

# A Domain Gap Aware Generative Adversarial Network for Multi-domain Image Translation

Wenju Xu, and Guanghi Wang, *Senior Member, IEEE.*

**Abstract**—Recent image-to-image translation models have shown great success in mapping local textures between two domains. Existing approaches rely on a cycle-consistency constraint that supervises the generators to learn an inverse mapping. However, learning the inverse mapping introduces extra trainable parameters and it is unable to learn the inverse mapping for some domains. As a result, they are ineffective in the scenarios where (i) multiple visual image domains are involved; (ii) both structure and texture transformations are required; and (iii) semantic consistency is preserved. To solve these challenges, the paper proposes a unified model to translate images across multiple domains with significant domain gaps. Unlike previous models that constrain the generators with the ubiquitous cycle-consistency constraint to achieve the content similarity, the proposed model employs a perceptual self-regularization constraint. With a single unified generator, the model can maintain consistency over the global shapes as well as the local texture information across multiple domains. Extensive qualitative and quantitative evaluations demonstrate the effectiveness and superior performance over state-of-the-art models. It is more effective in representing shape deformation in challenging mappings with significant dataset variation across multiple domains.

**Index Terms**—Generative adversarial network, image-to-image translation, multiple domains, self-regularization.

## I. INTRODUCTION

Image-to-image translation is to map input images from a source domain to a target domain through the domain-specific characteristic transformation, including inpainting [1], super-resolution [2], colorization [3], etc. The training images are collected separately in terms of domain, and the use of discriminator criticizing domain-specific characteristics provides domain-level supervision for the generator training. Combined with other regularization, the generator learns a translation mapping between the source domain and the target domain.

In recent years, there is a trend to make a trade-off between structure preservation and shape deformation in image translation tasks. One approach is to maintain the global structure and slightly change local textures [4], such as face attributes changing [5]. Another approach is to exploit shape deformation according to high-level semantic constraint [6], like cats to dogs. However, there exist various mappings between each domain pair given different regularization and

W. Xu is with the OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA. Email: xuwenju123@gmail.com.

G. Wang is with the Department of Computer Science, Ryerson University, 350 Victoria St, Toronto, ON M5B 2K3. Email: wangcs@ryerson.ca

This work was partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant no. RGPIN-2021-04244, and the United States Department of Agriculture (USDA) under Grant no. 2019-67021-28996.

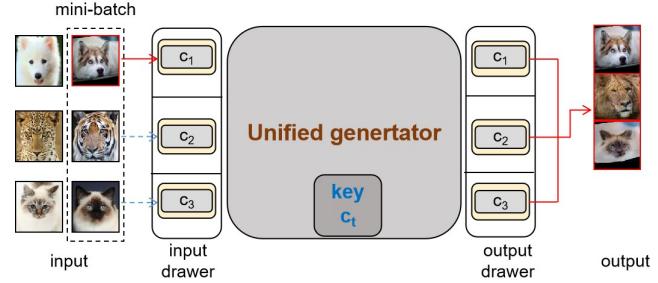


Fig. 1: Introduction of the proposed model. It works as a closet and realizes multiple-domain translation by stacking the drawers. Each input drawer corresponds to one input domain and each output drawer corresponds to one output domain. Users drop observed images into the input drawers, the unified generator can generate the translated images in the output drawers.

training schemes. How to make a balance between these two requirements is still an open problem.

The CycleGAN [7] has shown impressive results on patch-level image translation between two domains, such as translating stripes between zebras and horses. It introduces the most common regularization as the cycle-consistency restraint that learns a reverse mapping for additional supervision. Cycle consistency in the image space tends to preserve pixel-wise correspondence between the input and the output. It is helpful in tasks where the change is mostly on textures (e.g., horse to zebra) but recent findings [8], [9] have shown that being able to invert a mapping does not necessarily lead to successful translations for some domains, such as cats to dogs. It prevents the model from learning mappings that involve large shape differences. To promote shape deformation, previous works seek to increase the receptive field to discriminate on both the global and local features, match statistics of discriminatory features, penalizing distances in the latent space. However, this additional regularization introduces extra training parameters and computational costs, which are critical when dealing with multi-domains image translation.

In this work, we propose a unified model, named as UMIT, for multi-domain image translation. This model maintains consistency over both the global shape and the local texture in each domain by exploiting domain-specific information [10]. We introduce the input and output drawers to preserve local textures. The working mechanism is illustrated in Figure 1, where we put input images from one domain in the same drawer and generate the outputs from the corresponding output drawers. This mechanism facilitates disentangling generations in one framework. Furthermore, we efficiently integrate the discriminator with a multi-scale classifier and dilated convolutional layers. With a small portion of additional parameters,

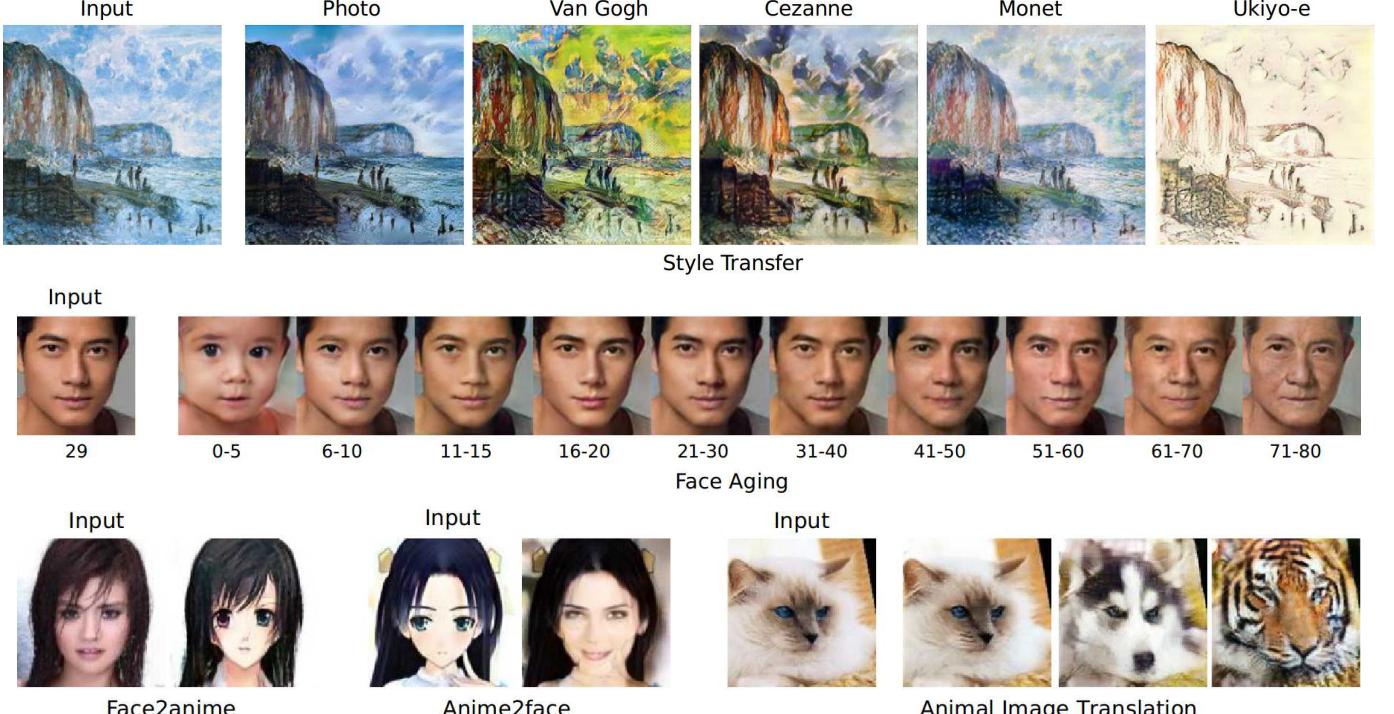


Fig. 2: Some examples of the translated images by the proposed model across multiple domains with large domain gaps. (*top*) Collection image transfer. (*center*) Face aging. (*bottom*) face transformation.

it significantly increases the receptive field for low-frequent shape change. Instead of exploring the cycle-consistency, we take a perceptual loss to minimize the distance between the input image and the translated image and expect the output is visually similar to the input in terms of foreground and background. Based on this observation, our model proposes to utilize a single generator that is trained with a self-regularization term to enforce perceptual similarity between the output and the input, making the training process simpler and more efficient.

