Unsupervised many-to-many image-to-image translation across multiple domains

Ye Lin1 | Keren Fu2 | Shenggui Ling1 | Peng Cheng3

1 National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, Sichuan, China
2 College of Computer Science, Sichuan University, Chengdu, Sichuan, China
3 School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan, China

Correspondence
Peng Cheng, School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065, China.
Email: chengpeng_scu@163.com

Funding information
Joint Funds of the National Natural Science Foundation of China, Grant/Award Number: U1833128; National Natural Science Foundation of China, Grant/Award Number: 61703077; Fundamental Research Funds for the Central Universities, Grant/Award Number: YJ201755; Sichuan Science and Technology Major Projects, Grant/Award Number: 2018GZDZX0029

Abstract
Unsupervised multi-domain image-to-image translation aims to synthesize images among multiple domains without labelled data, which is more general and complicated than one-to-one image mapping. However, existing methods mainly focus on reducing the large costs of modelling and do not pay enough attention to the quality of generated images. In some target domains, their translation results may not be expected or even cause the model collapse. To improve the image quality, an effective many-to-many mapping framework for unsupervised multi-domain image-to-image translation is proposed. There are two key aspects to the proposed method. The first is a many-to-many architecture with only one domain-shared encoder and several domain-specialized decoders to effectively and simultaneously translate images across multiple domains. The second is two proposed constraints extended from one-to-one mappings to further help improve the generation. All the evaluations demonstrate that the proposed framework is superior to existing methods and provides an effective solution for multi-domain image-to-image translation.

1 | INTRODUCTION

In image processing, image-to-image translation is regarded as an one-to-one mapping issue between two different image domains, which enables an input image to obtain features of another desired domain [16]. It is widely applied in computer vision industry such as, image segmentation [31], style transfer [17], image colorization [49], face synthesis [1, 18], image inpainting [32, 44] and super-resolution [21]. Without needing paired images, the unsupervised image-to-image translation methods [19, 26, 38, 46, 52] are more applicable than supervised ones [16, 28], since data preparation only involves dividing the images into different domains. For translations among multiple domains, namely multi-domain image-to-image translation, they are impractical due to the \( O(N^2) \) times of mutual training effort to achieve \( N(N - 1) \) different mappings, where \( N \) is the number of domains.

To simplify the training process for multi-domain image-to-image translation, a variety of schemes are adopted. One type of representatives [3, 15, 29, 30] is to divide each mapping into two separate processes, in which they use \( N \) domain-specialized encoders to separately process corresponding images from different domains and employ \( N \) domain-specialized decoders to finish the generations in different target domains. Thus, it only needs to choose the correct encoder and decoder to combine to complete the final translation. Another scheme [7, 23, 25, 39, 42] is to utilize a single generator with an auxiliary domain label vector to control translations among all domains. Although all these solutions can efficiently process multi-domain image translation, the quality of their translated images is not always good enough. There are still difficulties that cause the performance of their models to degrade. For the methods with splitting the generation, they have enough capacity to handle more general translations across any number of domains, but they are difficult...
to balance the training process between different domain pairs (input domain to output domain), possibly resulting in poor image quality or even model collapse. For the methods using an auxiliary vector, if the vector is not accurate enough to label different domains, the generated images may not conform to the target domain style. Besides, their single generator architecture makes them difficult to simultaneously handle the embeddings of different target domains, especially when there are large domain differences, possibly leading to unsatisfactory results.

To avoid the defects of the schemes mentioned above and improve the generations across all target domains, we present an unsupervised many-to-many image-to-image translation framework called multi-domain translator (MDT), based on generative adversarial networks (GANs) [9]. MDT has only one domain-shared encoder to reduce the interference of domain-specialized information from different source domains and has multiple identical domain-specialized decoders to translate images across all target domains. We also propose two general constraints extended from one-to-one mappings [19, 46, 52] to further help the multi-domain generation. The one named reconstruction constraint requires that, when the model generates images in all domains from a single input image, the generated domain-specialized images (except the result in the source domain) could be recovered to their original appearances in their source domain by feeding them to another cycled translation. The other enforces the input images to remain unchanged when they are processed by the decoders corresponding to their source domains, named identity consistency. With this architecture, MDT is easy to balance training and has sufficient capacity to process translations with large domain differences. An example of using MDT is shown in Figure 1.

Experiments are conducted on two types of image translation tasks with or without ground truth. Both qualitative and quantitative comparative evaluations demonstrate that MDT performs favourably against state-of-the-art methods [3, 7, 42] and offers an effective solution for unsupervised multi-domain image-to-image translation. In summary, our main contributions are listed as follows.

