ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240306001382
Original scientific paper

Bidirectional Image Translation for Robust Person Re-Identification in Unsupervised Domain Adaptation

Xiaohu HE*, Jing LIU

Abstract: Unsupervised domain adaptation in person re-identification (ReID) poses significant challenges due to domain shift, often leading to substantial performance degradation when models are deployed in new environments. To address this issue, we propose a novel framework that synergistically combines a bidirectional image translation network with a feature extraction network, augmented by an innovative id consistency loss. Our method leverages generative adversarial networks for image translation, ensuring style consistency between source and target domains, while preserving identity features. Extensive experiments on benchmark datasets demonstrate the superiority of our approach, setting new state-of-the-art results on multiple challenging tasks. The proposed framework represents a significant advancement in unsupervised domain adaptive person re-identification, with potential implications for real-world surveillance and security applications.

Keywords: domain adaptation; domain transfer; generative adversarial network; person re-identification

1 INTRODUCTION

In practical applications, the significance of person re-identification (ReID) systems is undeniable. They play a crucial role in the field of computer vision, seeking to match images of individuals across various camera views. However, despite notable advancements, the performance of ReID systems frequently experiences a sharp decline when deployed in novel, unfamiliar environments. This occurrence, termed domain shift, arises from disparities in data distributions between the source domain (where the model is trained) and the target domain (where the model is applied).

To illustrate the impact of domain shift, we present a comparison of ReID performance on the Market-1501 dataset when it is used as both the source and target domain, versus when it is the source domain and MSMT17 is the target domain. The results, shown in Tab. 1, highlight the stark contrast in model effectiveness due to domain shift.

Table 1 Performance degradation due to domain shift. The left side shows results within the same domain (Market-1501), while the right side shows results when transferring from Market-1501 to MSMT17

Market-1501→ Market-1501		Market-1501 → MSMT17	
mAP	Rank-1	mAP	Rank-1
92.6	96.7	16.6	38.0

As the table demonstrates, while the model achieves impressive results within the same domain, with a mean Average Precision (mAP) of 92.6% and a Rank-1 accuracy of 96.7%, its performance plummets when applied to a new domain, with the mAP dropping to 16.6% and Rank-1 accuracy to 38.0%. This significant drop in performance underscores the challenge of domain shift in ReID tasks.

Unsupervised Domain Adaptation (UDA) emerges as a necessary field of study to address the challenge of adapting a model trained on a labeled source domain to an unlabeled target domain, thereby mitigating the effects of domain shift. Specifically, many traditional methods commence by leveraging label prediction algorithms on unlabeled data to generate pseudo-labels for each instance. By leveraging UDA techniques, we can develop ReID systems that maintain high levels of accuracy even when deployed in new environments, making them more robust and practical for real-world applications. However, traditional UDA algorithms have limitations, particularly in handling the complex dynamics of real-world environments. These algorithms often struggle with the clustering process, crucial for generating reliable pseudo-labels. The clustering of unlabeled data can be imprecise due to the inherent variability and ambiguity in the data, which may not be well-represented by the source domain. The reliance on distance metrics for pseudo-label assignment can inadvertently introduce inaccuracies, as the nearest neighbors may not always belong to the same identity. This is particularly problematic in complex visual scenes with significant intra-class variations and inter-class similarities. The limitations of traditional clustering processes in UDA contribute to this issue, as they may fail to accurately capture the true relationships between different identities, further exacerbating the noise problem. Consequently, the noise in pseudo-labels can propagate through the training process, leading to a model that reinforces these inaccuracies, ultimately compromising the robustness and reliability of the ReID system.

To address the aforementioned challenges, we introduce a novel approach in this paper that circumvents the need for pseudo-labels entirely. Our method harnesses the power of generative adversarial networks to perform mutual image translation between the source and target domains. By doing so, we transform the problem of unsupervised domain adaptation into a supervised learning task, where the domain-translated images are paired with their corresponding labels from the source domain.

Specifically, the foundation of our approach is the synergistic integration of two neural networks: an image translation network and a feature extraction network, enhanced by the novel id consistency loss. The image translation network, based on a Generative Adversarial Network (GAN) architecture [18], is adept at translating images between domains while preserving the person's identity. This bidirectional translation is pivotal, enabling the feature extraction network to learn from images that are stylistically aligned with the target domain. The feature extraction network is fine-tuned to distill discriminative features from both original and translated images, thus

transforming the unsupervised domain adaptation challenge into a supervised learning paradigm without relying on potentially noisy pseudo-labels. During training, our adversarial training strategy concurrently refines both networks. The image translation network employs adversarial and cycle-consistency losses to ensure content preservation in translated images. The feature extraction network leverages a combination of triplet loss and cross-entropy loss to differentiate between identities. For testing, we preprocess the test images through the image translation network to align them with the training domain, ensuring consistency and improving the reliability of the feature extraction.

A key innovation in our method is the use of the perceptual loss and the id consistency loss during the training of the image translation network. On the one hand, the perceptual loss function helps the network to capture high-level semantic information, which is essential for maintaining the identity of the person during the image translation process. On the other hand, the introduction of the id consistency loss during training ensures that the identity is maintained throughout the image translation process, complementing the perceptual loss that captures high-level semantic information. This dual loss strategy results in translated images that are not only visually indistinguishable from the target domain but also retain critical identity features for accurate person re-identification.

Our experiments confirm the superiority of our method. On the DukeMTMC-reID to Market-1501 adaptation, our method achieves anmAP of 83.7% and a Rank-1 accuracy of 93.6%. For the reverse adaptation, we attain anmAP of 73.1% and a Rank-1 accuracy of 84.2%. Furthermore, on the challenging Market-1501 → MSMT17 [14] and DukeMTMC-ReID→MSMT17 [14] tasks, our method demonstrates its robustness with anmAP of 35.8% and 36.5%, and Rank-1 accuracies of 67.2% and 68.1%, respectively, setting new benchmarks for unsupervised domain adaptive person re-identification.

