Adaptive Transfer Network for Cross-Domain Person Re-Identification

Jiawei Liu[1], Zheng-Jun Zha[1*], Di Chen[1], Richang Hong[2], Meng Wang[2]
[1]University of Science and Technology of China, China
[2]HeFei University of Technology, China
{ljw368,cdrom000}@mail.ustc.edu.cn, zhazj@ustc.edu.cn, {hongrc,wangmeng}@hfut.edu.cn

Abstract

Recent deep learning based person re-identification approaches have steadily improved the performance for benchmarks, however they often fail to generalize well from one domain to another. In this work, we propose a novel adaptive transfer network (ATNet) for effective cross-domain person re-identification. ATNet looks into the essential causes of domain gap and addresses it following the principle of "divide-and-conquer". It decomposes the complicated cross-domain transfer into a set of factor-wise sub-transfers, each of which concentrates on style transfer with respect to a certain imaging factor, e.g., illumination, resolution and camera view etc. An adaptive ensemble strategy is proposed to fuse factor-wise transfers by perceiving the affect magnitudes of various factors on images. Such "decomposition-and-ensemble" strategy gives ATNet the capability of precise style transfer at factor level and eventually effective transfer across domains. In particular, ATNet consists of a transfer network composed by multiple factor-wise CycleGANs and an ensemble CycleGAN as well as a selection network that infers the affects of different factors on transferring each image. Extensive experimental results on three widely-used datasets, i.e., Market-1501, DukeMTMC-reID and PRID2011 have demonstrated the effectiveness of the proposed ATNet with significant performance improvements over state-of-the-art methods.

1. Introduction

Person re-identification is the task of matching a probe pedestrian image from a large-scale gallery collected by non-overlapping camera networks at diverse locations [17, 35, 18]. It has been widely investigated due to its importance for many practical applications, such as automated surveillance, content-based retrieval and behavior analysis *etc.* [20, 38, 43, 36]. Recently, deep learning technique has been applied for person re-identification, leading to steady performance improvement on popular benchmarks [34, 41].

Figure 1. Illustration of the domain disparity among Market1501, DukeMTMC-reID and PRID2011 benchmarks, presenting significant variances in illumination, resolution and camera viewpoint *etc*.

Despite remarkable progress on person re-identification [26, 31, 11], it still remains a challenging task due to dramatic variances in imaging device, condition and environment among different surveillance cameras/camera networks. In practice, visual appearance of a pedestrian observed by different cameras at various locations and times varies drastically due to different camera configuration, lighting condition and viewing angles *etc.* This results in intensive disparities across pedestrian image galleries known as the challenge of *domain gap* in literatures [23, 16], hindering the application of exiting person re-identification systems. Existing re-identification models trained on one domain often fail to generalize well to another and suffer from severe performance drop. For example, GoogleNet [29] trained on Market-1501 dataset achieves rank-1 recognition rate of only 5.0% on PRID2011. Figure 1 illustrates the domain disparity among three popular benchmarks for person re-identification. They are collected at different places (*e.g.*, supermarket, campus and street) and present

* Corresponding author.

significant variances in illumination, resolution and camera viewpoint, *etc*.

A promising solution for bridging domain gap is unsupervised domain adaptation (UDA), which is a class of techniques aiming to use a source domain with labeled samples to learn a classifier with good capability on a unlabeled target domain. Typical UDA approaches assume that the source and target domain contain the same set of classes. Hence, they could not applied directly to person re-identification task as different re-identification datasets consist of entirely different pedestrian identities (classes). Recently, a few of UDA approaches tailored for person re-identification [2, 45, 46, 33, 3] have bee proposed upon the domain translation model CycleGAN [48]. These approaches encompass two phases typically. First, pedestrian images labeled with identities from a source domain are transferred into the style of a target domain, while preserving pedestrian identities. Second, the style-transferred images with labels are used to train a re-identification model for the target domain. These approaches treat the domain gap as a "black box" and attempt to tackle it resorting to a single style transformer. Actually, inter-domain disparities arise from the variations in multiple essential factors (*e.g.* illumination, resolution and camera viewpoint) during imaging process [22]. Even for each different image, the factors may hold different impacts on its imaging, leading to various cases of discrepancy across domains. Such complexity of domain discrepancy mixed with various factors challenge existing approaches, resulting in suboptimal performance.