• We propose an unsupervised many-to-many framework to better improve the image quality of multi-domain image-to-image translation.

• Two general constraints are proposed to further improve the translations across all domains.

• Qualitative and quantitative evaluations demonstrate that the proposed method is superior to the state-of-the-art multi-domain image-to-image translation methods.

The remainder of the paper is organized as follows. Section 2 describes related work on image-to-image translation. Section 3 introduces the proposed MDT in details. The implementation is presented in Section 4. Experiments and analyses are included in Section 5. Conclusions are drawn in Section 6. More applications of the proposed framework are presented in Supporting Information.

2 | RELATED WORK

2.1 | Image-to-image translation in two domains

GANs [9] are widely used in solving image-to-image translation problems. The supervised approaches [16, 28] based on GANs train their model with paired data to generate high-quality images in desired domain. However, collecting labelled data is difficult, so an unsupervised setting is favoured by many researchers [19, 46, 52] which aim at achieving the same or even better results than the supervised ones. An effective practice is to utilize two generators to cooperate with each other to constraint and complete the entire image generation, which can be described as cycle consistency. The cycle consistency is an important constraint that requires the original image can be restored after successive two mappings by two different generators. Recent works, such as DiscoGAN [19], CycleGAN [52] and DualGAN [46], use cycle consistency to improve the quality of generated images. Shen et al. [35] further improve this constraint to obtain an accurate one-to-one mapping. On the other hand, Liu and Tuzel [27] propose a network with tied weights on the first few layers for shared latent representation, called CoGAN. Based on the idea of style transfer [17], Liu et al. use two pairs of encoders and decoders to embed input images into the same latent space and restore them to the target domain, called UNIT [26].

Multi-model image translation which focuses on increasing the diversity of translated images rather than generating images in different domains, is an extension of two-domain image translation. An image can be converted to different styles or appearances in the same target domain, such as translating a sketch to several color photos. Recently, many studies are proposed to tackle this problem by adding random noise [2, 53] or random style features [23] to the translation. Zhu et al. [53] use a low-dimensional latent vector that can be randomly sampled in testing to produce more diverse results, called BicycleGAN. Augmented CycleGAN proposed by Almahairi et al. [2], learns stochastic mappings which leverage auxiliary noise to capture multi-modal conditions. Li et al. [23] utilize an auxiliary variable to learn the extra information between two domains that have asymmetric information, and then produce diverse target
images. Huang et al. [14] directly extend UNIT [26] to multimodal scenarios called MUNIT, which encodes the images to a shared content space and combines a random domain-specialized style code for the generation.

2.2 Multi-domain image-to-image translation

Two-domain image-to-image translation models only handle an one-to-one mapping at a time. As the number of domains increases, the number of required mappings will exponentially increase. This makes them difficult to process the multi-domain image translation problems. Anoosheh et al. [3] who focus on reducing the modelling cost to linear complexity, divide the whole generator into $N$ encoders and $N$ decoders, and combine them into any pair to complete the translation between any two domains, called ComboGAN. A similar idea can be found in Domain-Bank proposed by Hui et al. [15]. They adopt a weight-sharing constraint in the last few layers of encoders and the first few layers of decoders. Additional shared layers for the discriminators are also used to tie weights before the final output. Unlike the partition of the generator, Choi et al. [7] use an auxiliary mask vector to label different domains and use only a single generator to translate multiple facial attributes, namely StarGAN. Following the same idea, Lin et al. [25] use relative attributes in the label vector to transfer target facial attributes and preserve other non-target attributes, called RelGAN. Yu et al. [47] utilize domain code to explicitly control the different generation tasks. To perform image translation with scalability and diversity, Wang et al. [42] condition the encoder with the target domain label and apply conditional instance normalization [40] in their network architecture, called SDIT which provides a solution for both multi-modal and multi-domain image translation. Similar work is also done by Xu et al. [45].

In general, existing strategies for multi-domain translations can be roughly classified as two types, splitting the generator or introducing an auxiliary label variable.

3 PROPOSED METHOD

Given different image domains represented by $X_i, i \in [1, N]$, the task of multi-domain image-to-image translation aims to find a representative mapping set $\{F_{ij}: X_i \rightarrow X_j\}, i, j \in [1, N], i \neq j$. To simplify such $O(N^2)$ complexity of mapping set, we in turn focus on functions $\{F_i: X \rightarrow X_i\}$ that can map any source domains to a target domain. A shared encoder $E$ together with several domain-specialized decoders $\{G_i\}_{i=1}^{N}$ can meet this requirement. This network can embeds images from different domains into a shared latent space, thus reducing the interference of domain-specialized features from different source domains and making the generated images more consistent with the styles of the target domains. We exploit GANs [9] to conduct our scheme. A translation to a domain $i$ is described as $G_i(E(X))$, and can be rewritten as $G_i(X)$ for the convenience of the following descriptions.

