2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Unsupervised Multi-Domain Image Translation with Domain-Specific Encoders/Decoders

Le Hui, Xiang Li, Jiaxin Chen, Hongliang He, Jian Yang
DeepInsight@PCALab
Nanjing University of Science and Technology
Email: {le.hui, xiang.li.implus, jiaxinchen, HongliangHe, csjyang}@njust.edu.cn

***Abstract*—Unsupervised Image-to-Image Translation achieves spectacularly advanced developments nowadays. However, recent approaches mainly focus on one model with two domains, which may face heavy burdens with the large cost of training time and the huge model parameters, under such a requirement that n ($n > 2$) domains are freely transferred to each other in a general setting. To address this problem, we propose a novel and unified framework named Domain-Bank, which consists of a globally shared auto-encoder and n domain-specific encoders/decoders, assuming that there is a universal shared-latent space can be projected. Thus, we not only reduce the parameters of the model but also have a huge reduction of the time budgets. Besides the high efficiency, we show the comparable (or even better) image translation results over state-of-the-arts on various challenging unsupervised image translation tasks, including face image translation and painting style translation. We also apply the proposed framework to the domain adaptation task and achieve state-of-the-art performance on digit benchmark datasets.**

I. Introduction

Image-to-image translation problem is a general formulation which involves a wide range of various computer vision problems. Just as a sentence may be translated in either English or French, an image may be rendered into another image with a different attribution. Many problems in image processing can be defined as "translating" an input image in one domain into a corresponding output image in another domain. Typically, denoising, super-resolution, and colorization all pertain to image-to-image translation where the input is a degraded image (noisy, low-resolution, or grayscale) and the output is a high-quality color image.

Recently, a series of attractive works ignite a renewed interest in the image-to-image translation problem by adopting Convolution Neural Networks (CNNs). Gatys et al. [2] first study how to use CNN to reproduce famous painting styles on natural images. Since the seminal work by Goodfellow et al. [3], GAN has been proposed for a wide variety of problems. Unlike past works, by utilizing GANs, [1], [4], [5], [6] are proposed to translate an image from a source domain X to a target domain Y in the absence of paired examples. These algorithms often produce more impressive results near to the corresponding target domain, since a joint distribution that can be learned from two different domains by using images from the marginal distributions in individual domains.

Notwithstanding their demonstrated success, the existing approaches basically focus on the setting of one model with two domains. Specifically, by learning through one fresh training, translation is limited to transfer one pair of different domains. After a careful examination of existing image-to-image translation networks, we argue that different marginal distributions can be projected into a common space in their learned network structures. To the best of our knowledge, a translation among n domains has not yet been proposed in previous works.

As a result, the network is only able to capture two specific domains' translation one at a time. For a new domain, the whole network has to be re-trained end-to-end, which leads to an unavoidable burden under the situation where $n \times (n-1)$ training time are required, given n domains. In practice, this makes these methods unable to scale to a large number of domains. Additionally, how to further reduce the training time, network model size and enable more flexibilities to control translation among domains, remain to be considered yet to be addressed.

To overcome these problems, we explore a multi-domain image translation in which we reconsider the joint distributions of multiple domains. From the perspective of a probabilistic modeling, the coupling theory [7] states there exists an infinite set of joint distributions that can arrive at the given marginal distributions in general. This highly ill-posed problem forces us to make an additional assumption on the structure of the joint distribution. By further considering the interaction of n domains, we make a global shared-latent space assumption that assumes every sampled image from one of the n domains can be mapped to a universal shared-latent space. Based on the universal assumption, we propose a compact, and easily extended *Domain-Bank* framework that learns every domain pairs' joint distribution simultaneously.

In details, the proposed *Domain-Bank* framework is composed of multiple domain-specific component banks and each component represents one specific domain. Specifically, a component bank consists of encoders and decoders for specific domains. For an input image, the corresponding component bank maps an image to a shared-latent space and then decodes it to the target image.