To succinctly encapsulate the innovative aspects of our research, we highlight the three primary contributions that distinguish our approach in the field of unsupervised domain adaptive person re-identification:

- Bidirectional Image Translation Network: Utilizing a generative adversarial network, our framework performs mutual image translation between the source and target domains. This not only ensures style consistency but also preserves the identity features across domains, effectively transforming unsupervised domain adaptation challenges into a more manageable supervised learning task.
- Innovative ID Consistency Loss: We introduce an id consistency loss that works in tandem with the perceptual loss to ensure that the identity features are not just preserved but are consistent across the translated images. This dual loss strategy is crucial for maintaining high-level semantic information and identity accuracy during the image translation process.
- Synergistic Integration of Networks: The seamless integration of the image translation network with the ReID feature extraction network allows for the extraction of robust and discriminative identity features from both original and domain-translated images. This integration enhances the adaptability and effectiveness of the ReID system under varying domain conditions without relying on noisy pseudo-labels.

2 RELATED WORK

In this section, we will explore the field of ReID focusing on the strides and stumbles in domain generalization and unsupervised domain adaptation. As ReID technologies evolve, the challenge of generalizing across diverse environments and reducing reliance on extensive labelled datasets has prompted innovative approaches in both domains.

2.1 Domain Generalization Person Re-identification

Despite significant advancements in person re-identification (ReID) with the advent of deep learning, the generalization ability of existing models across different scenarios remains limited. This limitation is primarily due to the inherent complexity of the task and the scarcity of large-scale labelled training data. To overcome these challenges, researchers have proposed solutions such as transfer learning and unsupervised domain adaptation, which have become mainstream research directions in ReID. While these methods have shown promise, they still necessitate extensive data collection in each application scenario for deep learning training, which is both time-consuming and labor-intensive.

Normalization techniques such as batch normalization (BN) [32] and instance normalization (IN) [33] have been explored to improve the generalization ability of deep models. For instance, Jia et al. [34] proposed a domain generalization ReID method based on normalization, which alleviated domain style and content bias, enhancing model robustness. However, these normalization techniques sometimes fail to completely eliminate the domain discrepancies, especially under complex environmental variations. Jin et al. [35] introduced a style-normalization and restitution (SNR) module that filtered out interference from style changes and restored identity-related features discarded by instance normalization, yet the challenge of maintaining high accuracy in diverse settings persists. Zhou et al. [36] proposed a lightweight CNN architecture, OSNet, for learning omni-scale feature representations in ReID, which offers scalability but may struggle with extreme variations in data distribution. Choi et al. [37] proposed a ReID framework, MetaBIN, that generalized normalization layers by pre-simulating unsuccessful generalization scenarios in the meta-learning process, though it requires careful tuning of simulation parameters. Jiao et al. [38] proposed dynamically transformed instance normalization (DTIN), which used dynamic convolution to allow non-normalized features to control the transformation of normalized features to a new representation, introducing adaptability but at the cost of increased computational complexity.

2.2 Unsupervised Domain Adaptive ReID

The task of person re-identification (ReID) has seen significant progress with the advent of deep learning, but

the requirement for manual annotation in multi-camera videos remains a challenge. To address this, recent research has focused on unsupervised ReID algorithms, which leverage pre-trained models, data augmentation strategies, and pseudo-label generation methods to reduce the need for manual annotation and improve performance. Pseudo-label generation methods, such as those based on K-nearest neighbors [19, 20, 39] and K-means clustering [6, 21, 28], have been widely used. These methods generate pseudo-labels for unlabeled data, which are then used to train the model. However, the presence of noise in pseudo-labels can significantly affect the training process. To mitigate this, researchers have proposed various methods to improve the robustness of the training process to label noise, such as model co-training [22-24] and mean teacher models [6, 25-28]. Despite these advancements, the accuracy and reliability of pseudo-labels remain problematic, often leading to suboptimal model performance.

Domain transfer methods have also been proposed to leverage labelled data from other scenes and transfer this information to the target scene. These methods often use generative adversarial networks (GANs) to transfer the style of labelled images from other scenes to the target scene [29, 30, 31, 39]. While effective, these GAN-based approaches require significant computational resources and can be unstable during training. In contrast to these methods, our approach relies on bidirectional translation between the source and target domains. During training, we perform supervised learning on the newly generated data. Furthermore, the testing process is also conducted on the newly generated data, ensuring domain consistency. This bidirectional translation strategy mitigates the domain shift, making it more effective in handling the challenges of unsupervised domain adaptive person re-identification.

3 RESEARCH METHODOLOGY

Our framework comprises two main components: the image translation network and the feature extraction network. In this section, we will first describe each network in detail. Following that, we will explain how we transform the unsupervised training process into a supervised one. Finally, we will discuss the testing procedure.

3.1 Image Translation Network

Before delving into the details of the networks, let's first introduce the two domains involved in our framework: $Domain_S$ and $Domain_T$. $Domain_S$ represents the source domain from which we start, and $Domain_T$ represents the target domain to which we aim to translate the images. The image translation network is responsible for learning the mapping between these two distinct domains.

Given the significant challenges and costs associated with obtaining paired data for this task, we have chosen to base our image translation network on CycleGAN [16]. CycleGAN is particularly advantageous for our application as it does not require paired images, unlike many other image-to-image translation methods. This capability is crucial in scenarios where paired training data is scarce or difficult to collect, such as in diverse real-world environments where exact correspondences between images across domains are not available.

Additionally, CycleGAN leverages a cycle consistency loss that enforces a unique constraint to bring the source domain closer to the target domain and vice versa, ensuring that the learned transformations preserve key attributes between the input and reconstructed images. This is particularly beneficial for tasks like person re-identification, where maintaining individual identity across domain translations is critical.

As demonstrated in Fig. 1, our image translation model consists of two main components: the generators ($Generator_{S\to T}$ and $Generator_{T\to S}$) and the discriminators ($Discriminator_S$ and $Discriminator_T$). The generators are responsible for translating images from one domain to another, while the discriminators aim to distinguish between real and generated images.