3.1 Objective

The overall objective of MDT consists of three parts:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{idt} \mathcal{L}_{idt}$$

(1)

where $\mathcal{L}_{adv}, \mathcal{L}_{rec}, \mathcal{L}_{idt}$ respectively represent the adversarial loss, two constraints of the reconstruction loss and the identity loss with their corresponding weights $\lambda_{rec}$ and $\lambda_{idt}$ to control their effects in training. The reconstruction loss and the identity loss are very important for unsupervised image-to-image translation. However, directly extending them from a two-domain scenario to that with multiple domains will cause an efficiency problem. To efficiently implement them across multiple domains, we propose to design a specific network architecture, which is actually the mapping functions $\{F_i: X \rightarrow X\}$. Here below, details of $\mathcal{L}_{adv}, \mathcal{L}_{rec}, \mathcal{L}_{idt}$ are described.

3.1.1 Adversarial loss

In unsupervised image translation with two image domains [52] represented as $X_i, X_j, i, j \in \{1, 2\}, i \neq j$, there are two mapping functions $F_{ij}: X_i \rightarrow X_j$. For each mapping, given two images $x_i \in X_i$ and $x_j \in X_j$, the adversarial loss [9] can be described as:

$$\mathcal{L}_{ij}(G_j, D_j, X_i, X_j) = E_{x_j \sim P_{X_j}}[f_1(D_j(x_j))] + E_{x_i \sim P_{X_i}}[f_2(D_j(G_j(x_i)))]$$

(2)

where $D_j$ and $G_j$ respectively represent the discriminator and the generator for a domain $j$, and $P_{X_i}$ is the data distribution of $X_i$, and $f_1, f_2$ are the functions for adversarial training, which can be specified as $f_1(D) = \log(D)$, $f_2(D) = \log(1 - D)$ or other forms in different models [11, 16, 52]. Following this objective, we can easily extend it to meet the requirements of translations among $N$ domains, where we only need to add up each item:

$$\mathcal{L}_{adv} = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \mathcal{L}_{ij}, i \neq j$$

(3)

3.1.2 Reconstruction loss

We generalize this loss from one-to-one mappings [19, 46, 52]. It means that the input image should remain unchanged when it is successively processed by two generators with opposite input and output. This is an important idea which helps constrain the generation so that the translated images can retain more original contents and have the styles of the target
FIGURE 2 The main process of training (left) and testing (right) across \( N \) domains for our method. During each training iteration, an unpaired bag consisting of images \( \{x_i\}_{i=1}^N \) from different domains \( \{X_i\}_{i=1}^N \) is fed to MDT. For an image of a certain domain, all decoders will participate in the generation and output \( N \) fake images which are used to compute the adversarial loss and implement the two constraints during training.

domain. If there are two domains \( X_1, X_2 \) and two mapping functions \( G_1 : X_2 \rightarrow X_1, G_2 : X_1 \rightarrow X_2 \), the reconstruction constraint requires \( G_1(G_2(X_1)) = X_1, G_2(G_1(X_2)) = X_2 \). To measure the degree of approximation between \( G_1(G_2(X_1)) \) and \( X_1 \), \( G_2(G_1(X_2)) \) and \( X_2 \), we use the \( L_1 \) distance as the metric because of the less blurring for generating images [16]. So the reconstruction loss for \( X_1, X_2 \) can be defined as:

\[
\|G_1(G_2(x_1)) - x_1\|_1, \|G_2(G_1(x_2)) - x_2\|_1
\]

For a multi-domain scenario, the underlying intuition is that an input image \( x_i \) from the domain \( i \) can be restored again after being transformed into all other domains except its source domain. Since \( x_i \) will be transformed into \( N - 1 \) domains, there are \( N - 1 \) reconstruction terms by feeding each generated image \( G_j(x_i) \) to the source domain generator \( G_i \). Thus, the reconstruction loss for \( x_i \) in a domain \( i \) can be defined as:

\[
\mathcal{L}_{\text{rec}}^i = \sum_{j=1}^{N} \|G_i(G_j(x_i)) - x_i\|_1, j \neq i \quad (4)
\]

For the total loss of reconstruction, we only need to add up all the losses of each input domain:

\[
\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} \mathcal{L}_{\text{rec}}^i \quad (5)
\]