In this work, we propose a novel Adaptive Transfer Network (ATNet) for effective cross-domain person re-identification. The ATNet looks into the "black box" of domain gap and proposes to address it following the principle of "*divide-and-conquer*". To the best of our knowledge, this work is the first one that looks into the essential factors of domain gap. It decomposes the complicated cross-domain transfer into a set of intermediate sub-tasks, each of which concentrates on style transfer at fine-grained level with respect to a certain factor. The sub-transformers are optimized jointly and assembled together to address the domain discrepancy. The ensemble of sub-transformers is self-adaptive to each image according to the affects of different factors. This gives ATNet the capability to transfer styles precisely with the perception of factor-wise affects. In particular, the proposed ATNet is built upon CycleGAN [48]. As illustrated in Figure 2, it consists of a transfer network composed by multiple factor GANs and an ensemble GAN as well as a selection network. Each factor GAN concentrates on transferring images to the target style of a certain imaging factor. The illumination, resolution and camera-view are three critical factors of domain disparity and are investigated in this work. It is noteworthy that ATNet is flexible to incorporate transfer modules of other factors. The ensemble GAN is designed to fuse the factor GANs adaptively towards painting precise style-transferred images. The selection network is to infer the affects of different factors on transferring each image, representing as sample-wise affect magnitudes which are used for the adaptive ensemble of factor GANs. We conduct extensive experiments to evaluate ATNet on three widely-used person re-identification datasets, *i.e.*, Market-1501 [42], DukeMTMC-reID [44] and PRID2011 [9], and report superior performance over state-of-the-art methods.

The main contributions of this paper are three-fold: (1) we propose a novel adaptive transfer network (ATNet) for effective cross-domain person re-identification following the principle of "*divide-and-conquer*" ; (2) we propose a flexible network architecture consisting of multiple factor GANs and an ensemble GAN. While the former performs factor-wise style transfer at more fine-grained level across domains, the latter synergizes factor GANs for effective domain transfer; (3) we propose an sample-wise adaptive ensemble of factor GANs by inferring the affects of various imaging factors on images.

2. Related Work

This work is closely related with unsupervised domain adaptation and feature learning in person re-identification. We will briefly summarize these two aspects of works.

2.1. Unsupervised Domain Adaptation

The proposed work relates to unsupervised domain adaptation (UDA) where images in the target domain are unlabeled. In the UDA community, most of the previous methods [25, 6, 5, 27, 28, 37, 32] attempt to align the source domain to the target domain by reducing the divergence of feature distributions. These methods assume that class labels are the same across domains, while different re-identification datasets contain different person IDs (classes). Thus, these approaches can not be applied directly for person re-identification.

Recently, a few cycle generative adversarial Networks (CycleGAN) [48, 1, 7] based UDA approaches [2, 45, 46, 33, 3] are proposed for person re-identification, which focus on learning a generator network that transforms samples in the pixel space from one domain to another. For example, Deng *et al.* [3] proposed a similarity preserving generative adversarial network (SPGAN) which preserved self-similarity of an image before and after translation, and domain-dissimilarity of a translated source image and a target image. Zhun *et al.* [45] introduced a Hetero-Homogeneous Learning (HHL) model, which enforced camera invariance, learned by positive pairs formed by unlabeled target images and their camera style transferred images, and domain connectedness, by regarding source / tar-