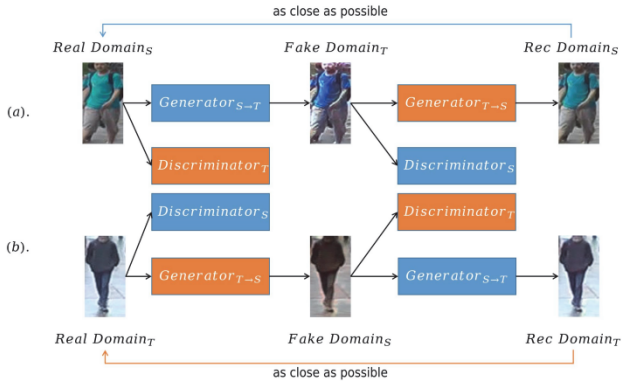

Figure 1 Illustration of the image translation network

As shown in Fig. 1a, the first branch works by taking an input pedestrian image from domain $Domain_S$ which is fed to our first generator $Generator_{S\to T}$ whose job is to transform a given image from source domain to an image in target domain $Domain_T$. This new generated image is then fed to another generator $Generator_{T\to S}$ which converts it back into an image in the original domain $Domain_S$. And as same as CycleGAN, this output image must be as close as possible to original input image to define a meaningful mapping that is absent in unpaired dataset. In addition, when an image from $Domain_S$ is transformed into $Domain_T$ by $Generator_{S\to T}$, $Discriminator_T$ is used to distinguish whether the transformed image is real (from the actual $Domain_T$) or fake (generated by $Generator_{S\to T}$). The goal of $Generator_{S\to T}$ is to fool $Discriminator_T$ into believing that the generated images are real.

Similarly, as depicted in Fig. 1b, the model also operates in the reverse direction. An image from $Domain_T$ is processed by $Generator_{T\to S}$ to generate an image in $Domain_S$. This image is then transformed back to $Domain_T$ by $Generator_{S\to T}$. The final output should closely resemble the initial input, akin to the first branch. Concurrently, $Discriminator_S$ is employed to differentiate real images in $Domain_S$ from those transformed by

$Generator_{T\to S}$. The objective of $Generator_{T\to S}$ is to deceive $Discriminator_S$ into classifying the transformed images as real.

The loss function of the image translation part is composed of two parts: the adversarial loss and the revised cycle consistency loss. The adversarial loss is used to match the distribution of generated images to the data distribution in the target domain. For simplicity, we will refer to the Generators and Discriminators as '*Gen*' and '*Dis'* in the following formulas. As a result, the adversarial loss can be defined as:

$$L_{GAN}(Gen, Dis, x, y) = E_{y\sim p_{data}(y)}[\log Dis(y)] + E_{x\sim p_{data}(x)}[\log(1 - Dis(Gen(x)))]. \quad (1)$$

where x is an image from one of $Domain_S$ or $Domain_T$, and y is an image from the other. The discriminators are trained to maximize the probability of correctly classifying real and fake images, while the generators are trained to minimize this probability, which is equivalent to maximizing the probability of the discriminators making a mistake.

In the revised cycle consistency loss, we incorporate a perceptual loss in addition to the original pixel-level loss to enhance the quality of the image translation by focusing on high-level content and structural integrity, rather than just pixel accuracy. The motivation for integrating perceptual loss stems from its ability to assess the perceptual similarity between images, which is crucial for maintaining the identity and semantic content during the translation process. This is particularly important in person re-identification, where preserving the distinct features of individuals across different domains is essential.

The perceptual loss is computed by comparing the feature maps of the generated image and the original image, as produced by a pre-trained model, specifically ResNet50. Let's denote ResNet50 [15] as F and $F_i(x)$ as the output of the i-th residual block of ResNet50 for image x. The new cycle consistency loss can be expressed as:

$$\begin{aligned} &L_{cyc}(Gen_{S\to T}, Gen_{T\to S}) = \\ &E_{y\sim p_{data}(y)}[\| Gen_{S\to T}(Gen_{T\to S}(y)) - y \|_1] + \\ &E_{x\sim p_{data}(x)}[\| Gen_{T\to S}(Gen_{S\to T}(x)) - x \|_1] + \\ &E_{y\sim p_{data}(y)}[\sum_i \| F_i(Gen_{S\to T}(Gen_{T\to S}(y))) - F_i(y) \|_1] + \\ &E_{x\sim p_{data}(x)}[\sum_i \| F_i(Gen_{T\to S}(Gen_{S\to T}(x))) - F_i(x) \|_1]. \end{aligned} \quad (2)$$

where $\|x\|_1$ denotes the L1-norm. In the above formulas, x and y represent the input images from $Domain_S$ and $Domain_T$ respectively. Additionally, the first two terms are the original pixel-level cycle consistency loss, and the last two terms are the newly added perceptual loss. Both perceptual loss terms compare the generated image and the original image at the output of each residual block in ResNet50, which is used as the backbone network in the following ReID feature extraction network. More details about F will be introduced later.

In addition to the above loss functions, which add constraints to each image independently, we propose an id consistency loss to ensure that the ID information from $Domain_S$ is preserved in the newly generated images in fake $Domain_T$. We first train a feature extraction network F_{id} on $Domain_S$ in a supervised manner. After training, all parameters of F_{id} (including the classifier) are frozen. Since the person IDs in fake $Domain_T$ are the same as those in $Domain_S$, we compute the ID loss for the generated images in $Domain_T$ using F_{id} . The formula is as follows:

$$L_{id}(Gen_{S\to T}, F_{id}, x) = E_{x\sim p_{data}(x)}[-\log F_{id}(Gen_{S\to T}(x))]. \quad (3)$$

In the above formula, the $-\log F_{id}(Gen_{S\to T}(x))$ term computes the negative log-likelihood of the correct class under the distribution predicted by F_{id} for the generated image $Gen_{S\to T}(x)$. This encourages the generated images to have similar ID features as the original images in $Domain_S$.

The total loss function is a weighted sum of the adversarial loss, the cycle consistency loss, and the id consistency loss:

$$\begin{aligned} &L(Gen_{S\to T}, Gen_{T\to S}, Dis_S, Dis_T, F_{id}) = \\ &L_{GAN}(Gen_{S\to T}, Dis_T, x, y) + \\ &L_{GAN}(Gen_{T\to S}, Dis_S, y, x) + \\ &\lambda L_{cyc}(Gen_{S\to T}, Gen_{T\to S}) + \\ &L_{id}(Gen_{S\to T}, F_{id}, x). \end{aligned} \quad (4)$$

where λ is a hyperparameter that controls the importance of the cycle consistency loss.

3.2 Feature Extraction Network

Following the image translation network, we will discuss the feature extraction network for person re-identification (ReID) in this subsection. The backbone of our feature extraction network is ResNet50 [15], a deep convolutional neural network known for its robust feature extraction capabilities.