In several challenging unsupervised multi-domain image translation tasks like face image translation and painting style translation, we comprehensively demonstrate the superior efficiency and at least comparable results to the state-of-the-art methods. Furthermore, as more domains' samples are engaged

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

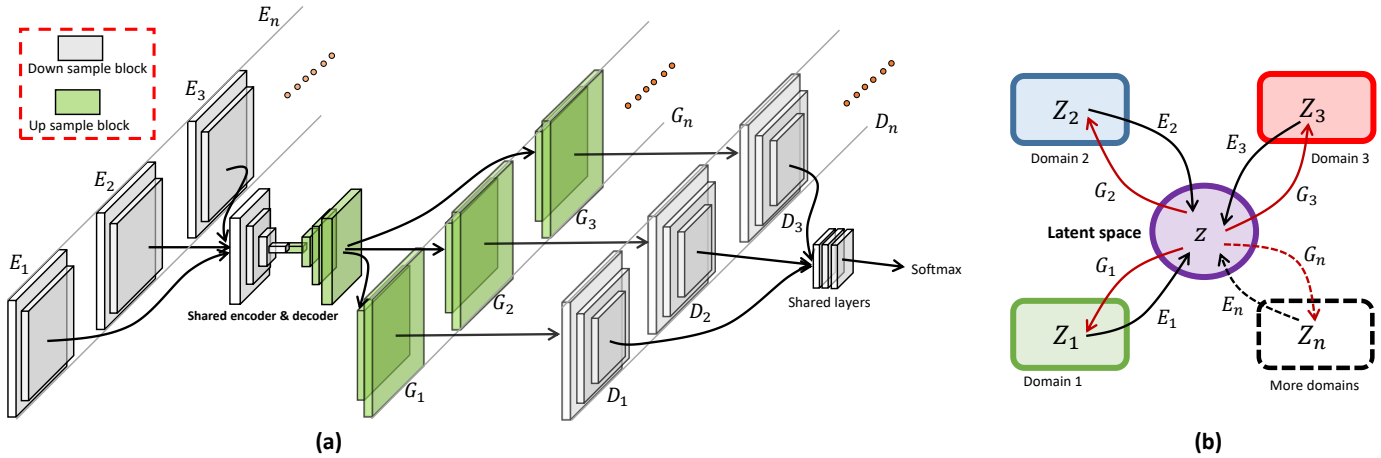

Fig. 1. (a) The proposed Domain-Bank framework. We declare tuples of $\{E_1, E_2, \cdots, E_n\}$ and $\{G_1, G_2, \cdots, G_n\}$ of n domains. By adopting a weight sharing constraint in the last few layers of $\{E_1, E_2, \cdots, E_n\}$ and the first few layers of $\{G_1, G_2, \cdots, G_n\}$, we implement the shared-latent space assumption. (b) The shared-latent space assumption. For arbitrary pairs of corresponding images (x_i^a, x_j^b), $a, b \in [1, n]$, they can be translated to a same latent code z. $\{E_1, E_2, \cdots, E_n\}$ encode functions that map the image to a latent code, and then $\{G_1, G_2, \cdots, G_n\}$ decode the latent code to images of the corresponding domain. $\{D_1, D_2, \cdots, D_n\}$ are adversarial discriminators for the corresponding domains in order to evaluate whether the translated images are realistic.

in *Domain-Bank* framework, the performance gain of domain adaptation tasks on digital recognition becomes consistently obvious.

Compared with existing models under the unsupervised image-to-image translation settings, our proposed *Domain-Bank* is unique in the following aspects:

- Our model is designed with a compact and clean structure, which also obtains a considerably huge reduction of training time and model parameters in case of n ($n > 2$) domains.
- The universal shared auto-encoder subnetwork is trained efficiently and effectively with multi-domain training samples/pairs, thus leading to a better generation which is confirmed in both quantitative and qualitative experimental results.

II. Related Work

Image-to-Image translation problem has already been promoted by the deep neural network and obtains some impressive results especially in the domain/style transfer fields. Neural generative models have recently received an increasing amount of attention. Several algorithms, including generative adversarial networks [3], variational autoencoders (VAEs) [8], [9], stochastic back-propagation [10] and diffusion processes [11], have demonstrated that a deep neural network can learn a domain distribution from examples.

Image-to-Image Translation. Many image processing and computer vision tasks can be viewed as an image-to-image translation problem, mapping an image in one domain to a corresponding image in another domain, e.g., image segmentation, stylization, super-resolution, and abstraction.