Extensive experiments demonstrate that the proposed model works surprisingly well in a wide variety of image translations involving multiple image domains, achieving superior qualitative and quantitative results as shown in Figure 2. We quantitatively and qualitatively demonstrate that our approach is more efficient and more effective by learning a unified model on challenging datasets, where the data comes from multiple resources and contains large shape deformations. We also carefully analyze the relation between our approach and CycleGAN and think that cycle-consistency is not a necessary assumption.

## II. RELATED WORK

Deep neural networks have shown great potential in dealing with real-world tasks [11], [12], [13], [14], [15], [16], [17]. Many deep learning based methods were proposed for image content understanding [18], [19] and image content generation tasks [20], [21], [22]. Especially, the content conditioned image synthesis task, which generates images based on the input content images via learning a mapping function between the input domain and target domain, has attracted a lot of

attentions. It requires to not only create realistic synthesized images but also preserve the content structural similarity. Lots of works now are being introduced for this purpose.

**Generative Adversarial Network.** Generative adversarial networks (GANs) [23] have shown impressive results in image generation [22], [24], image translation [25], [26], [27], super-resolution imaging [2], [28], and face image synthesis [29], [30]. These GAN models are trained to minimize the discrepancy between distributions of the training data and unobserved generations by adversarially learning between two neural networks. WGAN [31] and WGAN-gp [32] use W-distance to remedy the model collapse. Several recent efforts explore conditional GAN given in various image translation tasks to learn the mapping conditioned on prior constraints such as labels, text embedding, or images. Our work focuses on using a unified GAN model for multi-domain image translation.

**Image-to-Image Translation.** The goal of image-to-image translation is to learn a mapping between paired/unpaired image collections. Pix2pix [4] handles this task by learning a one-directional mapping from the source to the target domain with paired images as supervised signals. But it is not suitable for tasks where paired images are not available. To alleviate the burden of obtaining data pairs, unsupervised image-to-image translation approaches have been proposed [7], [33], [34], [35], which resort to the cycle consistency constraint for additional supervision. Similar ideas can also be found in [35], [36], [37].

Recent multimodal image translation methods, such as BiCycleGAN [25], MUNIT [38], DRIT [39], model a distribution of possible outputs by producing diverse translated samples instead of one deterministic sample for each single input image. This one-to-many mapping seeks to cover all

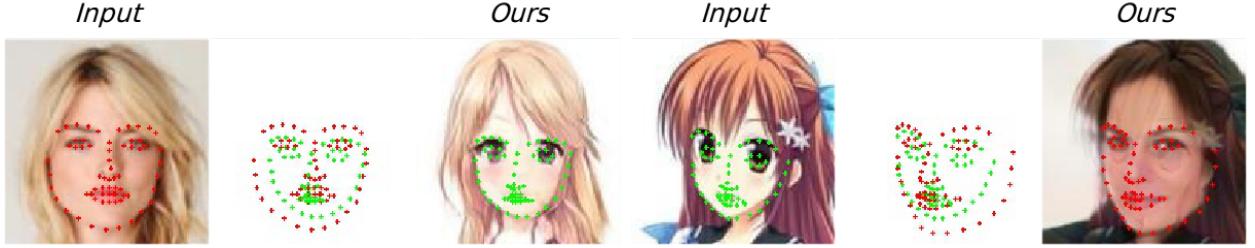


Fig. 3: Landmarks on face images. Our approach translates the structure and texture appearance variations between the human and Anime domains (left: input; right: translation).

possible mappings between two domains. However, few of them are effective to learn a map when the cross-domain gap (e.g., object shape) becomes large. The GANmorph [40] handles this task by utilizing dilated convolutions and multi-scale features in both discriminator and generator for larger receptive fields. WarpGAN [41] manipulates the predicted control points to warp the photo into a caricature. Our model does not rely on cycle-consistency and seeks to bridge the domain gap by exploring multi-scale information (Figure 3).

Another type of approaches [42], [43], [44] separately learns an appearance model and a structure model. The appearance model is to estimate the structure representation (landmarks or skeleton) of the object and synthesize a new sample-based on (manipulated) landmarks or skeleton. While the structure model is to manipulate the landmarks or skeleton based on pre-defined structural relation. In this sense, this type of approaches does not require the generative models to handle the shape deformation, but only synthesize samples conditioned on labels. Therefore, these approaches require image-label pairs to train both the appearance model and the structure model. TransGaGa [9] proposes to learn the appearance and geometry latent space separately. However, the learned geometry latent representation is dominant, tending to generate new images purely conditioned on the geometry latent code without appearance preserved. In addition, this type of approaches deals with the geometry and appearance separately and fully converts the appearance of the source domain into that of the target domain without preserving the consistency between the input images and output images. However, our proposed model aims at learning a unified model in an unsupervised or self-supervised manner. Our task is more challenging in terms of “unsupervision”. In a direct way, we seek to promote the capacity of the generative model for shape transformation and consistency preservation. For illustration, the landmarks of both the human face and the generated Anime face are demonstrated in Figure 3. The face shapes are depicted by the landmarks. Our model does not rely on the landmarks but is able to transform the face and preserve the semantic consistency.

**Multi-Domain Image Translation.** This task is to translate images into different domains using a unified model. Unlike learning the mapping between two domains, it deals with the scalability of unsupervised image translation methods. Recent works have shown impressive results in multi-domain image-to-image translation using one single model [45], [46], [47].

Based on CycleGAN, StarGAN [45] proposes a unified model that shares the generator and discriminator by all domains. Additionally, it integrates a classifier to predict domain probability. UFDN [46] and GANimation [48] learn a continuous latent manifold to perform a continuous cross-domain image translation and manipulation on human face dataset. However, these multi-domain models are more sensitive to the domain gap since training a classifier on the existing image translation dataset is challenging given the number of images. The ill-trained classifier fails to provide regularization when training the generator. To avoid this, ComboGAN [49] and singleGAN [50] employ multiple discriminators to train one unified generator, while they significantly increase the number of trainable parameters. The most recent FUNIT [6] focuses on few-shot learning. Based on MUNIT, it proposes a unified discriminator that works well on a large number of domains with limited samples. Nevertheless, the performance of these methods is limited to the dataset that contains easy-recognized image domains.