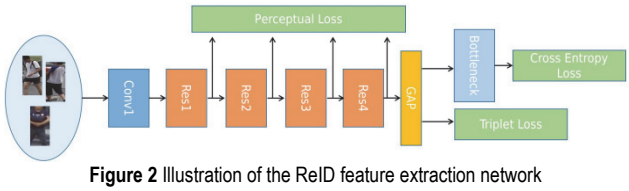

Figure 2 Illustration of the ReID feature extraction network

As presented in Fig. 2, the outputs of all four residual blocks in ResNet50 are utilized as inputs for the perceptual loss function described in Eq. (2). The output of the backbone network undergoes a Global Average Pooling (GAP) operation, resulting in a global feature vector. This global feature vector is optimized using a triplet loss.

The global feature vector is then passed through a fully connected (FC) layer with 2048 input channels and C_{out} output channels, followed by a Batch Normalization (BN) layer. The output of this sequence is the final output of the feature extraction network, which is then fed into a classifier and optimized using a Cross-Entropy (CE) loss.

3.3 Training on Generated Date

In the training phase, we initialize the ResNet50 denoted as F and F_{id} in the feature extraction network with parameters pre-trained on ImageNet. As illustrated in Fig. 3, we then train F_{id} on $Domain_S$ by minimizing the sum of the triplet loss and the cross-entropy loss, and freeze its parameters. During each epoch, we first freeze the parameters of F and train the image translation network by minimizing the sum of the adversarial loss, the cycle consistency loss, and the id consistency loss. After the image translation network is trained, we employ $Generator_{S\to T}$ to convert all data from $Domain_S$ into $Domain_{T'}$. We then unfreeze the parameters of the feature extraction network and train it using the newly generated data from $Domain_{T'}$, minimizing the sum of the triplet loss and the cross-entropy loss. This alternating training strategy continues for all T epochs.

Algorithm 1 Training procedure for our model

Initialize ResNet50 F and F_{id} with ImageNet pre-trained parameters
Define the adversarial loss L_{GAN}, cycle consistency loss L_{cyc}, id consistency loss L_{id}, triplet loss $L_{triplet}$, and cross-entropy loss L_{CE}
Set the hyperparameters: number of epochs T, learning rate η, final feature dimension C_{out}, and weight for cycle consistency loss λ
Train F_{id} on $Domain_S$ by minimizing $L_{triplet} + L_{CE}$, then freeze its parameters
for epoch = 1 to T **do**
    Freeze parameters of the feature extraction network F
    Train the image translation network with data from $Domain_S$ and $Domain_T$ by minimizing $L_{GAN} + \lambda L_{cyc} + L_{id}$
    Generate new data in $Domain_{T'}$ using the trained image translation network
    Unfreeze parameters of the feature extraction network F
    Train F with data from $Domain_{T'}$ by minimizing $L_{triplet} + L_{CE}$
end for

Figure 3 Illustration of training procedure for our model

3.4 Testing on Generated Date

In the testing phase, unlike traditional methods that directly test on $Domain_T$, we first preprocess the data in $Domain_T$. Specifically, we use $Generator_{T\to S}$ to transform the data from $Domain_T$ to $Domain_{S'}$. Subsequently, we use the $Generator_{S\to T}$ to transform the data from $Domain_{S'}$ to $Domain_{T'}$. This two-step transformation process allows us to leverage the learned mappings of our image translation network to align the data distribution of $Domain_T$ closer to that of $Domain_{T'}$, thereby improving the performance of our model in the testing phase. The transformed data in $Domain_{T'}$ is then fed into our trained model for final prediction.

3.5 Summary

In our model, the ResNet50, denoted as F, is shared between the feature extraction network and the image translation network. Another ResNet50, denoted as F_{id}, is trained on $Domain_S$ in a supervised manner and then its parameters are frozen. This allows F_{id} to extract ID-consistent features from the images. During the training of the image translation network, the parameters of F are frozen, meaning they are not updated. This allows the image translation network to learn to generate images that align with the feature space defined by F. In addition, an id consistency loss is introduced to ensure that the ID information from $Domain_S$ is preserved in the newly generated images in $Domain_T$. On the other hand, during the training of the feature extraction network, the parameters of F are updated. This allows F to learn to extract more discriminative features for the task at hand. This parameter sharing and alternating training strategy, along with the introduction of the id consistency loss, allows for a more efficient learning process, as the image translation network and the feature extraction network can benefit from each other's learning.

The proposed model, which includes dual generators and discriminators for the CycleGAN-based image translation, inherently demands substantial GPU memory. This is primarily due to the need to store multiple high-dimensional feature maps and model parameters during training and inference. Specifically, when the batch size is set to 6, the GPU memory occupied by the image translation network during the training process is 18G. In addition, since there is no need to save the gradient, the memory usage during testing will be greatly reduced.

The inference time of our model is significantly influenced by the complexity of the CycleGAN architecture. Specifically, each image undergoes translation and then reconstruction through the cycle consistency mechanism, effectively doubling the processing time compared to simpler translation models.

4 RESULTS

4.1 Datasets and Evaluation Metrics

Following the standard evaluation protocol on person re-identification (ReID) tasks, the evaluation metrics are the mean Average Precision (mAP) and Cumulated Matching Characteristic (CMC). We evaluated our method on three widely recognized benchmark datasets for ReID tasks, namely Market-1501 [12], DukeMTMC-reID [13], and MSMT17 [14]:

- Market-1501 [12]: This dataset is one of the most popular for person ReID. It contains 32,668 labelled images of 1,501 identities captured from six different camera views. The dataset is split into a training set of 12,936 images from 751 identities and a testing set of 19,732 images from 750 identities. The large number of identities and camera views makes it a challenging dataset for person ReID,
- DukeMTMC-reID [13]: This is a subset of the larger DukeMTMC dataset, specifically designed for person ReID. It consists of 36,411 labelled images of 1,812 identities captured from eight different camera views. The dataset is divided into a training set of 16,522 images from 702 identities and a testing set of 19,889 images from 702 identities. The diverse camera views and large number of identities make this dataset a challenging benchmark for person ReID,
- MSMT17 [14]: This is a large-scale multi-camera dataset for person ReID. It contains 126,441 images of

4,101 identities captured from 15 different camera views. The dataset is divided into a training set of 32,621 images from 1,041 identities and a testing set of 93,820 images from 3,060 identities. The large number of identities, camera views, and images make this dataset one of the most challenging benchmarks for person ReID.