Unsupervised Image-to-Image Translation. In unsupervised image-to-image setting, we only have two independent sets of images where one possesses images in one domain and so do the other. Note that there are no paired samples guiding how an image could be translated to a corresponding image in another domain. Several other approaches also adopt the unpaired setting, where the goal is to relate two data domains, domain X and domain Y. More Recently, [12] proposed the domain transformation network (DTN) and achieved promising results on translating small resolution face and digit images. Liu et al. [4] proposed CoGAN, which used a weight-sharing strategy to learn a common representation across two domains. Following, Liu et al. [1] first made a shared-latent assumption, and then they proposed an unsupervised image-to-image translation framework, which used VAEs and GANs to learn a mapping from input to output images. Our approach builds on this framework.

Generative Adversarial Networks (GANs). GANs have achieved great success in a wide variety of computer vision applications, enhancing both supervised tasks and unsupervised ones. The key of GANs is the introduction of the *adversarial loss*, which forces the generated images to be indistinguishable from real images substantially. Soon after, various GANs have been proposed to the image generation on class labels [13], attributes [14], [15], and images [16], [1], [6], [4], [5], [17], [18]. A list of training tricks of GANs is given in [19].

Variational Auto Encoders (VAEs). A VAE consists of two networks that encode a data sample to a latent representation and decode the latent representation back to data space. The key of VAEs is to optimize a variational bound. By enhancing the variational approximation, superior image generation results were obtained [20], [21]. Larsen et al. [9] proposed a VAE-GAN architecture to improve image generation quality of VAEs. VAEs also were applied to translate face image attribute in [15]. More recently, Liu et al. [1] extended the framework of VAE-GAN to unsupervised image-to-image translation problems.

III. Domain-Bank Networks

We construct a multi-domain image translation network based on variational autoencoders (VAEs) [8], [9], [10] and generative adversarial networks (GANs) [3], [4], [6], which

encodes an input image to the shared-latent space and can also reconstruct/transfer it.

A. Network Architecture

Figure 1 shows our multi-domain translation architecture, which is based on the universal shared-latent space assumption. Supposing we are considering arbitrary two domains of n domains, namely $X_a, X_b, a, b \in [1, n], a \neq b$, which contain training samples $\{x_i^a\}_{i=1}^{N_a}$ where $x_i^a \in X_a$ and $\{x_j^b\}_{j=1}^{N_b}$ where $x_j^b \in X_b$, respectively. We denote the corresponding marginal data distributions as $x^a \sim P_{X_a}$ and $x^b \sim P_{X_b}$. We aim to learn a joint distribution of images in the domain X_a and domain X_b by utilizing images from the marginal distributions in two individual domains. It can be easily extended to the case of n domains. That is, we can learn a joint distribution of n domains.

Everyone (supposed to be $a, a \in [1, n]$) has two functional paths, through which any given images x^a sampled from P_{X_a} can be projected into the shared-latent code z and then can be recovered back as well. That is, we suppose there are functions E_a, E_b, G_a, and G_b $(a, b \in [1, n])$ such that, given a pair of corresponding images (x_i^a, x_j^b) (where $x_i^a \in X_a$, $x_j^b \in X_b$, $i \in [1, N_a]$ and $j \in [1, N_b]$) from the joint distribution. We define $z = E_a(x_i^a) = E_b(x_j^b)$, on the contrary, $x_i^a = G_a(z)$ and $x_j^b = G_b(z)$. In our structures, we map domain X_a to domain X_b through the function $x_j^b = F_{a\to b}(x_i^a)$, which can be represented by the function $F_{a\to b}(x_i^a) = G_b(E_a(x_i^a))$. Equally, we define two reconstruction functions for domain X_a to domain X_a: 1) $x_i^a = F_{a\to a}(x_i^a)$ and 2) $x_i^a = F_{a\to b\to a}(x_i^a)$. The function 1) can be equivalently written as $F_{a\to a}(x_i^a) = G_a(E_a(x_i^a))$, and the 2) is written as $F_{a\to b\to a}(x_i^a) = G_a(E_b(F_{a\to b}(x_i^a))) = G_a(E_b(G_b(E_a(x_i^a))))$. Notably, for an input image in the domain X_a, the function 1) directly translate it to an image in the domain X_a. However, in function 2), the input image is first translated from domain X_a to domain X_b, and then the generated image in the domain X_b is converted back to the domain X_a. In addition, a necessary condition for translating domain X_a to domain X_b to exist is the cycle-consistency constraint [22], [1], [6]: $F_{a\to b\to a}(x_i^a) = F_{b\to a}(F_{a\to b}(x_i^a))$. In other words, we can reconstruct the input image from translating back to the translated input image. Therefore, the shared-latent space assumption indicates the cycle-consistency assumption.