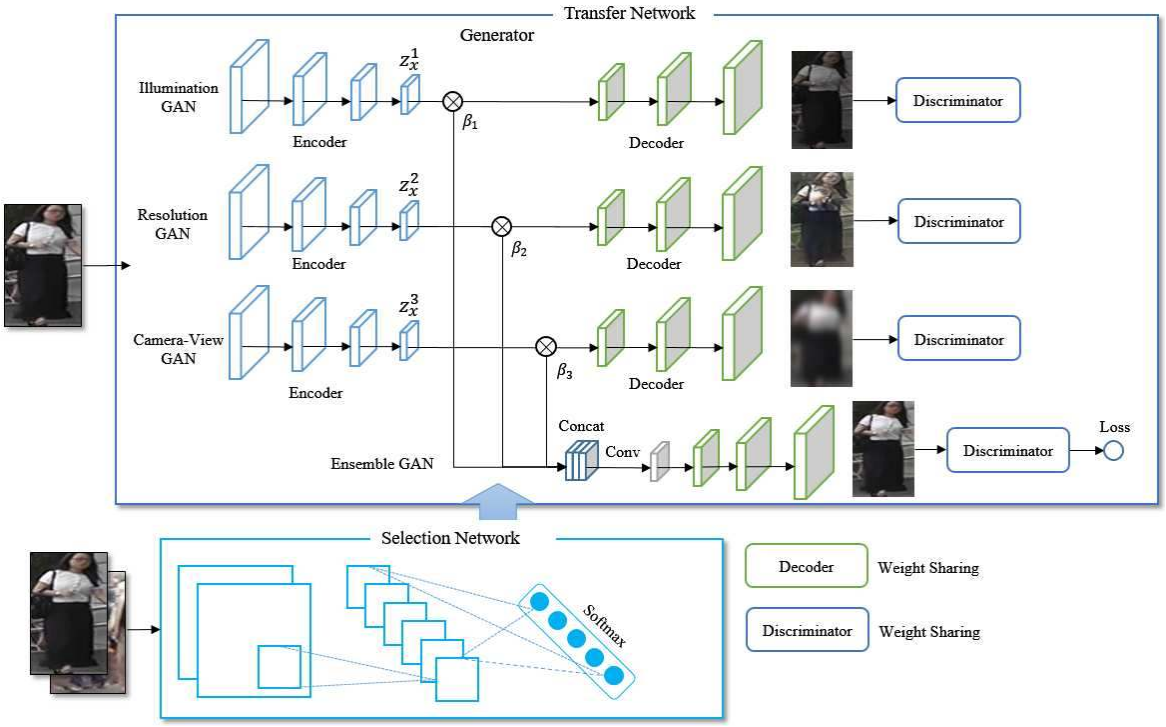

Figure 2. The overall architecture of the proposed ATNet approach. It consists of a transfer network for precise factor-wise style transfers and adaptive ensemble of them as well as a selection network for inferring affect magnitude of various imaging factors (e.g., *illumination, resolution and camera view*) on images.

get images as negative matching pairs. Slawomir *et al.* [2] proposed a three-step domain adaptation technique, which translated the Synthetic Person Re-Identification dataset to the target conditions by employing cycle-consistent adversarial networks. Wei *et al.* [33] proposed a Person Transfer Generative Adversarial Network (PTGAN) for bridging domain gap, which introduced a identity loss and a style loss to keep the identity of pedestrians and ensure the transferred images with similar style of target domain during transfer. Zhong *et al.* [46] proposed a camera style (CamStyle) adaptation method with a label smooth regularization (LSR) for person re-identification, which can serve as a data augmentation approach that smooths the camera style disparities and alleviate the impact of noise caused by the new generated samples.

2.2. Feature Learning

Deep learning based methods [47, 14, 26, 15, 40] for feature extraction have shown substantial advantage over traditional hard-crafted features on most of person re-identification datasets. For example, Xiao *et al.* [34] presented a pipeline for learning global full-body representations from multiple domains by a Domain Guided Dropout layer to discard useless neurons for each domain. Liu *et al.* [19] proposed a multi-scale triplet CNN which captures visual appearance of a person at various scales by a comparative similarity loss on massive sample triplets. McLaughlin *et al.* [21] presented a recurrent neural network architecture for video-based person re-identification, which utilized optical flow, recurrent layers and mean-pooling layer to learn video features containing appearance and motion information. Li *et al.* [14] formulated a method of jointly learning local and global features in a CNN model by optimizing multiple classification losses in different context.

3. The Proposed Method

In this section, we first present the overall architecture of the proposed ATNet and then elaborate its components.

3.1. Problem Formulation

Given an annotated dataset $\mathcal{S}$ from source domain and an unlabeled dataset $\mathcal{T}$ from target domain for person re-identification, the goal of unsupervised domain adaptation is to use the labeled source images to train a re-identification model that generalizes well to the unlabeled dataset on target domain. Considering that data bias caused by different influence factors $\mathbf{\Theta}$, we require a transfer model $G(\cdot)$ to translate the annotated dataset $\mathcal{S}$ from source domain to target domain, and learn effective generalized features of pedestrian with the new created dataset $G(\mathcal{S}; \boldsymbol{w}; \mathbf{\Theta})$. The unsupervised domain adaptation problem can be formulated

as:

$$\arg\min_{\boldsymbol{w}} \; D_{js}(P_{\mathcal{T}}(\boldsymbol{y}) \| P_G(\boldsymbol{x}; \boldsymbol{w}; \boldsymbol{\Theta})) \tag{1}$$

where D_{js} denotes the Jensen-Shannon divergence between two distributions, $P_{\mathcal{T}}$ denotes the distribution of the target domain over data $\boldsymbol{y}$, P_G denotes the distribution of the transfer model over data $\boldsymbol{x}$ from source domain $\mathcal{S}$. $\boldsymbol{w}$ and $\boldsymbol{\Theta}$ refer to the parameters of the transfer model and the factors (illumination, resolution, camera viewpoint, *etc*).

To learn an effective transfer model, we look into the "black box" of domain gap and address it following the principle of "*divide-and-conquer*". The complicated cross-domain transfer are decomposed into a set of factor-wise sub-transformers, each of which concentrates on style transfer at fine-grained level with respect to a certain factor, which are then assembled together to address the domain discrepancy. Moreover, the factors may hold different impacts on imaging process, the sub-transformers should be self-adaptive to each image according to the impacts of different factors for transferring style precisely. Thus, we propose a novel ATNet for effective cross-domain person re-identification. As shown in Figure 2, the ATNet consists of a transfer network containing multiple factor GANs and an ensemble GAN, and a selection network. Each factor GAN focuses on transferring images to the target style of a certain imaging factor. The ensemble GAN is designed to fuse the factor GANs adaptively towards painting precise style-transferred images. The selection network is to infer the weight scores of different factors on transferring each image, representing as sample-wise affect magnitudes which are used for the adaptive ensemble of factor GANs. After that, following the works [3, 33], we adopt the ResNet-50 [8] and GoogleNet [29] models as baseline to evaluate the performance on the target domain.