4.2 Implementation Details

For our framework, we train the model with the Adam strategy for 50 epochs and the weight decay is set as 0.0005. All experiments are implemented with PyTorch on an NVIDIA RTX 4090 GPU.

4.2.1 Image Translation Network

In CycleGAN, the generator employs a modified U-Net structure [18], which consists of an encoder-decoder framework with skip connections between mirrored layers in the encoder and decoder. The discriminator, on the other hand, is a PatchGAN classifier, designed to classify whether local image patches are real or fake. In our image translation network, we adopt the same network architectures as those used in CycleGAN for both the generator and discriminator. This choice is motivated by the proven effectiveness of these structures in capturing complex image transformations in an unsupervised setting.

In addition, we construct $Domain_S$ using the training set from the source domain, and similarly, $Domain_T$ is formed using the training set from the target domain. For all the experiments, we set $\lambda = 10$ in Eq. 4. We use the Adam solver with a batch size of 6. All networks were trained from scratch with a learning rate of 2×10^{-4}.

4.2.2 Feature Extraction Network

For fair comparison, the feature extraction network uses ResNet-50 pretrained by ImageNet dataset as the backbone network by default. Specifically, we set C_{out} to 1024 and the margin for the triplet loss to 0.3. Input images are resized to 256 × 128 and the same data augmentation (random horizontal flipping and cropping) as in [1] are used. The batch size is 256, which is made up of 64 pedestrians and 4 images for each pedestrian. Further, the initial learning rate is set to 3.5×10^{-4} and divided by 10 at the 30th epochs.

4.3 Ablation Study and Qualitative Results

Following the convention in other papers, we use the notation "Market-1501→DukeMTMC-reID" to denote an unsupervised domain adaptation task for person re-identification (ReID), where Market-1501 serves as the source domain and DukeMTMC-reID as the target domain.

4.3.1 Visualization

In this section, we present the training progression of our image translation network, using the task of adapting from Market-1501 to DukeMTMC-reID as an example. The data in $Domain_S$ is sourced from the training split of Market-1501, while the data in $Domain_T$ originates from the training split of DukeMTMC-reID.

As illustrated in Fig. 4, the display is organized into six columns, each representing a specific type of image. From left to right, the columns represent:

- $real_S$: These are the original images from $Domain_S$,
- $fake_T$: These are the images generated by transforming the $real_S$ images to $Domain_T$ via $Gen_{S \to T}$,
- rec_S : These are the reconstructed images obtained by transforming the $fake_T$ images back to $Domain_S$ via $Gen_{T \to S}$,
- $real_T$: These are the original images from $Domain_T$,
- $fake_S$: These are the images generated by transforming the $real_T$ images to $Domain_S$ via $Gen_{T \to S}$,
- rec_T : These are the reconstructed images obtained by transforming the $fake_S$ images back to $Domain_T$ via $Gen_{S \to T}$.

In addition, each row of images in Fig. 4 corresponds to the state of the model saved at one of these checkpoints, showcasing the evolution of the model's performance over the course of 1, 10, 20, 30, 40, and 50 epochs of training.

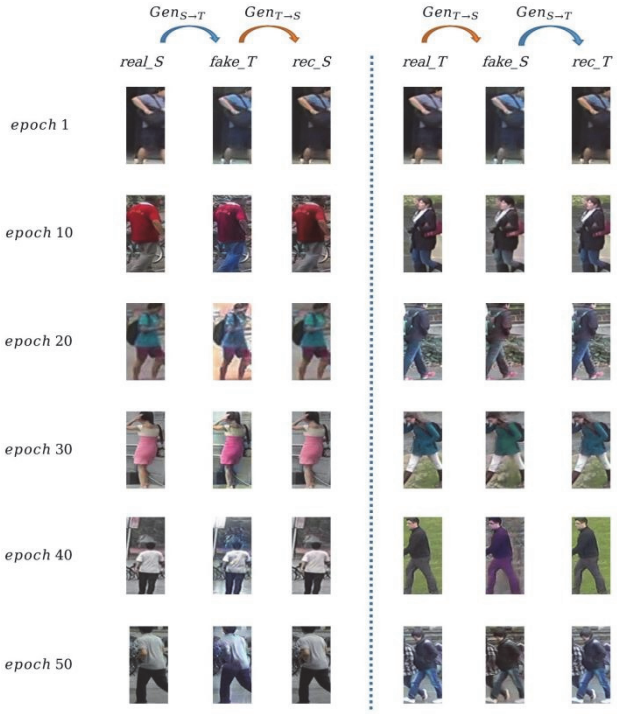

Figure 4 Progression of the image translation network training over different epochs

In the early stages of training (e.g., after 1 epoch), there are noticeable differences between rec_T and $real_T$ as well as between rec_S and $real_S$. This is due to the model still learning the mapping between the two domains. The translated images initially exhibit distortions and lack some domain-specific characteristics, which impacts the feature

extraction network's ability to accurately identify and extract pertinent features for person re-identification.

As training progresses, particularly evident by the 20th and 30th epochs, the quality of the translated images $fake_T$ and $fake_S$ improves markedly. The images begin to retain more of the essential attributes of the target domain, such as lighting, background, and style elements, which are crucial for the feature extraction network to perform effectively. This improvement in image quality enhances the network's ability to extract more discriminative and robust features, crucial for accurate re-identification across domains

By the 40th and 50th epochs, the differences between the reconstructed and original images decrease significantly. The translated images at these stages are almost indistinguishable from their real counterparts, indicating that the model has learned to effectively translate between the two domains and reconstruct the original images with high fidelity. This high level of accuracy in image translation ensures that the feature extraction network operates under optimal conditions, as it processes images that closely mirror the true distribution of the target domain.

4.3.2 Effectiveness of Feature Dimensionality

In this section, we will investigate the impact of the number of feature dimensions (C_{out}) on the performance of our model. We experiment with different settings of C_{out} , including 256, 512, 1024, and 2048. The results are presented in Tab. 2.