Domain-Specific Encoder and Decoder. Following the architecture used in [1], the image encoder E_a consists of 3 convolutional layers and 3 basic residual blocks [23], symmetrically, the image decoder G_a also consists of 3 basic residual blocks and 3 transposed convolutional layers. In our multi-domain image-to-image translation network, different domains have domain-specific encoders E_a and domain-specific decoders G_a. For instance, Monet's painting need to use Monet's specific encoder whilst Van Gogh's has to use Van Gogh's specific encoder. Similarly, this is also necessary for the domain-specific decoders. Different domains use domain-specific encoders to extract representations of the input images, and then domain-specific decoders responsible for decoding representations for reconstructing images in different domains. In other words, the encoder and the decoder can be seen as a domain-specific component, and we only need to train different components for different domains.

Universal Shared Auto-Encoder. Based on the shared-latent assumption, we enforce a weight-sharing constraint to relate the VAEs. Specifically, we further assume a share intermediate representation h that the process of generating corresponding images satisfy the formula

$$\{x^a, x^b, \cdots, x^n\} \to h \to z. \tag{1}$$

Therefore, we assume $E_a = E_{L,a} \circ E_H$ where $E_{L,a}$ is a low-level generation function that maps $X_a, a \in [1, n]$ to h, respectively. However, E_H is a common high-level generation function that maps h to z. From another view, z can be considered as the high-level representation of different domains, and h can be regarded as a special implementation of z through E_H. Similarly, h also admits us to represent G_a by $G_a = G_H \circ G_{L,a}$. In the implementation, we share the weights of the last few layers of E_a that are responsible for extracting high-level representations of input images in the n domains. Equally, the first few layers of $G_a, a \in [1, n]$ are shared, which responsible for decoding high-level representations for reconstructing the input images.

Domain-Specific Discriminator. Since we have n different domains, our framework has n adversarial networks: $GAN_a = \{D_a, G_a\}$. In GAN_a, for real images sampled from the domain X_a, D_a should output true, while for images generated by G_a, it should output false. In our framework, GAN_a can generate two types of images: 1) images from the reconstruction stream $F_{a\to a} = G_a(E_a(z))$ and 2) images from the translation stream $F_{a\to b} = G_b(E_a(z))$. The reconstruction stream can be trained with supervisions that we only apply adversarial training to images from the translation stream $F_{a\to b}$. Thus, we require train n domain-specific discriminators for n different domains.

B. Loss Functions

To better understand the losses applied in *Domain-Bank*, we first give a decomposed perspective of possible combinations of the key components in Figure 1. Basically, our framework is based on variational autoencoders (VAEs) and generative adversarial networks (GANs) including n domain image encoders, n domain image generators, and n domain adversarial discriminators.

VAE Loss. In our framework, we use variational autoencoder (VAE) to generate images in which VAE is supervised by the **KL** divergence. In the VAE, the encoder outputs a mean vector $E_{\mu,a}(x_i^a)$ where the input image $x_i^a \in X_a$. The distribution of the latent code z is written as $q_a(z_a|x_i^a) \equiv \mathcal{N}(z_a|E_{\mu,a}, I)$ where I is an identity matrix. We assume the distribution of $q_a(z_a|x_i^a)$ as a random vector of $\mathcal{N}(z_a|E_{\mu,a}, I)$ and sample from it. Thus, the reconstructed image is $x_i^{a\to a} = G_a(z_a \sim q_a(z_a|x_i^a))$. In addition, let η be a random vector with a multi-variate Gaussian distribution: $\eta \sim \mathcal{N}(\eta|0, I)$. In the VAE, the function $z_a \sim q_a(z_a|x_i^a)$ is implemented

via $z_a = E_{a,\mu}(x_i^a) + \eta$. The goal of VAE is to minimize a variational upper bound. The VAE object is written as

$$\mathcal{L}_{VAE_a}(E_a, G_a) = \lambda_1 \mathbf{KL}(q_a(z_a|x^a)\|p_\eta(z)) - \lambda_2 \mathbb{E}_{z_a \sim q_a(z_a|x^a)} \left[\log p_{G_a}(x^a|z_a)\right], \quad (2)$$

where the hyper-parameters λ_1 and λ_2 display the weights of the corresponding objectives and the $\mathbf{KL}$ divergence term penalizes deviation of the distribution of the latent code from the prior distribution. Note that the $L1$ loss used in VAE is to ensure similarity between the generated real image and the original rendering image.