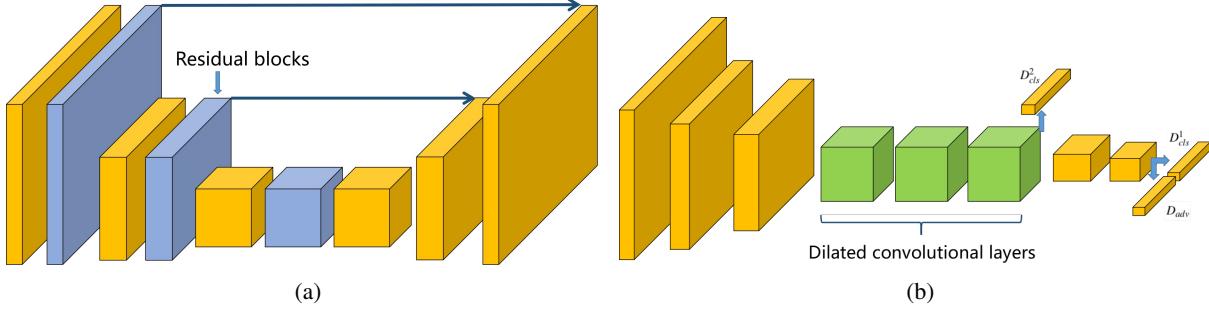
### III. PROBLEM FORMULATION

Given image sets  $\{X_c\}_{c=1}^N$  from  $N$  domains, our goal is to learn a pair of generator-discriminator  $\{G-D\}$  that transforms domain-specific properties across all of the domains. Additionally, we integrate a multi-scale classifier into the discriminator with most parameters shared. Let us use subscripts  $s$  and  $t$  to indicate the source and target domain, and  $c$  for domain label. This applies to represent transformation between any domain pair (*i.e.*, to map input image to one of the target domains and vice-versa). The problem is formulated as  $G(x_s, c_t) \rightarrow x_t$ . We will elaborate the proposed approach in this section.

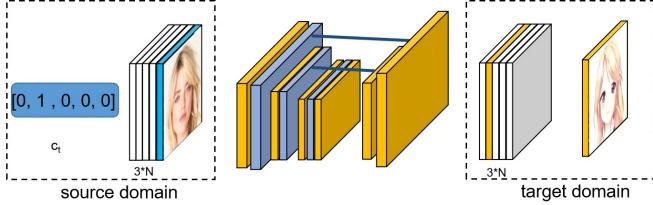
#### A. Network Architecture

**Unified Generator.** The architecture of our generator is illustrated in Figure 4a. It integrates residual blocks at multiple layers into a Unet, allowing the network to aggregate multi-scale information for the transformation of both higher- and lower-spatial resolution features. This structure takes both the images and the domain labels as inputs. In addition, we propose a novel input-output drawer mechanism, allowing the unified generator to disentangle local textures at pixel-level when all learnable parameters are shared.

**Input-output Drawer.** The input-output drawer structure, as shown in Figure 1, works as a domain container preserving



**Fig. 4:** (a) The generator is a combination of Unet and Resnet. The skip connections and residual blocks are combined via concatenation as opposed to addition. (b) The discriminator architecture is a fully-convolutional network. Each colored block represents a convolutional layer; block labels indicate filter sizes. In addition to the global context from the dilations, two additional classifiers are employed to enforce the network’s view of the domain-specific feature.



**Fig. 5:** The architecture of the input and output drawers.

domain-specific local textures. We stack the same number of drawers as involved domains for the input or output images. Let  $B$ ,  $C$ ,  $W$ , and  $H$  refer to the batch size, channel dimension, feature width, and height, respectively;  $c_t$  indicates the target domains and  $c_s$  represents the source domains. Given images  $x_s \in \mathbb{R}^{B \times C \times H \times W}$  and the initial input drawer  $T_{in} \in \mathbb{R}^{B \times [C \times N] \times H \times W}$ , we assign each image separately to the input drawer  $T_{in}$  as:

$$T_{in}[i, C \times c_s : C \times (c_s + 1)] = x_s[i] \quad (1)$$

Then the input of generator is formed as a concatenation of  $[T_{in}, c_t] \in \mathbb{R}^{B \times (C \times N + c_t) \times H \times W}$ . We perform a similar operation to receive output images from the output drawer given  $T_{out} = G(T_{in}, c_t)$ . The final images can be obtained by selecting the one indicated by the target domain label:

$$x_t[i] = T_{out}[i, C \times c_t : C \times (c_t + 1)] \quad (2)$$

where  $i$  is the index used to iterate in the mini-batch and  $C \times c_t$  is to select the target image channel-wisely. In this way, the generator can focus exclusively on the domain-specific feature translation (global shape information), leading to sharper and more realistic synthetic images. It does not need to learn separately a mapping for each domain pair. This process is depicted in Figure 5.

**Multi-scale Discriminator.** Using one unified model for Multi-domain image translation poses a significant challenge to the design of GAN discriminator. It requires to have a large receptive field for domain-specific global features. This would require a sophisticated architecture to capture the global features and differentiate the features according to the domain information. The patch discriminator has been proven to be effective for local texture translations. However, this patch-based view limits the network’s awareness of global spatial information, which limits the generator’s ability to perform

coherent global shape changes. The dilated discriminator encourages global shape changes with a large receptive field. But these structures are not effective in the scenario of multi-domain translation. To alleviate these issues, we combine classifiers at multi-scales with the discriminator for domain-specific feature recognition, as shown in Figure 4b. This structure can provide globally consistent information for the generator training.

### B. Objective Function

We aim at generating domain-consistent images with high visual quality in a unified model. We randomly assign the target domains for input images. The discriminator is to distinguish between the input images and the translated images, while the integrated classifier produces probability distributions over domain labels. We ignore the input-output drawer for expression simplicity as  $G(x_s, c_t) \rightarrow x_t$ . Our goal is to optimize the following objectives.

**Adversarial Loss.** The original GAN matches two distributions of the real and fake images. However, the adversarial loss is potentially not continuous with respect to the parameters in the generator. This can locally saturate and lead to vanishing gradients in the discriminator. As a result, it is extremely difficult to train the GAN model. We take advantage of the Wasserstein GAN with gradient penalty (WGAN-GP) [32], [31] to stabilize the training process. The adversarial loss is written as

$$\begin{aligned} \mathcal{L}_{adv} &= \mathbb{E}_{x \sim X}[D_{adv}(x)] + \mathbb{E}_{x \sim X}[-D_{adv}(G(x_s, c_t))] \\ &\quad + \lambda_{gp} \mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{adv}(\hat{x})\|_2 - 1)^2] \end{aligned} \quad (3)$$

where  $\hat{x}$  is the linear interpolation between a real image and the generated one.

**Multi-scale Domain Classification Loss.** The results using  $L_{adv}$  alone tend to have annoying artifacts, because the search for a local minimum of  $L_{cls}$  may go through a path that resides outside the manifold of natural images. Thus, we combine  $L_{cls}$  with  $L_{adv}$  as the final objective function for  $D$  to ensure that the search resides in that manifold and produce photo-realistic and domain-preserving images.

$$\mathcal{L}_D = \mathcal{L}_{adv} + \sum_{i=1}^2 \lambda_i \mathcal{L}_{cls}^i \quad (4)$$

We employ a typical softmax cross-entropy loss and regularize the outputs with the multi-class cross-entropy loss

$$\mathcal{L}_{cls}^i = \mathbb{E}_{x_s \sim X} [\log D_{cls}^i(c_s | x_s) + \log(D_{cls}^i(c_t | G(x_s, c_t)))] \quad (5)$$

where the term  $D_{cls}^i(x)$  represents a probability distribution over all domains.  $\tilde{x}$  represents the generated images. Theoretically, minimizing  $L_{cls}$  in the adversarial training scheme would learn a mapping  $G$  that produces outputs consistent to the target domain. However, if the capacity is large enough, a network can map the input images to any randomly generated samples in the target domain. If the classifier is ill-trained, the network fails to map the input images to the target domain. Thus, other regularizations need to be introduced to ensure the learned function  $G$  maps the input to the desired output.