3.2. Transfer Network

Inter-domain disparities arise from the variations in multiple essential factors during image processing. The transfer network decomposes the complicated cross-domain transfer into a set of factor-wise sub-transformers, each of which focuses on style transfer with respect to a certain factor. The proposed framework is generic and flexible to include sub-transfers of other factors. We select illumination, resolution and camera view in this work as they are common and critical factors as discussed in literature. It optimizes the sub-transformers for these factors jointly and assembles them together to address the domain discrepancy. Moreover, the ensemble of sub-transformers is self-adaptive to each image based on the affects of different factors for generating more realistic images with a similar style of target domain.

Specifically, the transfer network contains three factors GANs and a emsemble GAN. They are all based on the CycleGAN model, which contains two generator-discriminator pairs, $\{G, D_{\mathcal{T}}\}$ and $\{F, D_{\mathcal{S}}\}$, producing a translated sample that is indistinguishable from samples in the other domain. The two generators $G : \mathcal{S} \rightarrow \mathcal{T}$ and $F : \mathcal{T} \rightarrow \mathcal{S}$ are the mapping functions. The two adversarial discriminators $D_{\mathcal{T}}$,$D_{\mathcal{S}}$ are used to distinguish whether samples are translated from source (target) domain. For simplification, we only consider that mapping a sample from source domain $\mathcal{S}$ to target domain $\mathcal{T}$ and ignore the reverse process. Similar to [30], the overall loss of the four GANs for image-to-image translation is expressed as:

$$\mathcal{L}_{gan} = \mathcal{L}_{adv} + \lambda_1 \cdot \mathcal{L}_{cyc} + \lambda_2 \cdot \mathcal{L}_{ide} \tag{2}$$

where $\mathcal{L}_{adv}$ is used for matching the distribution of translated images to the data distribution in the target domain, $\mathcal{L}_{cyc}$ attempts to recover the original sample after a cycle of translation and reverse translation, and $\mathcal{L}_{ide}$ encourages the style transfer to keep the color consistency between the original sample and translated sample.

Different from the original CycleGAN model with the adversarial loss, cycle-consistent loss and identity mapping loss, the three factor GANs are elaborately designed to concentrate on transferring images to the target style of the imaging factors, *i.e.*, illumination, resolution and camera viewpoint. On the one hand, the three factor GANs are pre-trained on the pair of datasets whose inter-domain difference are mainly induced by the three factors respectively to provide a good initialization. For pre-training illumination GAN, a collection of images with different illumination conditions is created by utilizing random gamma correction [24] in source domain. The created collection together with souce dataset are used for pre-training. For pre-training the resolution GAN, we downsample images in source domain to create a collection of images with multiple resolutions. For pre-training camera-view GAN, we use images from any two different cameras in source domain for pre-training. All created images will not be used in subsequent end-to-end training procedure of the network. On the other hand, an illumination constraint and a resolution constraint are introduced to the illumination GAN and resolution GAN respectively, for further guaranteeing that the style difference between original images and translated images focuses on the variations of illumination and resolution. The formulation of the illumination constraint is shown as follows:

$$\mathcal{L}_{ill}(G, F, H) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})}[\|H(G(\boldsymbol{x})) - H(\boldsymbol{x})\|_1] \tag{3}$$

where $H(\cdot)$ denotes abstracting illumination insensitive features [39]. This constraint is able to enforce the style consistency between the original image and translated image except the illumination condition. Thus, the final overall loss of the illumination GAN is: $\mathcal{L}_{gan} + \eta_1 \cdot \mathcal{L}_{ill}$. The formulation of the resolution constraint is shown as follows:

$$\mathcal{L}_{res}(G, F, I) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\|I(G(\boldsymbol{x})) - I(\boldsymbol{x})\|_2^2] \quad (4)$$

where $I(\cdot)$ denotes extracting resolution-insensitive features [13]. This constraint keeps the style consistency except the resolution variation. The final overall loss of the resolution GAN is: $\mathcal{L}_{gan} + \eta_2 \cdot \mathcal{L}_{res}$. Moreover, the losses of the three factor GANs reflect the degree of style disparity between the translated samples and the target samples, which can be viewed as the different impacts of the three factors that result in the domain gap. If the associated loss is smaller, the factor is more critical for the domain gap. Therefore, the weight scores of the three factors are the reciprocal of the associated losses. Then the three weight scores $\beta = (\beta_1, \beta_2, \beta_3)$ are normalized by a softmax function, and are used for the emsemble GAN.