Adversarial Loss. The adversarial losses are applied to translation functions. Note that we have defined G_a and its discriminator D_a, where $a \in [1, n]$. We express the objective as:

$$\mathcal{L}_{GAN_{ab}}(E_a, G_a, D_a) = \lambda_0 \mathbb{E}_{x^a \sim P_{x^a}} \left[\log D_a(x^a)\right] + \lambda_0 \mathbb{E}_{z_b \sim q_b(z_b|x^b)} \left[\log\left(1 - D_a(G_a(z_b))\right)\right], \quad (3)$$

where the hyper-parameter λ_0 controls the impact of the GAN objective functions. In the adversarial part, G_a attempts to generate images $G_a(z_b)$ that look like images from domain X_a, while D_a tries to distinguish between translated samples $G_a(z_b)$ and real samples x^a. Finally, G_a aims to minimize this objective against an adversary D_a that tries to maximize it.

Cycle-consistency Loss. We use a VAE-like function to model the cycle-consistency constraint, which is written as

$$\mathcal{L}_{cyc_{ab}}(E_a, G_a, E_b, G_b) = \lambda_3 \mathbf{KL}(q_a(z_a|x^a)\|p_\eta(z)) + \lambda_3 \mathbf{KL}(q_b(z_b|F_{a\to b}(x^a))\|p_\eta(z)) - \lambda_4 \mathbb{E}_{z_b \sim q_b(z_b|F_{a\to b}(x^a))} \left[\log p_{G_a}(x^a|z_b)\right], \quad (4)$$

where the hyper-parameters λ_3 and λ_4 control the weights of the different objective terms.

Full Objective. As a result, our ultimate objective is written as:

$$\mathcal{L}(E, G, D) = \sum_a^n \sum_b^n \{\mathcal{L}_{VAE_a}(E_a, G_a) + \mathcal{L}_{GAN_{ab}}(E_a, G_b, D_b) + \mathcal{L}_{cyc_{ab}}(E_a, G_b, E_b, G_a)\}, \quad (5)$$

where $a, b \in [1, n]$ and $a \neq b$. We aim to solve:

$$E^*, G^* = \arg\min_{E,G} \max_D \mathcal{L}(E, G, D). \quad (6)$$

C. Training Strategy

We employ an alternative training strategy motivated by GAN's [3] solving a mini-max problem where the optimization aims to find a saddle point. The zero-sum game in our framework consists of two plays: the domain-specific discriminators as the first team, and the domain-specific encoders/decoders for the second. During training with a specific pair of images from domain X_a and X_b, we first train domain-specific X_b's discriminator with all other components fixed. Afterward the X_a's encoder/decoder and X_b's encoder/decoder are involved not only to minimize the VAEs losses and the cycle-consistency losses but also to defeat the first one.

TABLE I
THE COST OF PARAMETERS AND TIME IN THE PROCESS OF TRAINING. OBVIOUSLY, ADVANTAGES OF OUR METHOD IN RETRENCHING CALCULATING SPACE AND TIME ARE REVEALED IN THIS FORM.

Experiment	Type	UNIT [1]	CycleGAN [6]	**ours**
Face (3)	Time	6 day	13 day	**3 day**
	Param	54.06M	68.28M	**25.58M**
Painting (4)	Time	12 day	27 day	**4 day**
	Param	19.62	136.56M	**5.94M**

TABLE II
UNSUPERVISED DOMAIN ADAPTION PERFORMANCE. THE REPORTED NUMBERS ARE CLASSIFICATION ACCURACIES.

Method	CoGAN [4]	UNIT [1]	**ours**
SVHN → MNIST	-	0.9053%	**0.9146%**
MNIST → USPS	0.9565%	0.9597%	**0.9645%**
USPS → MNIST	0.9315%	0.9358%	**0.9412%**

IV. EXPERIMENTS

We first give qualitative results on various tasks. Further, we present the quantitative performance gain on the digital domain adaptation tasks.