**Perceptual Self-Regularization Loss.** To further constrain the learned mapping, the translated image should preserve the visual characteristics of the input image including color and pose, etc. In other words, the output and input need to share perceptual similarities. To this end, we impose a perceptual self-regularization constraint, which is modeled by minimizing the distance between the input and output images. The perceptual loss is used for semantic consistency, which promotes the model to preserve the semantically related features, such as the color and the pose, while allowing us to change the domain-specific features. In contrast, the cycle loss seeks for pixel-level consistency. It either fails to maintain the semantic consistency or prevents the changes in the local patch.

We use a pre-trained VGG-19 network to compute style losses at multiple levels and one content loss in a way similar to the prior work. At the training stage, the target domain labels are randomly generated as  $c_t = c_s[perm]$ , where  $perm$  stands for randomly generated permutation of data in the mini-batch. We attempt to minimize the perceptual loss of the translated image  $x_t$  by considering the input source image  $x_s$  as the content image while the  $x_s[perm]$  as the style image. Similar to [2], [51], we use the pre-trained VGG-19 to compute the loss.

$$\mathcal{L}_{vgg} = \mathcal{L}_c + \lambda \mathcal{L}_s \quad (6)$$

where the content loss is the Euclidean distance between the target features and the features in the output image.

$$\mathcal{L}_c = \|\phi(x_t) - \phi(x_s)\| \quad (7)$$

In our experiments,  $\phi$  represents the features of  $relu_{41}$  layers. Our style loss only matches the mean and standard deviation of the style features.

$$\mathcal{L}_s = \sum_{i=1}^L \|\zeta(\psi_i(x_t)) - \zeta(\psi_i(x_s[perm]))\|_2 \quad (8)$$

where  $\zeta$  is the *Gram matrix* and each  $\psi_i$  denotes a layer in VGG-19 used to compute the style loss. In our experiments we use  $relu_{21}$ ,  $relu_{31}$ ,  $relu_{41}$  layers with equal weights.

**Identity Preserving Loss.** As observed by previous works, the image translation GAN models lean to recognize the difference of color distributions instead of the texture/shape

difference. To ensure the color consistency, we take the identity-preserving loss.

$$\mathcal{L}^{identity} = \mathbb{E}_{x \sim X} \|x - G(x, c_t)\|_1 \quad (9)$$

**Final Objective Function.** Through the above analysis, we train the proposed network by optimizing the following objective function.

$$\min_G \max_D \mathcal{L}_D + \lambda_{identity} \mathcal{L}_1^{identity} + \lambda_{vgg} \mathcal{L}_{vgg} \quad (10)$$

**Remarks.** To demonstrate the effect of different losses, we visualize the difference between the cycle loss and our perceptual self-Regularization loss in Figure 6. The cycle loss seeks for pixel-level consistency. The main issue of cycle-consistency is that it assumes the translated images contain all the information of the input images in order to reconstruct the input images. This assumption leads to the preservation of source domain features, such as traces of objects and texture, resulting in unrealistic results. Moreover, the cycle-consistency constrains shape changes of source images. Our model proposes to use a single generator to learn directed mappings with a self-regularization term that enforces perceptual similarity between the output and the input, together with an adversarial term that enforces the output to appear like drawn from the target domain. The translated images preserve both high-level attributes and local texture from the target domain, relaxing the requirement of preserving source domain features for pixel-level consistency.

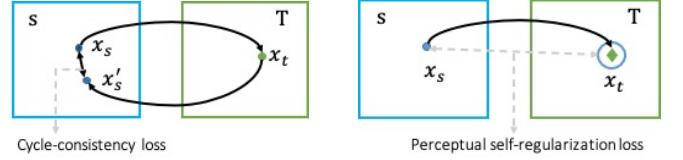


Fig. 6: Comparison between the cycle consistency loss and perceptual self-regularization loss. Given two image domains S and T represented by blue and green rectangles, any point inside a rectangle represents a specific image in that domain. (Left): given a source image  $x_s$ , cycle-consistency loss requires the image translated back,  $x'_s$ , to be identical to the source image  $x_s$ . (Right): given a source image  $x_s$ , we synthesize images  $x_t$  with shape variants (from a round dot to a rhombus) in the target domain. Instead of requiring the image to be translated back, we minimize the distance between  $x_s$  and  $x_t$ , so that  $x_t$  is encouraged to preserve the features of the original image  $x_s$ .

## IV. EXPERIMENT

### A. Datasets

We evaluate our model on two types of datasets. The first type of datasets consists of two domains used to demonstrate the effectiveness of the model. These include:

**Anime  $\leftrightarrow$  Human.** The Anime images are obtained from the Getchu dataset. We use anime-face detector to extract a total of 8,034 anime face images. The human face images are selected from the CelebA dataset. We only use the frontal face images without hats and glasses, and randomly select 14,000 human face images.

**Photo  $\leftrightarrow$  Portrait.** This dataset is used by DRIT. We follow the train and test split used by the authors.

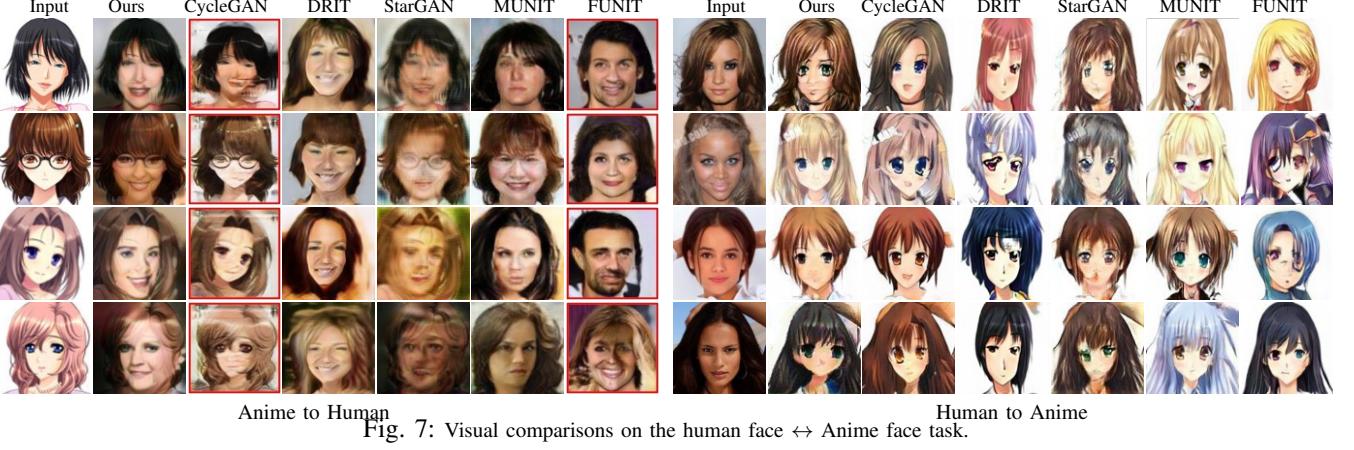


Fig. 7: Visual comparisons on the human face  $\leftrightarrow$  Anime face task.