Table 2 Results for our models with different numbers of feature dimensions

C_{out}	DukeMTMC-reID→ Market-1501		Market-1501→ DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
256	80.1	92.7	70.3	82.9
512	82.9	93.1	72.6	84.0
1024	84.6	94.5	73.5	85.2
2048	84.0	94.1	73.1	84.3

From the table, it can be observed that increasing the number of feature dimensions generally leads to improved performance. Specifically, when C_{out} is increased from 256 to 1024, the mAP for the task of DukeMTMC-reID→ Market-1501 improves from 80.1% to 84.6%, and the Rank-1 accuracy improves from 92.7% to 94.5%. Similarly, for the task of Market-1501→DukeMTMC-reID, the mAP improves from 70.3% to 73.5%, and the Rank-1 accuracy improves from 82.9% to 85.2%. However, further increasing C_{out} to 2048 leads to slight decreases in performance. This suggests that a feature dimensionality of 1024 is sufficient to capture the necessary information for the person re-identification tasks, and further increasing the dimensionality may lead to overfitting and increased computational cost without significant benefits.

4.3.3 Effectiveness of Loss Functions

In Fig. 5, we illustrate the importance of incorporating id consistency loss into our model. Each row corresponds to a randomly selected sample from DukeMTMC-reID. The first column shows the original images from the source domain. The second column presents the translated images generated by our model using the id consistency loss, while the third column displays the translated images generated by our model trained without the id consistency loss. As can be observed, the use of id consistency loss leads to a significant visual improvement in the generated images compared to those generated without it. The images are clearer and retain more details, demonstrating that the id consistency loss helps the model to preserve the identity information during the translation process.

In addition to the visual comparison in Fig. 5, we also provide a quantitative comparison in Tab. 3. This table compares the performance of our model under different settings: without any additional loss, with perceptual loss, with id consistency loss, and with both. The performance metrics in the table show a significant improvement in mAP when using the id consistency loss. Specifically, for the task of DukeMTMC-reID→Market-1501, the mAP improves from 81.5% without any additional loss to 83.6% with id consistency loss. Similarly, for the task of Market-1501→DukeMTMC-reID, the mAP improves from 70.8% to 72.9% with id consistency loss. When both perceptual loss and id consistency loss are used, the performance improves even further, reaching an mAP of 84.6% for the task of DukeMTMC-reID→Market-1501, and an mAP of 73.5% for the task of Market-1501→DukeMTMC-reID. These results demonstrate the effectiveness of both perceptual loss and id consistency loss in improving the performance of our model.

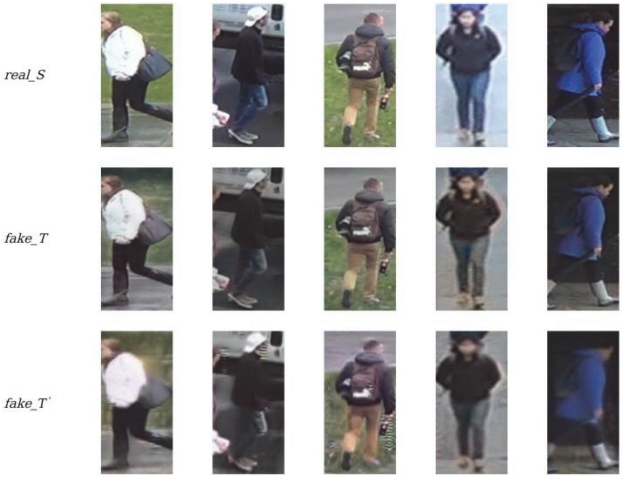

Figure 5 Illustration of the effect of the id consistency loss

The perceptual loss helps to capture high-level semantic information, which is crucial for maintaining the visual integrity and the contextual details of the identities across different domains. This loss function compares the feature maps of the generated image and the original image, as produced by a pre-trained model, to ensure that the essential characteristics and the overall structure are preserved during the translation process. This is particularly important in scenarios where the visual cues are subtle yet critical for the identification process.

On the other hand, the id consistency loss ensures the preservation of identity information during the image translation process. It works by maintaining a consistent representation of identity features across the original and translated images, thereby ensuring that the identity-

specific attributes are not lost or distorted. This loss is crucial for the re-identification task as it directly impacts the model's ability to correctly match the same individual across different images and domains.

Together, these two loss functions complement each other effectively. While the perceptual loss ensures that the high-level semantic and structural integrity is maintained, the id consistency loss safeguards the identity-specific features. This dual approach leads to more discriminative features and improved re-identification performance, as evidenced by the enhanced mAP and Rank-1 scores across different domain adaptation tasks.

Table 3 Performance comparison with different loss functions

Loss Function	DukeMTMC-reID→ Market-1501		Market-1501→ DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
None	81.5	92.8	70.8	83.1
Perceptual Loss	82.3	93.0	72.5	83.8
ID Consistency Loss	83.6	93.5	72.9	84.2
Both	84.6	94.5	73.5	85.2

4.3.4 Effectiveness of the Image Translation during Testing Phase

To further understand the contribution of the image translation network during the testing phase, we conduct an ablation study where we compare the performance of our model with and without the use of the image translation network during testing. The results are presented in Tab. 4.

Table 4 Comparison of model performance with and without the use of the image translation network during the testing phase

Image Translation	DukeMTMC-reID→ Market-1501		Market-1501→ DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
With	84.6	94.5	73.5	85.2
Without	81.9	93.0	71.4	83.6

As can be seen from the table, the use of the image translation network during the testing phase leads to a significant improvement in performance on both tasks. This demonstrates the importance of the image translation network in adapting the images to $Domain_{T'}$, which helps to ensure that both training and testing are conducted in the same domain ($Domain_{T'}$). However, it's worth noting that the use of the image translation network during the testing phase also increases the testing time, specifically doubling the inference time. Despite this trade-off, we believe that the benefits in terms of performance improvement justify the additional computational cost.

Given the increased inference time, a strategic approach to deploying this technology involves choosing when to utilize the image translation network based on the specific requirements of the task. For tasks where real-time processing is not critical, such as offline processing of surveillance footage or batch processing in forensic applications, the use of the image translation network during the testing phase is advisable. This strategy allows for the maximization of accuracy without the constraints of immediate response times. On the other hand, for real-time applications where speed is crucial, alternative strategies that optimize between speed and accuracy might be more appropriate, such as pre-processing the images in a less computationally intensive manner or using a simplified version of the network. This selective application ensures that the system remains versatile and effective across different operational scenarios.