A. Qualitative Analysis

Face Attributes Translation. The CelebA dataset [24] is exploited for attribute-based face images translation. There are many different attributes of face images including hair, smiling and eyeglass. Particularly, we select a domain of hair with different colors including blond, brown, black, etc. Specifically, the hair with blond color constitutes the 1st domain, brown hair constitutes the 2nd domain, while the black hair constitutes the 3rd domain. In the experiments, we set λ_0=10, λ_1=λ_3=0.1 and λ_2=λ_4=0.1. In Figure 2, we visualize the results where we display the transitions between hair with different colors. It is not difficult to see that we obtain comparable results to other algorithms in hair translation.

Painting Style Translation. We further utilize landscape photographs downloaded from Flickr and WikiArt, which is also used in [6]. The size of the dataset for each artist/style is 526, 1073, 400, and 563 for Cezanne, Monet, Van Gogh, and Ukiyo-e. We use the same settings with λ_0=10, λ_1=λ_3=0.1 and λ_2=λ_4=0.1. Figure 3 shows our results in comparison with the other methods. Compared with UNIT, our results are superior to it. Particularly, Our generated images are more clearly and with higher contrast ratios than UNIT, whilst comparable to CycleGAN.

Summary of Complexity Comparisons. To demonstrate the advantages of the complexity of our proposed framework, we compare the training time and model parameters with those baselines in Table I. It can be clearly seen that our *Domain-Bank* has fewer parameters and training time. This is because we consider modeling n domains in a model so that n domains have a common shared-latent space. For models that only handle two domains, they need $n \times (n-1)$ separate shared-latent spaces. Thus, we have fewer model parameters in the shared-latent part when facing n domains. Because of the usage of different n domains during a training process, the shared-latent part of the network can better learn the

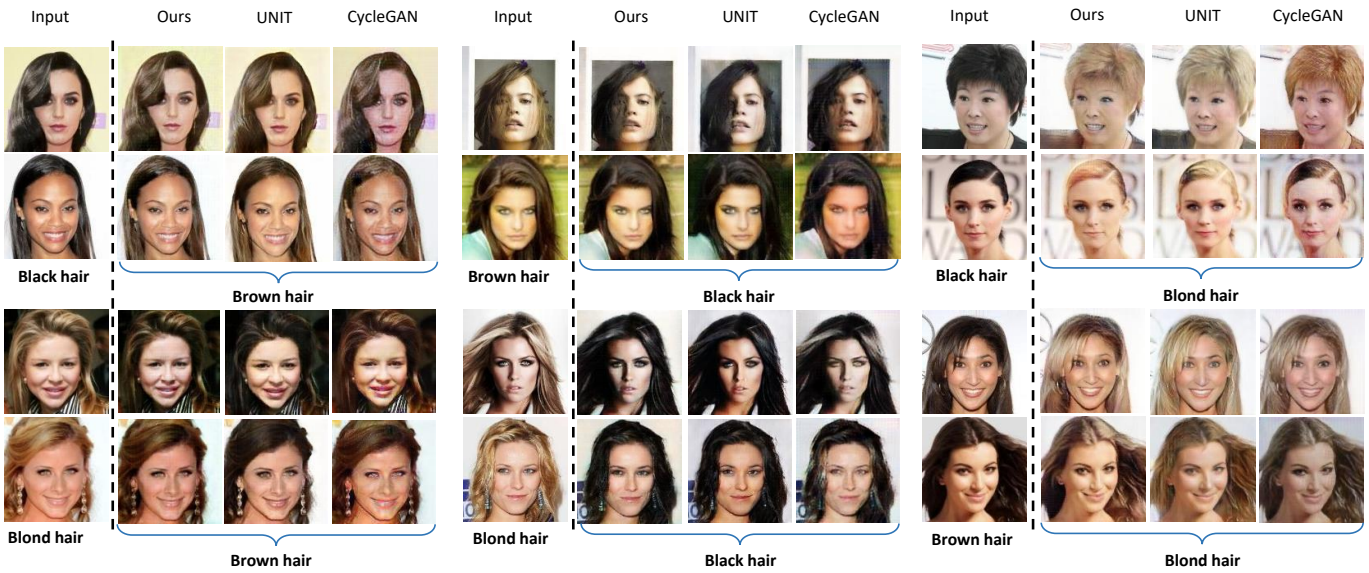

Fig. 2. The results of Face attributes translation. It contains the results of three sets of hair translation. In each group, the far left is the input image, and the results from left to right are Ours, UNIT and CycleGAN. The original quality of the human face and facial identity are better preserved by our method. Specifically, it can be seen that only the color of hair region changes, and the rest remains even in details. Additionally, our results are generated by passing end-to-end training only once, while others require training three models for translations between different pairs of domains respectively.