The emsemble GAN takes the adaptive fused image feature $\boldsymbol{z_x}$ as input, which is computed by:

$$\boldsymbol{z_x} = [\beta_1 \cdot \boldsymbol{z_x^1}; \beta_2 \cdot \boldsymbol{z_x^2}; \beta_3 \cdot \boldsymbol{z_x^3}],\ \boldsymbol{z_x} \in \mathbb{R}^{64\times64\times768} \quad (5)$$

where $\boldsymbol{z_x^1}, \boldsymbol{z_x^2}, \boldsymbol{z_x^3} \in \mathbb{R}^{64\times64\times256}$ refer to the image features extracted from the associated encoders of the three factor GANs. Afterwards, the fused image feature are send to a convolution layer with $1 \times 1 \times 256$ filters and a decoder to generate the final translated image. The emsemble GAN also has a discriminator to distinguish whether the sample is real or fake. Moreover, a Jensen-Shannon divergence constraint is added to the image features $\boldsymbol{z_x^1}, \boldsymbol{z_x^2}, \boldsymbol{z_x^3}$ for enforcing the learned features possessing different semantic information, which is formulated as follows:

$$\mathcal{L}_{js}(\boldsymbol{z_x^1}, \boldsymbol{z_x^2}, \boldsymbol{z_x^3}) = f(\overline{\boldsymbol{z_x^1}}, \overline{\boldsymbol{z_x^2}}) + f(\overline{\boldsymbol{z_x^1}}, \overline{\boldsymbol{z_x^3}}) + f(\overline{\boldsymbol{z_x^2}}, \overline{\boldsymbol{z_x^3}}) \quad (6)$$

where f denotes the reciprocal of Jensen-Shannon divergence between two distributions, $\overline{\boldsymbol{z_x^1}}$, $\overline{\boldsymbol{z_x^2}}$, and $\overline{\boldsymbol{z_x^3}}$ are the normalized image features by a softmax function. The overall loss of the emsemble GAN is $(\mathcal{L}_{gan} + \eta_3 \cdot \mathcal{L}_{js})$, which is used to optimize the parameters of the transfer network.

The emsemble GAN and the three factor GANs have the similar architecture, in which the generators contain 9 residual blocks [8] and four convolution layers, while the discriminators are 70×70 PatchGANs [10]. More details can be founded in [48]. The decoders and the discriminators in the emsemble GAN and the three factor GANs share same parameters.

3.3. Selection Network

The selection network is developed to infer the weight scores of different factors $\beta = (\beta_1, \beta_2, \beta_3)$ on transferring each image, representing as sample-wise affect magnitudes which are used for the adaptive ensemble of factor GANs. We use the selection network to infer β. This allows ATNet to avoid going through the process of generating fake images and calculating the losses during testing, thus greatly reduces computational cost. The selection network contains four convolution layers and one fully connected layer. Specifically, the kernel size of the four convolution layers is $4 \times 4 \times 64$, $4 \times 4 \times 128$, $4 \times 4 \times 256$, $4 \times 4 \times 256$, respectively, the padding and the stride of these layers are 1 and 2. Each convolution layer is followed by a batch normalization (BN) and a rectified linear unit (ReLU) layer. The last fully connected layer has 6 hidden units. The output feature of the fully connected layer represents the three weight scores of a pair of images, which are then passed through two softmax operations, respectively. In the training stage, the selection network takes a pair of images from a source domain and a target domain as input, the weight scores of the pair of images calculated from the transfer network is viewed as the ground-truth. We optimize the selection network with MSE loss. In the testing stage, the output weight scores of the selection network is provided to the emsemble GAN for generating the final style-transferred image.

3.4. Feature Learning

Once we obtain the style-transferred dataset $G(\mathcal{S})$, which is composed of the translated images with the associated labels, the feature learning step is the same as supervised person re-identification methods. Since we mainly focus on the step of source-target image translation, we adopt the ResNet-50 and GoogleNet models as baseline, following the works [3, 33]. During testing, we can extract the 2048-dim pedestrian feature from ResNet-50 model and 4096-dim pedestrian feature from GoogleNet model for retrieval under the Euclidean distance, and test the performance on the target domain.

4. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed ATNet on three widely used person re-identification benchmarks and compare the ATNet to state-of-the-art methods. The experimental results show that ATNet achieves superior performance of UDA in person re-identification over the state-of-the-art methods. Moreover, we investigate the effectiveness of the proposed ATNet including the three factor GANs and the emsemble GAN.

4.1. Experimental Settings

Datasets - In this work, extensive experiments are conducted on three widely used datasets, i.e, Market-1501, DukeMTMC-reID and PRID2011 for fair comparison and evaluation.

The Market-1501 dataset contains 32,643 images of 1,501 identities captured by 6 cameras. All images are automatically detected by the Deformable Part Model (DPM) detector [4]. The dataset is fixedly divided into two parts respectively, one part contains 12,936 images of 750 identities as training set and the other contains 19,732 images of 751 identities as testing set.