4.3.5 Comparisons with the State-Of-The-Art Methods

In this part, we benchmark our proposed method against a range of state-of-the-art approaches on the tasks of DukeMTMC-reID→Market-1501 and Market-1501 →DukeMTMC-reID, as well as the tasks of Market-1501 →MSMT17 and DukeMTMC-ReID→MSMT17. Our evaluation encompasses a comprehensive set of metrics, including mAP and Rank-1, with the results detailed in Tabs. 5 and 6.

Table 5 Comparisons with the State-Of-The-Art Methods

Methods	DukeMTMC-reID→ Market-1501		Market-1501→ DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
MMCL [2]	60.4	84.4	51.4	72.4
ACT [3]	60.6	80.5	54.5	72.4
ECN-GPP [5]	63.8	84.1	54.4	74.0
AD-Cluster [4]	68.3	86.7	54.1	72.6
MMT [28]	71.2	87.7	65.1	78.0
MEB-Net	76.0	89.9	66.1	79.6
SpCL [7]	77.5	89.7	68.8	82.9
Dual-Refinement [8]	78.0	90.9	67.7	82.1
UNRN [9]	78.1	91.9	69.1	82.0
GLT [10]	79.5	92.2	69.2	82.0
RESL [11]	83.1	93.2	72.3	83.9
Ours	84.6	94.5	73.5	85.2

Our method outperforms existing models across most metrics on all tasks, as evidenced by the data in the tables. To elucidate why our method surpasses other approaches, it is crucial to highlight the unique integration and functionality of our framework components. The bidirectional image translation network effectively mitigates the domain shift by ensuring style consistency between the source and target domains, which is pivotal for maintaining the visual coherence of identity features across different environments. Simultaneously, our ReID feature extraction network is finely tuned to distill and enhance identity-specific features that are critical for accurate person re-identification.

Table 6 Comparisons with the State-Of-The-Art Methods

Methods	Market-1501→ MSMT17		DukeMTMC-reID→ MSMT17	
	mAP	Rank-1	mAP	Rank-1
ECN-GPP [5]	16.2	43.6	15.1	40.8
MMT [28]	23.3	50.1	22.9	49.2
SpCL [7]	26.8	53.7	26.5	53.1
Dual-Refinement [8]	25.1	53.3	26.9	55.0
UNRN [9]	25.3	52.4	26.2	54.9
GLT [10]	26.5	56.6	27.7	59.5
RESL [11]	33.6	64.8	34.2	65.2
Ours	35.8	67.2	36.5	68.1

Moreover, the introduction of the id consistency loss plays a fundamental role in our framework's success. This loss ensures that the identity features are not only preserved but also consistently represented across the translated images. This consistency is vital for the ReID system's ability to recognize the same individuals across varied domain-specific distortions, thereby significantly boosting the performance.

In summary, the integration of the image translation network with the ReID feature extraction network, bolstered by the id consistency loss, results in a robust framework that excels in domain adaptation for person re-identification. Our method not only achieves high-quality image translation between different domains but also extracts robust and discriminative features for person re-identification, setting a new benchmark on the DukeMTMC-reID→Market-1501, Market-1501→Duke MTMC-reID, Market-1501→MSMT17, and DukeMTMC -ReID→MSMT17 tasks.

5 CONCLUSIONS

In this work, we introduced a novel approach for unsupervised domain adaptation in person re-identification (ReID) tasks. Our framework integrates an image translation network with a ReID feature extraction network, augmented by the newly proposed id consistency loss. This combination not only bridges the domain gap but also ensures the preservation of identity-specific features, which are crucial for ReID. The image translation network adeptly transfers images across domains, aligning the data distributions, while the ReID feature extraction network focuses on distilling discriminative identity features. The use of the image translation network, however, doubles the inference time, making the process more time-consuming. Typically, training an image translation network requires a substantial amount of data from each domain, with each domain having over 500 images. This large dataset helps the network to learn the intricate details and variations wit hin each domain, leading to more accurate and efficient tr anslations. The id consistency loss, a key innovation in our method, further ensures that the identity is consistently maintained across the translated images, complementing the perceptual loss that preserves high-level semantic content. Our method sets new benchmarks on the DukeMTMC-reID→Market-1501, Market-1501→Duke MTMC-reID, Market-1501→MSMT17, and DukeMTMC -ReID→MSMT17 tasks, demonstrating its effectiveness. Nonetheless, there are trade-offs, such as increased inference time due to the image translation network and substantial GPU memory requirements during training. Future research could focus on enhancing computational efficiency and reducing memory demands, thereby extending the practicality of our approach. Despite these challenges, the significant performance gains on ReID tasks affirm the value of our method. We are confident that our contributions will inspire further innovations in the domain adaptation for person re-identification.

6 REFERENCES

[1] Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1487-1495. https://doi.org/10.1109/CVPRW.2019.00190

[2] Tang, Q. & Jo, K.-H. (2022). Unsupervised Person Re-identification via Mining Label Homogeneity. *2022 IEEE International Conference on Industrial Technology (ICIT)*, 1-6. https://doi.org/10.1109/ICIT48603.2022.10002807

[3] Yang, F., Li, K., Zhong, Z. et al. (2020). Asymmetric Co-Teaching for Unsupervised Cross-Domain Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 12597-12604. https://doi.org/10.1609/aaai.v34i07.6950

[4] Zhai, Y. et al. (2020). AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9018-9027. https://doi.org/10.1109/CVPR42600.2020.00904

[5] Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2021). Learning to Adapt Invariance in Memory for Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(8), 2723-2738. https://doi.org/10.1109/TPAMI.2020.2976933

[6] Zhai, Y., Ye, Q., Lu, S. et al. (2020). Multiple Expert Brainstorming for Domain Adaptive Person Re-Identification. *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 594–611. https://doi.org/10.1007/978-3-030-58571-6_35

[7] Ge, Y., Zhu, F., Chen, D. et al. (2020). Self-Paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-Id. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 11309-11321. https://doi.org/10.48550/arXiv.2006.02713

[8] Dai, Y., Liu, J., Bai, Y., Tong, Z., & Duan, L.-Y. (2021). Dual-Refinement: Joint Label and Feature Refinement for Unsupervised Domain Adaptive Person Re-Identification. *IEEE Transactions on Image Processing*, *30*, 7815-7829. https://doi.org/10.1109/TIP.2021.3104169

[9] Zheng, K., Lan, C., Zeng, W., Zhang, Z., & Zha, Z.-J. (2021). Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3538-3546. https://doi.org/10.1609/aaai.v35i4.16468