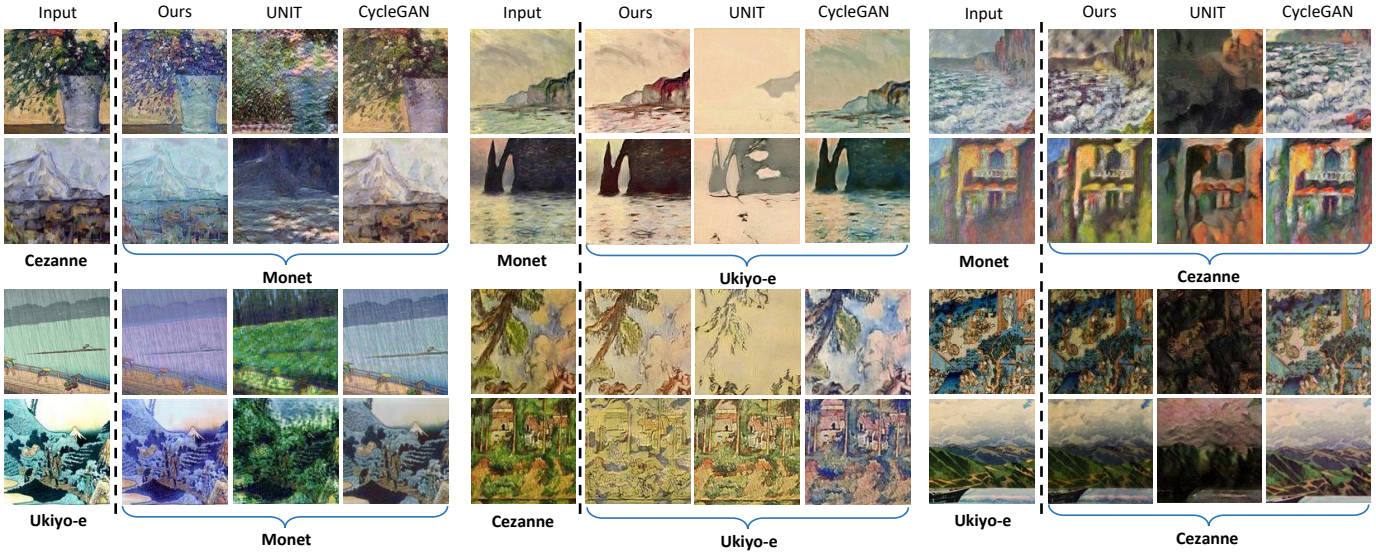

Fig. 3. The results show the painting style translation of famous painters: Cezanne, Monet, and Ukiyo-e. On the image quality, our generated images are superior to UNIT's and are comparable to those generated by CycleGAN. Note that all our results are obtained by training only once avoiding heavy training burden while others not.

commonalities and the characteristic among them. Due to the better commonalities and the characteristic, the convergence of network is speeded up, leading to less training time.

B. Quantitative Performance

In order to better understand the performance gained by sharing more information through more than two domains, we adopt our framework to the domain adaptation task, which adapts a classifier trained using labeled samples in one domain (source domain) to classify samples in a new domain where labeled samples in the new domain (target domain) are unavailable during training. In our case, we append additional auxiliary domains by applying our framework with minimal efforts to check whether it can boost the system's performance. The network structure is omitted.

More specifically, we utilize three datasets for digits: the Street View House Number (SVHN) dataset [25], the MNIST dataset [26] and USPS dataset [27], and perform multi-task learning where our framework is supposed to 1) translate images between any two of three domains and 2) classify the samples in the source domain using the features extracted by the discriminator in it. In practice, we adopt a small network because the digit images have a small resolution. In the experiment, we find that the cycle-consistency constraint is not necessary for this problem, and that is why we remove the cycle-consistency stream from the framework.