The DukeMTMC-reID dataset contains 36,411 hand-drawn bounding boxes of 1,812 identities from 8 high-resolution cameras. It is fixedly divided into two parts respectively, one part contains 16,522 images of 702 identities as training set and the other contains 17,661 gallery images of 702 identities as testing set. In addition, there are 2,228 query pedestrian images.

The PRID2011 dataset is captured from two static surveillance camera views. Camera view A contains 385 persons, camera view B contains 749 persons, with 200 of them appearing in both views. Therefore, there are 200 person image pairs in the dataset. These image pairs are randomly split into a training and a testing set of equal size.

Evaluation Metrics - Evaluation Metrics Cumulative Matching Characteristic (CMC) is adopted for quantitative evaluation of person re-identification. The rank-k recognition rate in the CMC curve indicates the probability that a query identity appears in the top-k position. The other evaluation metric is the mean average precision (mAP), considering person re-identification as a retrieval task.

Implementation Details - The implementation of the proposed method is based on the Pytorch framework with eight NVIDIA Titan XP GPUs. Images in the three datasets are resized to $256 \times 256 \times 3$, the number of mini-batches is 8. The proposed architecture is optimized by 20,000 iterations in each epoch, and 20 epochs in total. For the transfer network, we adopt the Adam optimizer [12] with a learning rate of 0.0002. The learning rate remains unchanged for the first 10 epochs and linearly decay to zero over the last 10 epochs. The parameters λ_1, λ_2, η_1, η_2, η_3 are set to 10, 5, 2, 1, 1, respectively. The three factor GANs are pre-trained on the source dataset and the generated dataset with the variations of the three factors (illumination, resolution and camera viewpoint), the emsemble GAN is trained from scratch. For the selection network, the stochastic gradient descent (SGD) algorithm is started with learning rate lr of 0.01, the weight decay of $1e^{-5}$ and the Nesterov momentum of 0.9.

4.2. Comparison to State-of-the-Arts

Transfer from large dataset to large dataset. Table 1 shows the performance comparison of the proposed ATNet against 5 methods in terms of CMC accuracy and mAP on the large target datasets (DukeMTMC-reID and Market-1501). We employ ResNet-50 model as the baseline for feature learning, following the work [3]. When tested on DukeMTMC-reID, Market-1501 is used as the source dataset, and vice versa. "Supervised learning" denotes using labeled training sets from the target datasets. "Direct Transfer" means directly applying the source-trained model to the target datasets. CycleGAN(based), CycleGAN (base+ $\mathcal{L}_{ide}$) and SPGAN are the state-of-the-art methods. When comparing the supervised learning method and the direct transfer method (66.7% vs 33.1%, 75.8% vs 43.1%), it can be observed that a large performance drop when using the direct transfer method on the target domain, due to the bias of data distributions in different domains. When tested on DukeMTMC-reID, the proposed ATNet achieves 45.1% rank-1 recognition rate and 24.9% mAP score. We can see that our method improves the 2nd best compared method SPGAN by 3.7% rank-1 recognition rate and 2.6% mAP score. When tested on Market-1501, the proposed ATNet achieves 55.7% rank-1 recognition rate and 25.6% mAP score. It can be observed that our method improves the 2nd best compared method SPGAN by 4.2% rank-1 recognition rate and 2.8% mAP score. The comparison indicates that the effectiveness of the proposed ATNet to generate more realistic translated images for bridging domain gap. An illustration of some generated results is given in Figure 3.

Figure 3. Examples of original images and their style-transferred images after image-to-image translation. (Best viewed in color)

Transfer from large dataset to small dataset. Table 2 shows the performance comparison of the proposed ATNet against 3 methods in terms of CMC accuracy on the small target dataset (PRID2011). We employ GoogleNet model as the baseline for feature learning, following the work [33]. Market-1501 and PRID2011 are used as the source dataset

Method	Market-1501→DukeMTMC-reID					DukeMTMC-reID→Market-1501				
	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
Supervised Learning	66.7	79.1	83.8	88.7	46.3	75.8	89.6	92.8	95.4	52.2
Direct Transfer	33.1	49.3	55.6	61.9	16.7	43.1	60.8	68.1	74.7	17.0
CycleGAN (base) [48]	38.1	54.4	60.5	65.9	19.6	45.6	63.8	71.3	77.8	19.1
CycleGAN (base+ $\mathcal{L}_{ide}$) [48]	38.5	54.6	60.8	66.6	19.9	48.1	66.2	72.7	80.1	20.7
SPGAN [3]	41.4	56.6	63.0	69.6	22.3	51.5	70.1	76.8	82.4	22.8
ATNet	**45.1**	**59.5**	**64.2**	**70.1**	**24.9**	**55.7**	**73.2**	**79.4**	**84.5**	**25.6**

Table 1. Performance comparison to the state-of-the-art methods in terms of rank-k recognition rate and mAP scores on DukeMTMC-reID and Market-1501 datasets, respectively.

