J. Qi, Q. Tian, and H. Li, “Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization,” IEEE Trans. Multimedia, vol. 23, pp. 597–610, 2021.

[30] F. Yang et al., “Part-aware progressive unsupervised domain adaptation for person re-identification,” IEEE Trans. Multimedia, vol. 23, pp. 1681–1695, 2021.

[31] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Unsupervised domain adaptation by backpropagation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 153–168.

[32] J. Hoffman et al., “CyCADA: Cycle-consistent adversarial domain adaptation,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 1989–1998.

[33] Y. Li, H. Yuan, X. Wang, and X. Liu, “Bidirectional learning for domain adaptation of semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6936–6945.

[34] Y.-C. Chen, M.-H. Yang, and J.-B. Huang, “CrDeCo: Pixel-level domain transfer with cross-domain consistency,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1791–1800.

[35] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[36] B. Shaw and T. Jebara, “Structure preserving embedding,” in Proc. 26th Ann. Int. Conf. Mach. Learn. (ICML), 2009, pp. 937–944.

[37] F. Tung and G. Mori, “Similarity-preserving knowledge distillation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1365–1374.

[38] W. Park, D. Kim, Y. Lu, and M. Cho, “Relational knowledge distillation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3967–3976.

[39] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, arXiv:1703.07737.

[40] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2794–2802.

[41] Y. Taiyan, A. A. Poljak, and L. Wolf, “Unsupervised cross-domain image generation,” in Proc. Int. Conf. Learn. Represent., 2017, pp. 1–8.

[42] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, “Pose transferrable person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4099–4108.

[43] Y. Ge et al., “FD-GAN: Pose-guided feature distilling GAN for robust person re-identification,” in Proc. Adv. Neural Inf. Process. Syst., vol. 2018, pp. 1223–1233.

[44] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2138–2147.

[45] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 17–35.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1116–1124.

[47] X. Sun and L. Zheng, “Dissecting person re-identification from the viewpoint of viewpoint,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 608–617.

[48] J. Ricciotti, “John ricciotti sets out to identify the engine of growth for unity technologies (interview),” in VentureBeat, Interview With Dean Takahashi, vol. 18, no. 3, VentureBeat, 2015. [Online]. Available: https://venturebeat.com/2014/10/23/john-ricciotti-sets-out-to-identify-the-engine-of-growth-for-unity-technologies-interview
GE et al.: SDA WITH ONLINE RELATION REGULARIZATION

[75] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. 2nd Int. Conf. Knowl. Discovery Data Mining, 1996, pp. 226–231.

[76] W. Liang, G. Wang, J. Lai, and J. Zhu, “M2M-GAN: Many-to-many generative adversarial transfer learning for person re-identification,” 2018, arXiv:1811.03768.

[77] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.

[78] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 1–8.

[79] X. Pan, P. Luo, J. Shi, and X. Tang, “Two at once: Enhancing learning and generalization capacities via IBN-Net,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 464–479.

[80] X. Zhu, Z. Luo, P. Fu, and X. Ji, “VOC-ReID: Vehicle re-identification based on vehicle-orientation-camera,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2020, pp. 602–603.

[81] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with K-reciprocal encoding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1318–1327.

Yixiao Ge received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2017, and the Ph.D. degree from the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, in 2021. She is with the ARC Lab of Tencent PCG, Hong Kong. Her research interests include computer vision and deep learning, with focus on representation learning.

Feng Zhu received the B.S. degree in electronic information engineering and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China, Hefei, China, in 2011 and 2017, respectively. He is with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. His research interests include computer vision, face analysis, and machine learning.

Dapeng Chen (Member, IEEE) received the B.S. degree in biomedical engineering and the Ph.D. degree in control science and engineering from Xi’an Jiaotong University, Xi’an, China, in 2010 and 2016, respectively. He is with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. His research interests include computer vision, machine learning, and large-scale retrieval/clustering.

Rui Zhao (Member, IEEE) received the B.S. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong, Hong Kong, in 2010 and 2015, respectively. He is with the Multimedia Laboratory, The Chinese University of Hong Kong. His research interests include computer vision and machine learning, with focus on face analysis, deep learning, and visual surveillance.

Xiaogang Wang (Member, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2001, the M.S. degree from The Chinese University of Hong Kong, Hong Kong, in 2003, and the Ph.D. degree from the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009. He is a Professor with the Multimedia Laboratory, The Chinese University of Hong Kong. His research interests include computer vision and machine learning.

Hongsheng Li (Member, IEEE) received the B.S. degree in automation from the East China University of Science and Technology, Shanghai, China, and the M.S. and Ph.D. degrees in computer science from Lehigh University, Bethlehem, PA, USA, in 2006, 2010, and 2012, respectively. He is an Assistant Professor with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. His research interests include computer vision, medical image analysis, and machine learning.