3.1.3 Identity loss

In addition to reconstruction loss, we also propose another generalized loss as an extra constraint to improve the generation. The key is that when a domain generator processes an image from its own domain, it should allow the image to pass through without any changes. Thus the identity constraint means the invariance of generation in the generator’s own domain. It helps enforce the generator to learn its domain features. Following the same metric in the reconstruction loss, we define this constraint for a domain \( i \):

\[
\mathcal{L}_{\text{idt}}^i = \|G_i(x_i) - x_i\|_1 \quad (6)
\]

The total identity loss is thus:

\[
\mathcal{L}_{\text{idt}} = \sum_{i=1}^{N} \mathcal{L}_{\text{idt}}^i \quad (7)
\]

3.2 Procedure

Figure 2 illustrates the procedure of training and testing in MDT. During a training iteration, an unpaired bag consisting of images randomly selected from each domain is fed to the networks. Each image in this bag is processed by MDT to meet the two constraints and the adversarial loss. The coloured lines respectively show the constraints of reconstruction (green) and identity (yellow), the processes of adversarial training (red), and the cross-domain translations (black). A common encoder (the red block) is used to embed any single input \( x_i \in X_i \) to the shared latent space, and a corresponding decoder \( G_j \) (the blue block) is provided to decode the embeddings to the desired domain \( j \). For domain-specialized decoders and their corresponding discriminators, they are initially the same as each other, which means the labels of different domains are not
AlGORITHM 1 MDT training procedure. The functions $f_i, f_2$ in Equation 2 can be specialized into different adversarial forms. An image $x_i$ from the domain $i$ translated to the domain $j$ can be described as $\hat{x}_{ij} = G_j(x_i)$. Note that each generator always has a common encoding part with other generators, due to the shared encoder.

| Require: | $N$ image sets, $N$ generators, and $D(N)$ discriminators, respectively denoted as $\{X\}_i^{N}$, $\{G(x;\omega_g)\}_{i=1}^N$ with parameters $\{\omega_g\}_{i=1}^N$, and $\{D(x;\omega_d)\}_{i=1}^N$ with parameters $\{\omega_d\}_{i=1}^N$. |
| --- | --- |
| 1: | Randomly initialize all $\omega_g, \omega_d, i \in [1, N]$. |
| 2: | for number of training iterations do |
| 3: | Randomly get an unpaired bag $\{x_i\}_{i=1}^N = \{x_i \in X_1, \ldots, X_N \in X_N\}$. |
| 4: | Feed $\{G_1, \ldots, G_N\}$ with $\{x_i\}_{i=1}^N$ to obtain $\{\hat{x}_{ij}\}_{i=1}^N = \{G_1(x_1), \ldots, G_1(x_N), \ldots, G_N(x_1), \ldots, G_N(x_N)\}$. |
| 5: | for $i = 1, \ldots, N$ do |
| 6: | update $\omega_d$ to minimize $\sum_{j=1}^N \left[ E[f_j(D_j(x_i))] + E[f_j(D_j(\hat{x}_{ij}))] \right]$. |
| 7: | end for |
| 8: | for $i = 1, \ldots, N$ do |
| 9: | Compute $\ell_{rec} = \sum_{j=1}^N \left[ |G_i(\hat{x}_{ij}) - x_i|_1, j \neq i \right]$. |
| 10: | Compute $\ell_{adv} = ||\hat{x}_{ij} - x_i||_1$. |
| 11: | update $\omega_g$ to minimize $\sum_{j=1}^N \left[ E[f_2(D_j(G_j(x_i)))] + \lambda_{rec} \ell_{rec} + \lambda_{adv} \ell_{adv} \right]$. |
| 12: | end for |
| 13: | end for |

special, so the only preparation for training data is to divide the images into different domains. For testing, we only need to feed an image from any domain to the generator, and then we will get the translated results in all domains, which is actually a many-to-many mapping. It seems unnecessary to generate an image in its source domain, but considering that MDT ignores the input domain label for practical utility, this is an inevitable result.

3.3 Training algorithm

Here below is the training algorithm of MDT without considering the specific implementation or any algorithm used in each step. It does not separately train $N$ adversarial models but ties the encoder and all decoders together as a whole. Then, we accumulate all the losses obtained from different domains for backpropagation.

4 IMPLEMENTATION

4.1 Network architecture

Figure 3 shows the architecture of our method. For the generator, the encoder and decoders respectively contain several convolutional and deconvolutional layers with size-4, stride-2, and padding-1 to down-sample and up-sample, and have 6 residual blocks [12] appended at the tail and header of them. The leaky rectified linear units (LReLU) with a slope of 0.2 is selected as the none linearity activation after instance normalization (IN) [40]. To make full use of different feature maps in the generation, skip connections [34] are adopted between the same size features of encoder and decoders. For the discriminator, we selectively follow the conduction of Pix2Pix [16], using Markovian patchGAN [22] architecture to discriminate whether $70 \times 70$ overlapping image patches are real or fake. To enhance the discriminator, the output of the original version is also retained. This means there are two outputs in a discriminator, an array composed of the discrimination results of all local image patches, and a scalar representing the discrimination result of the entire image.