Method	Market-1501→PRID2011			
	cam1/cam2		cam2/cam1	
	Rank-1	Rank-10	Rank-1	Rank-10
Supervised	13.0	43.0	11.0	38.5
Direct Transfer	5.0	26.0	11.0	40.0
PTGAN(cam1)[33]	17.5	50.5	8.5	28.5
PTGAN(cam2) [33]	10.0	31.5	10.5	37.5
ATNet(cam1)	**24.0**	**51.5**	**21.5**	**46.5**
ATNet(cam2)	**15.0**	**51.0**	**14.0**	**41.5**

Table 2. Performance comparison to the state-of-the-art methods in terms of rank-k recognition rate on PRID2011 dataset.

Method	Market-1501→DukeMTMC-reID			
	Rank-1	Rank-5	Rank-20	mAP
ResGAN	37.9	53.9	64.0	21.3
CamGAN	38.1	53.8	63.9	21.4
illumGAN	39.8	54.3	65.2	21.7
ATNet w/o illumGAN	41.2	55.5	66.4	22.9
ATNet w/o CamGAN	42.1	55.6	66.2	23.1
ATNet w/o ResGAN	43.3	57.8	68.8	23.7
ATNet w/o adaptive	42.6	56.6	67.5	23.4
ATNet	45.1	59.5	70.1	24.9

Table 3. Evaluation of the effectiveness of each component within ATNet on DukeMTMC-reID dataset.

and the target dataset, respectively. The subscript $cam1$ and $cam2$ represent the transferred target dataset PRID-$cam1$ and PRID-$cam2$. "cam1/cam2" means using samples in PRID-$cam1$ as query set and samples from PRID-$cam2$ as gallery set, and vice versa. "Supervised learning" denotes using labeled training set of the target dataset. "Direct Transfer" means directly applying the source-trained model to the PRID2011 datasets. PTGAN is the state-of-the-art method. GoogLeNet trained on the Marker-1501 dataset, only achieves the Rank-1 accuracy of 5.0% on PRID2011, which implies substantial domain gap between Market-1501 and PRID2011. When transfered on PRID-$cam1$, the proposed ATNet achieves 24.0% and 21.5% rank-1 recognition rate for PRID-$cam1$ and PRID-$cam2$ as query set, respectively. It can observed that our method improves the compared method PTGAN by 6.5% and 13.0% rank-1 recognition rate, respectively. When transfered on PRID-$cam2$, the proposed ATNet obtains 15.0% and 14.0% rank-1 recognition rate for PRID-$cam1$ and PRID-$cam2$ as query set, respectively, boosting the compared method PTGAN by 5.0% and 3.5% rank-1 recognition rate, respectively. The comparison indicates that the effectiveness of the proposed ATNet and it can achieve reasonable re-identification performance on PRID2011 dataset, training on the other dataset. An illustration of some generated results is given in Figure 3.

4.3. Ablation Studies

To demonstrate the effectiveness and contribution of each component of the ATNet, we conduct a series of ablation experiments on DukeMTMC-reID dataset, using Market-1501 dataset as the source domain.

The impact of the proposed three factor GANs. We conduct the experiment to verify the influence of the three factor GANs on performance in Table 3. ATNet w/o illumGAN, ATNet w/o CamGAN and ATNet w/o ResGAN refer to the ATNet model without the illumination GAN, the camera viewpoint GAN and resolution GAN, respectively. These models achieve 41.2%, 42.1% and 43.3% rank-1 recognition rate, as well as 22.9%, 23.1%, 23.7% mAP score, respectively. From Table 3, we can observe that their performances are inferior to the ATNet, which shows the effectiveness of ATNet by incorporating the physical priors into UDA and utilizing the multiple factor GANs to decompose the complicated problem of bridging domain gap into handling the inter-domain the discrepancy caused by different factors. Moreover, by comparing the performance of the three models, it shows that the illumination GAN is the most important network branch to bridge domain gap.

The impact of the proposed emsemble GAN. We also conduct the experiment to verify the effectiveness of the

emsemble GAN with the adaptive ensemble strategy in Table 3. ResGAN, CamGAN and illumGAN donates only using the individual resolution GAN, the camera viewpoint GAN and resolution illumination GAN for UDA. ATNet w/o adaptive refers to the ATNet without the adaptive ensemble strategy ($\beta_1 = \beta_2 = \beta_3 = 1/3$). From table 3, it can observed that the ATNet w/o adaptive obtains better performance of 42.6% rank-1 recognition rate and 23.4% mAP scores as compared to the other three models, which indicates that the effectiveness of the ensemble GAN for handling the inter-domain discrepancy caused by multiply factors over one factor GAN for one factor. Moreover, the performance of the ATNet w/o adaptive model is inferior to the ATNet, demonstrating that the effectiveness of the adaptive ensemble strategy based on the different weight scores of the three factors, since the factors may hold different impacts on imaging process for each different sample. We also show some generated results from the three factor GANs with their associated weight scores in Figure 4. The image style of the translated images from the three factor GANs are different, as compared to the source images. We can see that the images from the illumination GAN slant dark, the images from the resolution GAN is more ambiguous, which show the factor GANs is able to handle the inter-domain differences caused by the different factors. By comparing the weight scores of the factors, it can be the observed that illumination condition are dominant for the domain gap.