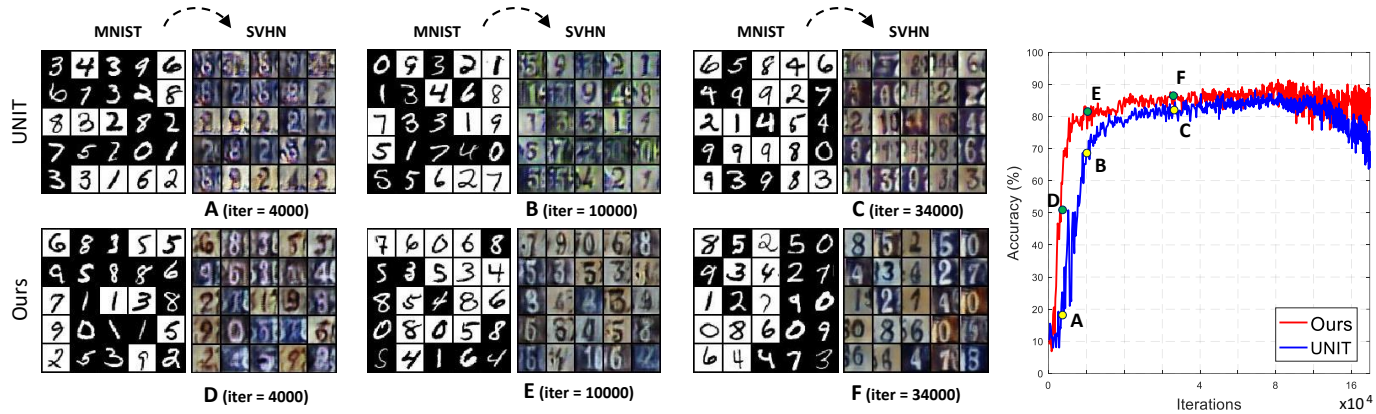

Fig. 4. The visualization of digital image translation. The curve on the right represents the classification accuracy of MNIST (In the case of test, we classify the MNIST dataset using the features extracted by the discriminator in the SVHN dataset.), and the six images numbered from **A** to **F** are the visualization of different iterations. Visibly, our results outperform in terms of image quality and details for describing digits. Due to the generation of clearer digits, our results obtain higher accuracy for classification by unsupervised learning without a hitch. Worth mentioning, our approach speeds up the training process.

As a result, Figure 4 shows the visualization of digit and Table II reports the achieved performance with comparison to the competing algorithms. We achieve better performances for SVHN $\rightarrow$ MNIST, MNIST $\rightarrow$ USPS and USPS $\rightarrow$ MNIST domain adaption, which are the state-of-the-art right now.

V. Conclusion

In this paper, we have proposed a novel multi-domain image translation framework, namely *Domain-Bank*. We show it learned to translate from multiple domains to multiple domains in one training process. Particularly, our *Domain-Bank* explicitly reduces the training time and model parameters, given n ($n > 2$) domains. The universal shared auto-encoder subnetwork leads to a better generation which is confirmed in both quantitative and qualitative experimental results.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant Nos. U1713208 and 61472187, the 973 Program No.2014CB349303, and Program for Changjiang Scholars.

References

[1] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *arXiv preprint arXiv:1703.00848*, 2017.
[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
[4] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016.
[5] Z. Yi, H. Zhang, P. T. Gong *et al.*, "Dualgan: Unsupervised dual learning for image-to-image translation," *arXiv preprint arXiv:1704.02510*, 2017.
[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.
[7] T. Lindvall, *Lectures on the coupling method*. Courier Corporation, 2002.
[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[9] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.
[10] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and variational inference in deep latent gaussian models," in *ICML*, 2014.
[11] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *arXiv preprint arXiv:1503.03585*, 2015.
[12] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.
[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[14] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.
[15] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *ECCV*, 2016.
[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.
[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.
[19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016.
[20] D. P. Kingma, T. Salimans, and M. Welling, "Improving variational inference with inverse autoregressive flow," *arXiv preprint arXiv:1606.04934*, 2016.
[21] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," *arXiv preprint arXiv:1602.05473*, 2016.
[22] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *arXiv preprint arXiv:1703.05192*, 2017.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[24] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.
[25] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop*, 2011.
[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[27] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics New York, 2001, vol. 1.