4.2 Training details

Instead of training generator (G) and discriminator (D) to minimize and max Equation 1, we turn to respectively minimize the loss for G and D. Binary cross entropy (BCE) is used to measure the adversarial loss, and the real or fake image is labelled as 1 or 0 for the D. In detail, with the measurement of BCE, Equation 2 is equal to choose $f_1(D) = -\log D$, $f_2(D) = -\log (1 - D)$ for training the D to minimize Equation 3. For the G, since the D only needs to regard fakes as reals, the measurement is equal to choosing $f_1(D) = 0$, $f_2(D) = -\log D$, and then the G is trained to minimize Equation 1.

We use the Adam optimizer [20] with momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$ and a batch size of 1, suggested in CycleGAN [52]. All the decoders are bundled with the encoder as a whole for backpropagation to update the weights, as are the discriminators. All models are trained from scratch with a variable learning rate which is a constant value 0.0002 in the first half of epochs and then is linearly decayed to zero over the rest of epochs. For all experiments, the images are resized to $256 \times 256$. We choose $\lambda_{rec} = \lambda_{adv} = 10$ for Equation 1 because we find this setting is helpful to balance the image quality of each translated domain.

5 EXPERIMENTS

5.1 Experimental setting

We adopt three unsupervised multi-domain methods as the baseline models. The first one is ComboGAN [3], partitioning the generation. The others are StarGAN [7] and SDFT [42], introducing an auxiliary label vector into the generation. All of them are representatives and have achieved state-of-the-art results.

We conduct experiments on two types of datasets, images in each domain with or without ground truth in the other domains. All datasets contain multiple image domains. The first one is Artistic Painting Styles, which includes five different unpaired images, dividing into four artistic painting styles and one scenic photo style, collected from Flickr and WikiArt [8]. Each domain
FIGURE 3 The architecture of MDT with the corresponding kernel size (k), stride (s), padding (p), and the number of feature maps (n). For the generator, the black lines are the flow of cross-domain translations, and the encoder or decoder is surrounded by a rectangle with red or blue dots. Here we show an example of two decoders labelled as i and j. For the discriminator, there are two outputs, an array and a scalar respectively representing the discrimination results of all image patches and the entire image.

has hundreds or thousands of images for training, as well as tens or hundreds for testing. The other is Multi-PIE [10], which has more than 750, 000 face images under 15 poses, 20 illuminations, and 6 expressions, taken from 337 subjects of different ages, genders, races, and whether they wear glasses or not. In this database, we can divide the images into different domains, in which each image has ground truth in other domains. We choose different illuminations from session-1 of this database for face re-lighting synthesis. In each illumination domain, there are 249 face images corresponding to different subjects with the same pose and expression, divided into 150 and 99 for training and testing.

Here we evaluate all methods through the translations among three domains because it is a typical multi-domain scenario with the minimum domain numbers. If a method performs unsatisfactory translations across three domains, it will also be ineffective in performing translations with more domain numbers. The style transfer among three different artistic styles and the face re-lighting under three illuminations namely normal, shadow, and dark, are selected as the tasks which represent two typical application scenarios of image-to-image translation, and respectively have large and small differences among their domains. Each task contains six subcases, which are one-to-one mappings and are used for evaluation together with the whole task. We trained all models for 200 epochs with their default recommended super-parameters.

5.2 Qualitative evaluation

Figure 4 presents the translations among three different artistic painting styles, including the samples of original paintings (left top) for the convenience of clear comparison. ComboGAN is failed and collapsed in some domain pairs and is also not good enough in its relatively successful results. StarGAN and SDIT also have unsatisfactory translation results with obvious artefacts and ambiguities. Compared to these state-of-the-art methods, our approach transforms the images more realistically and accurately without obvious artefacts, and the transferred styles are more in line with the target styles.

The results of face re-lighting are shown in Figure 5. ComboGAN is still unable to translate all subcases well. The other three methods perform equally well, but according to the ground truth, StarGAN and MDT seem to retain more identity features after mapping.

5.3 Quantitative evaluation

Since the experiments are conducted on two types of datasets, one of which contains unpaired images, and the other in which all images have ground truth, we use corresponding metrics to quantitatively measure the results.