[10] Zheng, K., Liu, W., He, L., Mei, T., Luo, J., & Zha, Z.-J. (2021). Group-aware Label Transfer for Domain Adaptive Person Re-identification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5306-5315. https://doi.org/10.1109/CVPR46437.2021.00527

[11] Li, Z., Shi, Y., Ling, H., Chen, J., Wang, Q., & Zhou, F. (2022). Reliability Exploration with Self-Ensemble Learning for Domain Adaptive Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(2), 1527-1535. https://doi.org/10.1609/aaai.v36i2.20043

[12] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1116-1124. https://doi.org/10.1109/ICCV.2015.133

[13] Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3774-3782. https://doi.org/10.1109/ICCV.2017.405

[14] Wang, G., Yuan, Y., Chen, X. et al. (2018). Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *Proceedings of the 26th ACM International Conference on Multimedia*, 274–282. https://doi.org/10.1145/3240508.3240552

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. https://doi.org/10.1109/CVPR.2016.90

[16] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242-2251. https://doi.org/10.1109/ICCV.2017.244

[17] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241. https://doi.org/10.1109/ACCESS.2021.3053408

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. (2014). Generative Adversarial Nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *2*, 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

[19] Zhong, Z., Zheng, L., Luo, Z. M., Li, S. Z., & Yang, Y. (2019). Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 598–607. https://doi.org/10.1109/CVPR.2019.00069

[20] Yu, X.R., Han, B., Yao, J.C., Niu, G., Tsang, I., & Sugiyama, M. (2019). How Does Disagreement Help Generalization Against Label Corruption? Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA, 7164–7173. https://doi.org/10.48550/arXiv.1901.04215

[21] Huang, Y. R., Peng, P. X., Jin, Y., Xing, J. L., Lang, C. Y., & Feng, S. H. (2019). Domain Adaptive Attention Model for Unsupervised Cross-Domain Person Re-Identification. arXiv, 2019, arXiv:1905.10529. https://doi.org/10.48550/arXiv.1905.10529

[22] Yu, H. X., Zheng, W. S., Wu, A. C., Guo, X. W., Gong, S. G., & Lai, J. H. (2019). Unsupervised Person Re-Identification by Soft Multilabel Learning. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2143–2152. https://doi.org/10.1109/CVPR.2019.00225

[23] Han, B., Yao, Q. M., Yu, X. R., Niu, G., Xu, M., Hu, W. H., Tsang, I. W., & Sugiyama, M. (2018). Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 8536–8546. https://doi.org/10.48550/arXiv.1804.06872

[24] Ke, Z. H., Wang, D. Y., Yan, Q., Ren, J., & Lau, R. (2019). Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 6727–6735. https://doi.org/10.1109/ICCV.2019.00683

[25] Tarvainen, A. & Valpola, H. (2017). Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1195-1204. https://doi.org/10.48550/arXiv.1703.01780

[26] Jiang, L., Zhou, Z. Y., Leung, T., Li, L. J., & Li, F. F. (2018). MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. *Proceedings of the 35th International Conference on Machine Learning, Stockholm*, 2304-2313. https://doi.org/10.48550/arXiv.1712.05055

[27] He, K. M., Fan, H. Q., Wu, Y. X., Xie, S. N., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729-9738. https://doi.org/10.1109/CVPR42600.2020.00975

[28] Ge, Y., Chen, D., & Li, H. (2020). Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-Identification. arXiv, 2020, arXiv:2001.01526. https://doi.org/10.48550/arXiv.2001.01526

[29] Wei, L. H., Zhang, S. L., Gao, W., & Tian, Q. (2018). Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 79-88. https://doi.org/10.1109/CVPR.2018.00016

[30] Chen, Y. B., Zhu, X. T., & Gong, S. G. (2019). Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 232-242. https://doi.org/10.1109/ICCV.2019.00032

[31] Liu, J. W., Zha, Z. J., Chen, D., Hong, R. C., & Wang, M. (2019). Adaptive Transfer Network for Cross-Domain Person Re-Identification. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7195-7204. https://doi.org/10.1109/CVPR.2019.00737

[32] Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the International Conference on Machine Learning*, 448-456. https://doi.org/10.48550/arXiv.1502.03167

[33] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance Normalization: The Missing Ingredient for Fast Stylization. arXiv, 2016, arXiv:1607.08022. https://doi.org/10.48550/arXiv.1607.08022

[34] Jia, J. R., Ruan, Q. Q., & Hospedales, T. M. (2019). Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice. arXiv, 2019, arXiv:1905.03422. https://doi.org/10.48550/arXiv.1905.03422

[35] Jin, X., Lan, C. L., Zeng, W. J., Chen, Z. B., & Zhang, L. (2020). Style Normalization and Restitution for Generalizable Person Re-Identification. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3140-3149. https://doi.org/10.1109/CVPR42600.2020.00321

[36] Zhou, K. Y., Yang, Y. X., Cavallaro, A., & Xiang, T. (2022). Learning Generalisable Omni-Scale Representations for Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5056-5069. https://doi.org/10.1109/TPAMI.2021.3069237

[37] Choi, S., Kim, T., Jeong, M., Park, H., & Kim, C. (2022). Meta Batch-Instance Normalization for Generalizable Person Re-Identification. *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3424-3434. https://doi.org/10.1109/CVPR46437.2021.00343

[38] Jiao, B. L., Liu, L. Q., Gao, L. Y., Lin, G. S., Yang, L., Zhang, S. Z., Wang, P., & Zhang, Y. N. (2022). Dynamically Transformed Instance Normalization Network for Generalizable Person Re-Identification. Proceedings of the 17th European Conference on Computer Vision, 285-301. https://doi.org/10.1007/978-3-031-19781-9_17

[39] Xie, H., Luo, H., Gu, J., & Jiang, W. (2022). Unsupervised Domain Adaptive Person Re-Identification via Intermediate Domains. *Applied Sciences*, *12*, 2022. https://doi.org/10.3390/app12146990

Contact information:

Xiaohu HE, Associate Professor
(Corresponding author)
Weinan Normal University,
Chaoyang Street, Weinan, Shaanxi, 714099, PR China
E-mail: 254223964@qq.com

Jing LIU, Associate Professor
Weinan Normal University,
Chaoyang Street, Weinan, Shaanxi, 714099, PR China
E-mail: liujing8318@126.com