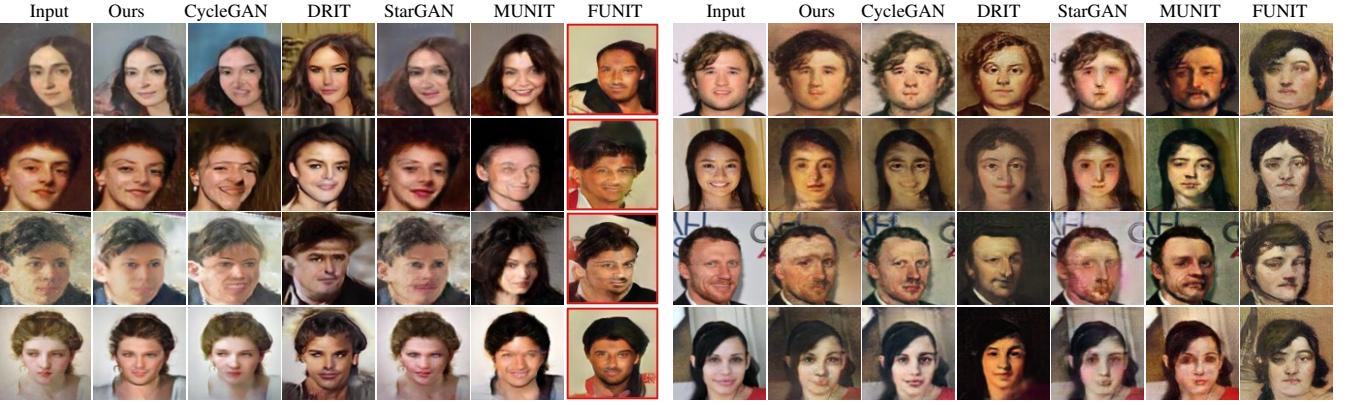


Fig. 8: Visual comparisons on the human  $\leftrightarrow$  Portrait task.



Fig. 9: Qualitative ablation study on the Anime  $\rightarrow$  human task. The ablation study is performed by removing parts of the network. The patch Dis indicates the model is trained with a patch-GAN [4] instead of our discriminator.

The second type of datasets involves multiple image domains to evaluate the efficiency of our model. They include

**Animal Image Translation.** To create a multi-domain translation task, we extend the cat $\leftrightarrow$ dog dataset used by DRIT with another image collection, the big cat. We collect the big cat images from the ImageNet dataset.

**Face Aging Dataset.** The dataset is provided in [30]. The age is divided into ten categories, i.e., 0-5, 6-10, 11-15, 16-20,

21-30, 31-40, 41-50, 51-60, 61-70, and 71-80. Therefore, we can consider each of the 10 elements as one image domain. In total, there are 10 domains involved.

**Collection Style Transfer.** We use the dataset from CycleGAN [7], which contains Monet, Cezanne, Van Gogh, Ukiyo-e, and natural scene images. We train our model once to learn all the mapping among these 5 image domains.

### B. Baseline Methods

We compare our model with different types of state-of-the-art-models: CycleGAN [7], DRIT [39], MUNIT [38], StarGAN [45] and FUNIT [6]. All of these methods can perform unsupervised image-to-image translation without paired

TABLE I: Quantitative comparison. We use FID (lower is better) and user study (higher is better) to evaluate the quality of generated images.

| Setting             | Human2Anime |      | Anime2Human |      | photo2portrait |      | portrait2photo |      |
|---------------------|-------------|------|-------------|------|----------------|------|----------------|------|
|                     | FID         | AMT  | FID         | AMT  | FID            | AMT  | FID            | AMT  |
| real data           | 1.84        | 0.50 | 2.43        | 0.50 | 1.12           | 0.50 | 2.35           | 0.50 |
| CycleGAN            | 17.37       | 0.17 | 38.74       | 0.00 | 10.13          | 0.01 | 18.46          | 0.09 |
| DRIT                | 23.61       | 0.01 | 37.56       | 0.02 | 8.23           | 0.21 | 20.63          | 0.03 |
| StarGAN             | 22.18       | 0.01 | 32.57       | 0.00 | 9.26           | 0.02 | 19.14          | 0.08 |
| MUNIT               | 17.34       | 0.10 | 39.67       | 0.05 | 9.45           | 0.07 | 20.54          | 0.12 |
| FUNIT               | 22.45       | 0.01 | 27.25       | 0.12 | 9.84           | 0.01 | 24.25          | 0.02 |
| Ours                | 15.47       | 0.20 | 24.47       | 0.31 | 8.97           | 0.18 | 17.12          | 0.16 |
| w/o drawer          | 18.34       | -    | 28.57       | -    | 9.54           | -    | 17.81          | -    |
| w/o $\mathcal{L}_c$ | 16.25       | -    | 26.13       | -    | 10.23          | -    | 17.47          | -    |
| w/o $\mathcal{L}_s$ | 17.34       | -    | 27.63       | -    | 9.85           | -    | 17.24          | -    |
| patch Dis [4]       | 21.44       | -    | 31.32       | -    | 11.54          | -    | 20.19          | -    |

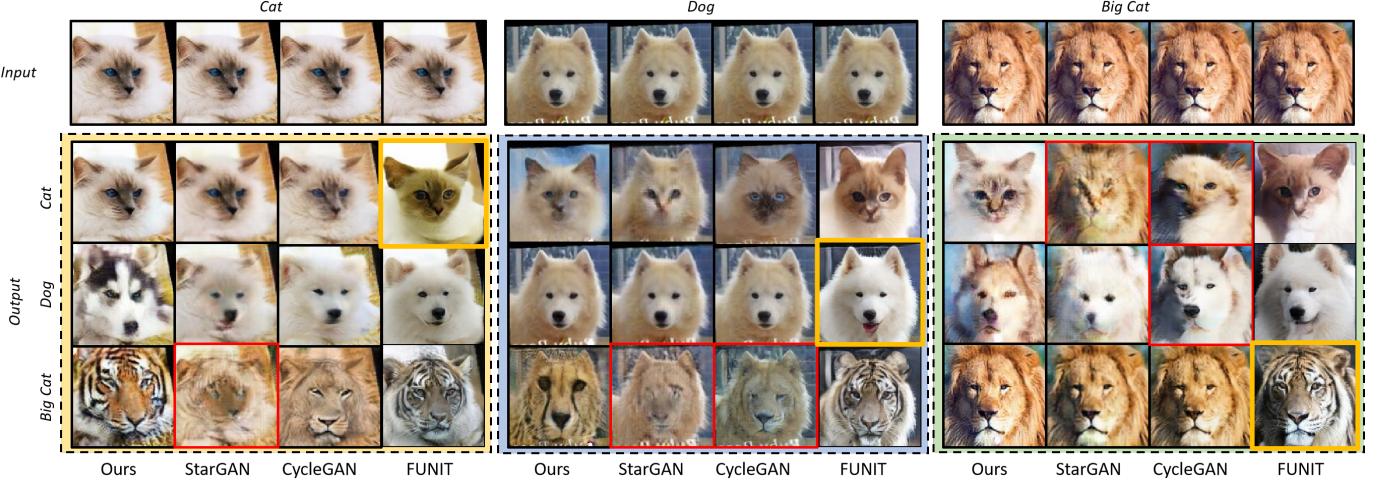


Fig. 10: Results on animal dataset. The red rectangles highlight the failure cases in translation, while the yellow rectangles highlight the failure cases in reconstruction.