5.3.1 Evaluation of unpaired samples

Fréchet Inception Distance (FID) [13] is usually used to evaluate the performance of GANs [9]. It measures the distance between two samples which are the real images and the generated images, with the principle that if the distributions of two samples are more similar, the FID value is lower. Similar to FID, Kernel Inception Distance (KID) [5] is another metric to evaluate the distance of two images respectively selected from the real sample and the generated sample, so we can use the mean measurement as the indicator. Both of these two metrics utilize the Inception Network [37] to obtain the image features to compute their own scores. Table 1 illustrates the scores of FID and KID on the style transfer task. It is obvious that MDT has good performance in both the subcases and the entire task against
To further evaluate whether the model correctly generates the target images, we employ the VGG-16 [36] which is pre-trained on the ImageNet [8] database and fine-tuned on our dataset, to classify the generated images. To test whether there is a model collapse, we utilize Learned Perceptual Image Patch Similarity (LPIPS) [50] which measures the perceptual distance of output images, to reveal the diversity of different generations. The measurements are shown in Table 2. According to the classification scores, all baseline models almost fail in translating images to the Cezanne domain, and they obviously are biased towards the Monet domain. For our method, it has a more balanced performance in processing all domains. The measurements of LPIPS demonstrate that model collapse may possibly occur in ComboGAN than other methods. For StarGAN and SDIT, the phenomenon of low classification accuracy with high LPIPS value in some domains implies that these two methods are not effective enough in this style transfer task even if they output diverse images against model collapse. For our method, it stably translates the correct target images with good quality across all domains.
TABLE 2  Classification accuracy and mean LIIPS on the style transfer task for each target domain—the higher is the better

| Metric               | Classification accuracy | LIIPS                      |
|----------------------|------------------------|----------------------------|
|                      | Method                 |                           |
|                      | ComboGAN               | StarGAN                   | SDIT | MDT | Real image | ComboGAN | StarGAN | SDIT | MDT | Real image |
| Monet                | 0.9669                 | 0.9583                    | 0.7250 | 0.8083 | 0.9833     | 0.3872   | 0.3964  | 0.3402 | 0.4023 | 0.4586     |
| Cezanne              | 0.0592                 | 0.1908                    | 0.1579 | 0.8487 | 0.9775     | 0.2594   | 0.3226  | 0.3124 | 0.3529 | 0.3737     |
| Van Gogh             | 0.5810                 | 0.3571                    | 0.1048 | 0.8381 | 0.9677     | 0.3929   | 0.3916  | 0.3442 | 0.4007 | 0.4052     |
| Across all domains   | 0.5357                 | 0.5021                    | 0.3292 | 0.8317 | 0.9762     | 0.3465   | 0.3702  | 0.3323 | 0.3853 | 0.3992     |

FIGURE 5  Different methods for face re-lighting among three different illuminations of normal, shadow, and dark. From left to right: input, ComboGAN [3], StarGAN [7], SDIT [42], MDT, and ground truth.

5.3.2  Evaluation of paired samples

In Multi-PIE [10] database, each image has its ground truth. Since some methods synthesized faces with good identity preservation, resulting in 100% face recognition rate which cannot provide a meaningful comparison, we turned to focus on Full Reference Image Quality Assessment (FR-IQA) and utilized two metrics of Feature Similarity Index (FSIM) [48] and Structural Similarity Index (SSIM) [43] for evaluation. The principle of these two metrics is if two images are more similar, the measurement is more close to 1, which in our experiment represents the quality of generated images referenced to their ground truth. Table 3 lists the mean values and standard deviations of FSIM and SSIM for the six subcases and the entire translation. It is obvious that ComboGAN is still ineffective in translating some subtasks, and StarGAN and MDT have comparable performance. For more details of the overall assessment, we present the evaluation curves to illustrate the quality distribution of the generated images. As shown in Figure 6, the vertical axis indicates the percentage of images whose FR-IQA values are higher than the values on the horizontal axis. Our method almost has a higher percentage on all IQA-values.

We do not list the evaluation results of classification accuracy and IPISP for this task, because almost all methods obtained 100% accuracy, and the diversity of the original test sample is low making it meaningless to measure IPISP score. Following the work [24], we use the cosine distance between the features of a fake face and its corresponding real one to measure whether a translated face is meet the target domain style and whether the original identity information is preserved. The ResNet-50 [12] which is a high-performance off-the-shelf face recognition network and is pre-trained on the VGGFace2 database [6], is employed to extract features for all real and fake images. The measurements of mean feature distance are shown in Table 4. It is clear that ComboGAN still has performance bias in some domains, and the other methods are successful in translations and have good preservation of the face identity information after mapping. On the whole, our method has the best results than the baseline models.