Figure 4. Visual examples of image-to-image translation in Market-1501 with the weight scores. The images in the first column are from Market-1501. The images in the middle three columns are the translated images from the illumination, camera viewpoint and resolution GANs. The images in the last column is the final generated images. (Best viewed in color)

Sensitivity of ATNet to key parameters. The parameters η_1, η_2 and η_3 are key parameters for the proposed ATNet, which controls the relative importance of the proposed illumination constraint, resolution constraint and the Jensen-Shannon divergence constraint, respectively. We conduct experiment to evaluate the impact of η_1, η_2, η_3 respectively, and the results are shown in Figure 5. When adjusting the value of one parameter, the other parameters are fixed. From Figure 5, we can see that when $\eta_1 = 2$, $\eta_2 = 1$, $\eta_3 = 1$, the ATNet yields the best re-identification performance, which is superior to the ATNet without the three additional constraints ($\eta_1 = 0$, $\eta_2 = 0$, $\eta_3 = 0$). This comparison verifies the effectiveness of the proposed ATNet by using the three additional constraints.

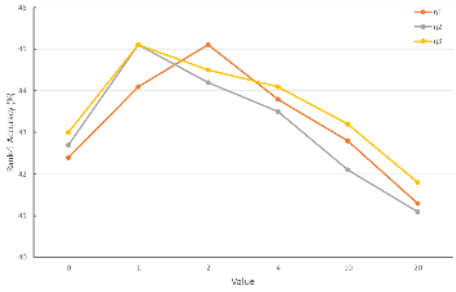

Figure 5. Evaluation of the proposed ATNet with different values of parameter η_1, η_2, η_3.

5. Conclusion

In this paper, we have addressed the cross-domain person re-identification problem by proposing a novel adaptive transfer network (ATNet), which looks into the essential imaging factors that engender dramatic inter-domain discrepancy. We proposed a "decomposition-and-ensemble" solution to tackle the complicated cross-domain transfer. ATNet was designed to contain multiple factor GANs, an ensemble GAN and a selection network. While each factor GAN concentrates on factor-wise precise style transfer at fine-grained level, the ensemble GAN adaptively fuses the factor GANs for effective domain transfer. The factor and ensemble GANs are jointly optimized in an end-to-end manner. The selection network was developed to perceive the affects of various factors on transferring different images to the target domain. Extensive experiments on multiple benchmarks have shown that the proposed ATNet outperforms state-of-the-art methods by a large margin.

Acknowledgement

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61622211 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

[1] Cycada: Cycle consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Ma-*

chine Learning, pages 1989–1998, 2018.

[2] S. Bak, P. Carr, and J.-F. Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[3] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6–16, 2018.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2008.

[5] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[6] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian Conference on Image Analysis*, pages 91–102. Springer, 2011.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2500–2510, 2017.

[11] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4–14, 2017.

[14] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *Proceeding of the International Joint Conference on Artificial Intelligence*, pages 2194–2200, 2017.

[15] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–12, 2018.

[16] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. F. Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[18] G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*, page 10. ACM, 2014.

[19] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the ACM Conference on Multimedia Conference*, pages 192–196. ACM, 2016.

[20] J. Liu, Z.-J. Zha, H. Xie, Z. Xiong, and Y. Zhang. Ca3net: Contextual-attentional attribute-appearance network for person re-identification. In *Proceedings of the ACM Conference on Multimedia Conference*, pages 737–745. ACM, 2018.

[21] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[22] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[23] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[24] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision*, pages 213–226, 2010.

[26] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2018.

[27] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8–18, 2016.

[28] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 443–450, 2016.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[30] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[31] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[32] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification.

[34] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[35] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 536–551. Springer, 2014.

[36] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016.

[37] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 994–1002, 2017.

[38] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[39] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu. Face recognition under varying illumination using gradientfaces. *IEEE Transactions on Image Processing*, 18(11):2599–2606, 2009.

[40] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3239–3248, 2017.

[41] K. Zheng, X. Fan, Y. Lin, H. Guo, H. Yu, D. Guo, and S. Wang. Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2877–2885. IEEE, 2017.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[43] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 649–656. IEEE, 2011.

[44] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[45] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[46] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[47] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 6, 2017.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2242–2251, 2017.