TABLE II: Quantitative comparison on the Animal set. The FID and CA scores of the FUNIT are not reported by the authors.

| Setting             | AMT   | IS   | FID   | CA   |
|---------------------|-------|------|-------|------|
| real data           | 0.500 | 2.93 | 0.47  | 0.99 |
| CycleGAN            | 0.08  | 2.26 | 25.74 | 0.82 |
| StarGAN             | 0.02  | 1.92 | 33.57 | 0.68 |
| FUNIT               | 0.24  | 2.82 | -     | -    |
| Ours                | 0.16  | 2.62 | 17.47 | 0.93 |
| w/o drawer          | -     | 2.47 | 19.54 | 0.78 |
| w/o $\mathcal{L}_c$ | -     | 2.51 | 18.31 | 0.83 |
| w/o $\mathcal{L}_s$ | -     | 2.53 | 18.15 | 0.84 |
| patch Dis [4]       | -     | 2.36 | 20.45 | 0.69 |

training dataset. In particular, StarGAN and FUNIT belong to multi-domain image translation methods. DRIT and MUNIT are multimodal models. These two models generate diverse samples that are not faithful to the input image. Thus, they are not compared in the multi-domain image translation tasks. We train these baseline models using their public implementation.

### C. Evaluation Metrics

In unsupervised image-to-image translation, there is no per-pixel ground truth. We utilize the following metrics for evaluation: the Classification Accuracy (CA) and Inception Score (IS) [52], which is designed to assess multi-classes generation by measuring the probability of the generated images belonging to the distribution of the real data; the FID score [53] and AMT perceptual test [3] are used to measure the visual quality of the generated images.

### D. Implementation Details

All the models are trained using Adam optimizer [54] with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ , and we train our model for 400,000 iterations with a batch size 16, learning rate 0.0001 and the learning rate is reduced by a factor of 10 after 200,000 iterations. In our experiments, We set  $\lambda_{gp} = 10$ ,  $\lambda_1 = 1$ ,  $\lambda_2 = 0.5$  and  $\mathcal{L}_1^{identity} = 10$  and  $\mathcal{L}_{vgg} = 0.5$  for all of the tasks. For baseline methods, we use their original default parameters for training.

### E. Transformation between Two Domains

We first analyze model effectiveness on the unsupervised tasks involving two domains. These are photo $\leftrightarrow$ portrait and human $\leftrightarrow$ anime.

**Qualitative Comparison.** Learning a mapping between these domains is a more challenging task than facial attribute translation due to the domain difference between these different types of faces. As illustrated in Figure 7, the synthesis of StarGAN and CycleGAN suffers from artifacts and has noticeable trace of the source objects and textures. With the cycle-consistency constraint, the generated images are able to preserve the head direction but not successfully translated to the target domain. We can observe a big improvement in our approach for head pose and color consistency, and promote a large degree of shape deformation such as changing the vertical height of the faces. The multimodal methods, MUNIT and DRIT, yield diverse samples without preserving the color and pose of input images. The comparison on photo $\leftrightarrow$ portrait task is shown in Figure 8, StarGAN, and CycleGAN generate recognizable results in the target domain due to the similarity between these two domains. Our approach achieves competitive results with all of the baselines.

**Quantitative Comparison.** To quantify the performance of visual quality, we first compute the FID scores between the generated data and the real data. Table I lists the quantitative comparisons. Overall, our model achieves a large degree of shape transformation on a challenging dataset. In particular, even though CycleGAN obtains reasonable performance in translating human to anime, it fails to translate the anime to human, indicating a limitation of dealing with large domain gaps. We proceed to conduct a user study to evaluate the proposed algorithm against the state-of-the-art style transfer methods. Participants are shown the translated images and asked to select the best-translated images. All test images of each dataset are rated by different participants. Our method obtains the best scores.

**Ablation Study.** We perform the ablation study by removing parts of the network to validate the effectiveness of the

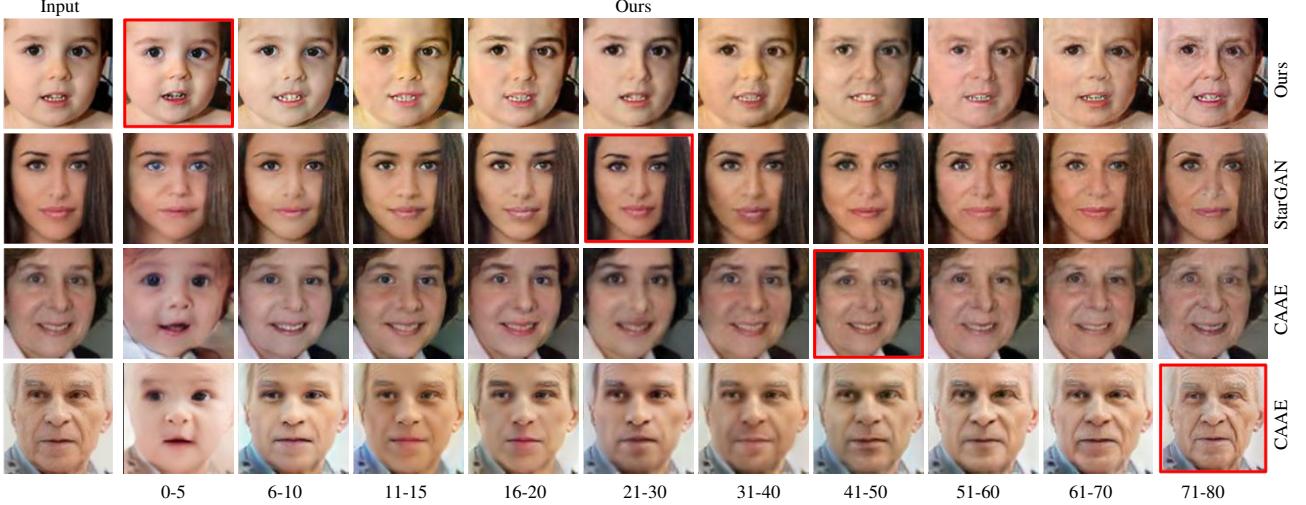


Fig. 11: Samples generated by our method. The red rectangles indicate the input images are belong to the corresponding age range.