5.3.3  Evaluation summary

In summary, ComboGAN has failed in some subcases of the two tasks, which may be due to the difficulty in balancing the
TABLE 3  The mean and standard deviations of FSIM and SSIM under different methods in the face re-lighting task (including each subcase)—for all scores, the higher is the better

| Task                | Metric | FSIM          | SSIM          |
|---------------------|--------|---------------|---------------|
| Normal → shadow     | ComboGAN | 0.602 ± 0.015 | 0.875 ± 0.019 |
| Normal → dark       | ComboGAN | 0.676 ± 0.014 | 0.890 ± 0.004 |
| Shadow → normal     | ComboGAN | 0.865 ± 0.016 | 0.916 ± 0.005 |
| Shadow → dark       | ComboGAN | 0.891 ± 0.005 | 0.931 ± 0.003 |
| Dark → normal       | ComboGAN | 0.907 ± 0.007 | 0.941 ± 0.003 |
| Dark → shadow       | ComboGAN | 0.946 ± 0.007 | 0.957 ± 0.005 |
| Across all domains  | ComboGAN | 0.815 ± 0.129 | 0.920 ± 0.028 |

FIGURE 7  The schemes of different methods to efficiently process multi-domain image-to-image translation problems. The traditional strategy of one-to-one mapping methods is also drawn in this figure.

TABLE 4  Mean feature distance on the style transfer task for each domain. The lower is the better

| Domain   | Method | ComboGAN | StarGAN | SDIT  | MDT  |
|----------|--------|----------|---------|-------|------|
| Normal   | ComboGAN | 0.5781   | 0.1977  | 0.2125 | 0.1717 |
| Shadow   | ComboGAN | 0.5861   | 0.1575  | 0.1622 | 0.1730 |
| Dark     | ComboGAN | 0.1767   | 0.1428  | 0.1591 | 0.1384 |
| Across all domains | ComboGAN | 0.4470   | 0.1660  | 0.1779 | 0.1610 |

5.4  Analysis of different modelling schemes

Existing representative modelling schemes for multi-domain image-to-image translation are shown in Figure 7. Compared to traditional one-to-one mappings, all the methods have improved the modelling efficiency. StarGAN and SIDT have the same and the best efficient modelling scheme due to the introduction of a domain label vector to discriminate different domains. However, there is a trade-off between effectiveness and efficiency, and the most efficient translators do not mean the most effective translated results, as the evaluations mentioned in Section 5.3 supported.

The detailed attributes of these schemes used in different methods are shown in Table 5. The scheme used in ComboGAN, splitting the generator into \( N \) encoders and \( N \) decoders to compose the corresponding processing for mappings among \( N(N-1) \) domain pairs, provides enough capacity for different translations. It is efficient in its testing stage since an input image only needs to be encoded once, actually \( O(d + Ne) \) where \( d, e \) are respectively the complexity of its encoder and decoder. However, there are still too submodels (totally \( 2N \)) making it difficult to balance the training of each domain pair, which may possibly cause failure or model collapse in the generation for some domain pairs. In addition, due to its one-to-many training of each encoder and decoder, SDIT and StarGAN have achieved good visual effects in the face synthesis task, but they are not good enough in the style transfer task. This may be mainly because their single generator architecture has insufficient capacity to control the correct generation in the corresponding domain when the differences among the target domains are large. For MDT, it is successful in the two tasks, and the translated results in each domain are almost superior to those of state-of-the-art methods. However, we observe that, for each method conducted in the two tasks, there are always some subcases with relatively lower performance (e.g. ‘Monet → Van Gogh’, ‘normal → shadow’), which implies it is still a challenge to balance the training for each domain pair.
mapping property, it needs the source domain label to pick up the correct encoder, limiting its real applications.

For the scheme used in StarGAN and SDIT, it introduces a label vector into a single generator, and extremely improves the training efficiency, but it costs much in the testing stage than the other schemes because its indivisible generator needs to encode the input image for each translation every time. Although it shows the ability against model collapse, it faces the difficulty of training domain classification for the label vector. If the label vector cannot accurately discriminate different domains, it may lead to the same dilemma as ComboGAN. In addition to this, the single generator may not have sufficient capacity to simultaneously handle translations across multiple domains with large differences but may be effective in processing translations with small differences, so StarGAN and SDIT are relatively successful in face synthesis but are not effective enough in style transfer. It is possible when there are conflicts for embedding requirements among different translations, a single mapping is unable to reconcile these conflicts, which can be avoided by using separate mappings.