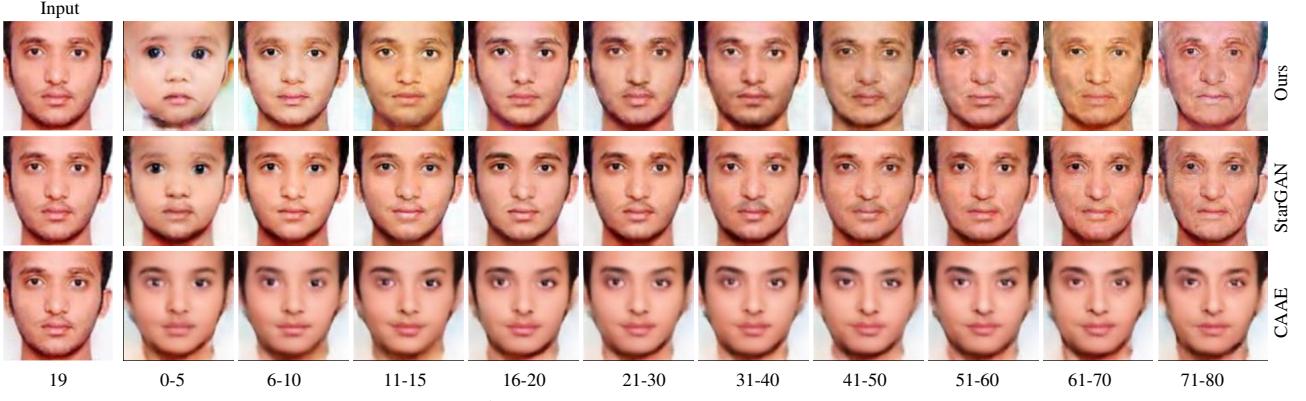


Fig. 12: Results on face aging dataset.

TABLE III: Quantitative comparison on face aging dataset.

| Setting             | FID   | CA   | IS   | AMT  | Inter-class | Intra-class |
|---------------------|-------|------|------|------|-------------|-------------|
| CAAE                | 10.95 | 0.42 | 3.24 | 0.06 | 18.44       | 8.45        |
| StarGAN             | 4.18  | 0.78 | 4.56 | 0.13 | 23.28       | 5.25        |
| Ours                | 4.26  | 0.86 | 5.12 | 0.31 | 25.87       | 4.43        |
| w/o drawer          | 4.55  | 0.76 | 4.54 | -    | -           | -           |
| w/o $\mathcal{L}_c$ | 4.45  | 0.81 | 4.63 | -    | -           | -           |
| w/o $\mathcal{L}_s$ | 4.34  | 0.79 | 4.85 | -    | -           | -           |
| patch Dis [4]       | 5.23  | 0.69 | 4.11 | -    | -           | -           |

proposed network. We show the results without the input-output drawer, content perceptual loss  $\mathcal{L}_c$ , or style perceptual loss  $\mathcal{L}_s$ , respectively. Without using the input-output drawer, the performance is significantly degraded as shown in Figure 9. The content perceptual loss works to preserve visual characteristics of the input image including color and pose, etc. The style perceptual loss increases the domain consistency between the generated images and images in the target domain. Training with the patch discriminator [4] neither exhibits the shape deformation in the content nor faithful style variation. Quantitative results of the FID score is reported in Table I, which indicates that these techniques are important for the generation quality and pose preservation.

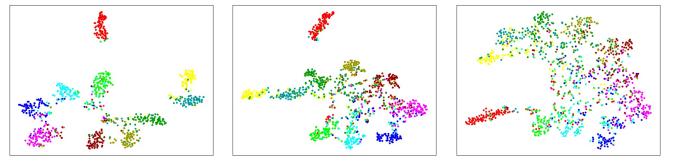


Fig. 13: Feature distribution comparison. The feature is extracted from the pre-trained classification network on the translated images. one color per class.

#### F. Translations across Multiple Domains

We proceed to evaluate our model on applications that contain multiple domains.

**Animal Image Translation.** The motivation of our work is to learn a unified model for multi-domains image translation with large domain gaps. We hope our model is both efficient and effective to handle multi-domain image translation tasks. This animal image translation task consisting of three domains is characterized by significant cross-domain shape variation. In Figure 10, we demonstrate the multi-domain translation results of different methods. Note that we do not train the UNIT model on this dataset, instead, we employ the pre-

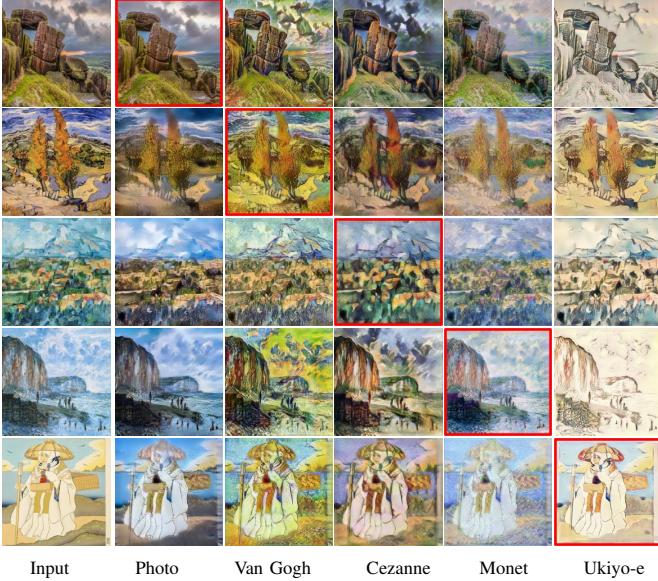


Fig. 14: The multi-domain translation task of multi-style image generation. The first column is the real images and the rest columns are the generated results in different artistic styles.

trained model, which was trained on 139 categories of animal images by the authors. The FUNIT generates appealing results in translation, but the reconstructed animal faces are much different from the original inputs. Our proposed model achieves competitive results against the baselines. First, our model captures domain-related details. Given an input image, the translation outputs cover multiple fine-grained animal characteristics in the target domain. The shape of an animal has undergone significant transformations, but the pose is largely preserved. Second, our model faithfully reconstructs the input images while FUNIT generates a new similar sample.

The quantitative results are listed in Table II. We compare the overall model capacity with other baselines. The CycleGAN model is inefficient for tasks involving multiple domains. For this task, it needs to train 3 times separately to cover all the mappings between each domain pair. The StarGAN is efficient in learning a unified mapping. However, it fails to disentangle domain-specific features. Our model advances in translating domain-consistent textures across multiple domains. Overall, the FUNIT achieves the best metric scores, we consider this as the benefits of large numbers of training data across the 139 domains.

**Human Face Aging.** The Age Progression/Regression requires the generative model to directly produce images with the desired age attribute. To successfully generate the aging images, this task requires both textures and global shapes transformation. The conditional adversarial autoencoder (CAAE [30]) learns a face manifold for age progression and regression. In this work, we consider it as a domain translation task. Instead of learning a latent manifold, we consider each age period as one domain, and translate the input image to the corresponding age domain. We take the StarGAN and the CAAE model as baselines. We present the qualitative performance in Figure 12. The results show that our model can translate both the texture and shape. It is noteworthy that

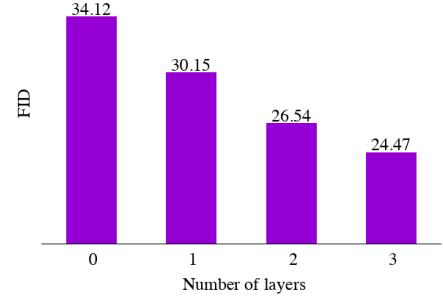


Fig. 15: The accuracy (FID score) for the discriminator with respect to different numbers of dilated convolutional layers on the anime to human task.