Compared with the scheme used in ComboGAN, MDT needs to train for \( N \) domain pairs, which reduces the difficulty of training balance among the \( 2N \) submodels. Furthermore, the shared encoder can reduce the interference of different domain-specialized information, making the \( N \) decoders better complete their own translation. Compared with the scheme used in StarGAN and SDIT, the decoders also can be regarded as independent label vectors that do not need to train to classify different domains, thus avoiding the difficulty of controlling the correct generation. For the defects of ours compared to other schemes, they are the medium network capacity and medium training complexity, but it is effective and efficient enough for practical applications. Though the scheme quality also depends on the detailed model architecture, it is a key factor in determining the performance of a method.

Since our scheme has only one encoder, we actually assume that all domains should have at least the same latent space. This assumption is reasonable because we can always find an encoder to compress the data to a low-dimensional space to reduce the feature differences among domains. But if the shared embedding will cause a large loss of image content information, it may make the quality of the translated images degrade, or even make the translation fail. On the other hand, even if there is always an effective embedding space among target domains, we cannot assert that our scheme is stable for any number of domains because as the domain number increases, there is less or even no information to share in embedding. We wonder if there will be a maximum number of translatable domains, over which MDT will not converge no matter what specific network structure or training technique is. Fortunately, MDT is sufficient to handle the number of domains involved in practical image-to-image translations.

In general, the typical advantages of the two schemes adopted in the baseline models are the large network capacity and the low modelling cost, respectively. While the typical disadvantages are the unbalanced training and the low network capacity, respectively. For our scheme, it has a trade-off between efficiency and effectiveness, making it more suitable for practical application. Certainly, the common difficulty for all of these solutions is how to further improve the unbalanced performance across all target domains.

### 5.5 Analysis of the objective

To investigate the effectiveness of the proposed two generalized constraints, we isolate the three items of Equation 1 and respectively train the networks to perform the three-domain face relighting task.

Figure 8 shows an example of face re-lighting among dark, normal, and shadow by using four different components to train. It is obvious that the image quality is improved by using reconstruction and identity consistency. We respectively measure the mean values and standard deviations of FSIM [48] and SSIM [43] for different components as shown in Table 6, where indicates the effectiveness of adding the reconstruction loss and
identity loss in training. More details for analysis are drawn in Figure 9. It can be seen that if these two constraints are not involved in training, the image generation quality of MDT will be substantially degraded.

In particular, compared to the identity loss, the reconstruction loss is more conducive to improve the generation, possibly because it has more processing and learning objectives than the other. In identity consistency, a decoder only focuses on learning the feature of its domain through one input image, but in reconstruction constraint, it needs to use the learned domain features to restore $N-1$ fake images from all other domains.

6 | CONCLUSIONS

We propose an effective framework for unsupervised image-to-image translation across multiple domains, called MDT. It consists of a shared encoder and $N$ identical decoders, aiming to reduce the training complexity and the interference of the special source domain information. We also propose two general constraints extended from one-to-one mappings to meet the requirement of multi-domain scenario, which can significantly improve the quality of generated image. According to qualitative and quantitative evaluations, MDT performs favourably against the state-of-the-art multi-domain image translators [3, 7 42] in both the entire task and each sub task, which suggests MDT provides an effective solution for image-to-image translations across multiple domains. In future work, we would like to extend MDT to handle other mixed domains, such as text, video or even audio.

ORCID

Ye Lin https://orcid.org/0000-0003-3247-7084
Keren Fu https://orcid.org/0000-0002-3195-2077
Shenggui Ling https://orcid.org/0000-0002-4667-6431
Peng Cheng https://orcid.org/0000-0002-9632-0636

REFERENCES

1. Abboud, B., et al.: Expressive Face Recognition and Synthesis. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA (2003)
2. Almahairi, A., et al.: Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In: ICML 2018: Thirty-fifth International Conference on Machine Learning, Stockholm, Sweden, pp. 195–204 (2018)
3. Anoosheh, A., et al.: ComboGAN: Unrestrained scalability for image domain translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, Utah, USA, pp. 783–790 (2018)
4. Bhatt, H.S., et al.: Memetically optimized MCWLD for matching sketches with digital face images. IEEE Transactions on Information Forensics and Security 7(5), 1522–1535 (2012)
5. Buikowski, M., et al.: Demystifying MMD GANs. In: ICLR 2018: International Conference on Learning Representations 2018, Vancouver, British Columbia, Canada (2018)
6. Cao, Q., et al.: VGGFace2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi’an, China, pp. 67–74 (2018)
7. Choi, Y., et al.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 8789–8797 (2018)
8. Deng, J., et al.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, Florida, USA, pp. 248–255 (2009)
9. Goodfellow, I.J., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27, 2672–2680 (2014)
10. Gross, R., et al.: Multi-PIE. Image and Vision Computing 28(5), 807–813 (2010)