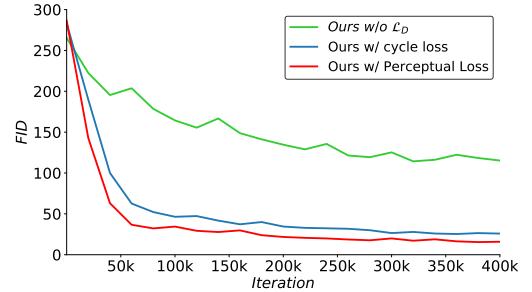


Fig. 16: Quantitative results of the FID scores at the training stage.

our mode translates the input image to an infant of age 0-5 with face shape transformation while other baselines suffer the difficulties. We compare the Feature distribution comparison using t-SNE [55] in Figure 13. We observe that the translated images generated by our method are clustered in the feature space.

**Collection Style Transfer.** The previous tasks consist of domains that are distinct from each other. To demonstrate the effectiveness of our model. We employ our model for style transfer. We use images from 4 artists and consider each painting style as one domain. This task is challenging because the style images of each artist comprised several distinct styles. It is hard to extract domain-specific features. We present stylized results of the evaluated methods in Figure 14. Our model generates impressive images given input images from different domains. As we can see, the results of season are acceptable but the style translation of paintings is not so good accompanied by artifacts and blurriness. Compared to previous tasks, image style translation needs more variations on texture and color, a single model might be difficult to simultaneously handle all styles with large variation. However, our model has the potential for further explorations and extensions.

## V. DISCUSSION

In this paper, we propose a unified model to handle different issues in image-to-image translation tasks. Our multi-scale discriminator is the driven power to enforce the output to appear like drawn from the target domain (*e.g.* mappings between

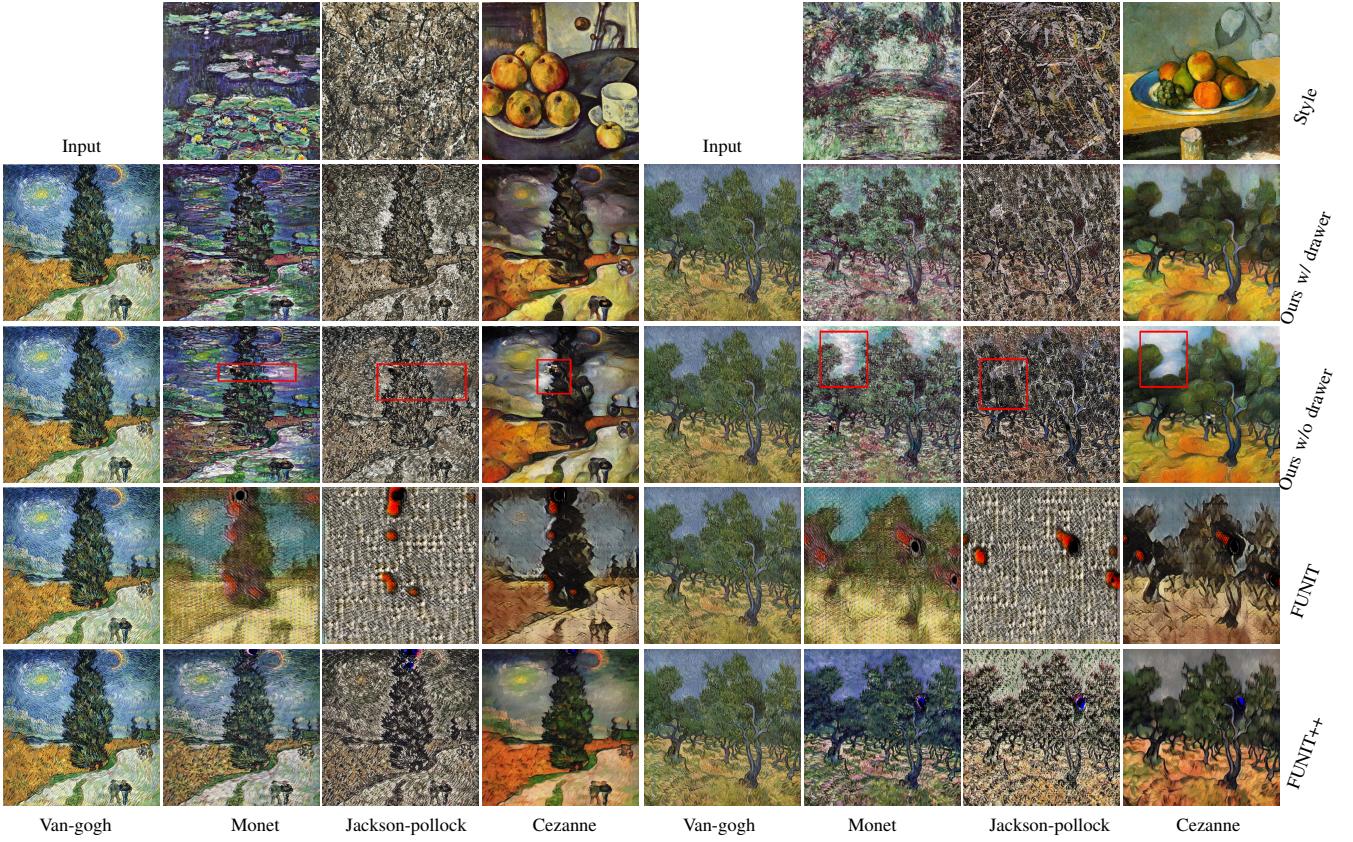


Fig. 17: The artistic style images synthesized in high-resolution ( $500 \times 500$ ). From left to right: the input images and different artistic styles. Images in the second row are results from the model with the input-output drawer; Images in the third row are results from the model without the input-output drawer; Images in the fourth row are results from FUNIT. FUNIT++ refers to a FUNIT variant trained under a stronger reconstruction constraint. Specifically, we increase the weight of reconstruction loss from 0.1 to 1.0. The red rectangles highlight the regions with artifacts or missing details. Please zoom in for better visualization.

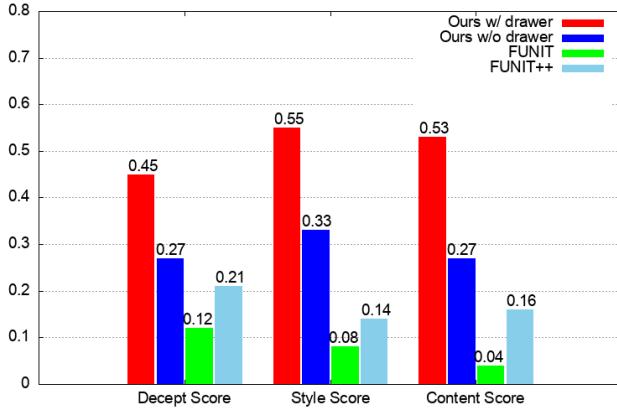


Fig. 18: Quantitative comparison on style transfer task.

different domains). Since there could be an unlimited number of mappings, we introduce the perceptual self-regularization to encourage perceptual similarity and the input-output drawer to improve the quality of local texture. To verify the effectiveness of each module in the proposed framework, we have performed a further study to compare the performance of our full model with different variants.

#### A. Effect of the Dilated Convolutional layer in Discriminator

The dilated discriminator encourages global shape changes with a large receptive field. When combining classifiers at multi-scales with the discriminator for domain-specific feature recognition, this discriminator can provide globally consistent information for the generator training. To clarify, we compare our model to its variants of discriminator with different numbers of dilated convolutional layers. More dilated convolutional layers indicate a larger receptive field in the discriminator. Because our focus is to learn a mapping for large shape deformation, we evaluate our model on anime to human task. Note that the anime to human task is more challenging than the human to anime task. This is because human face consists of local textures that are missing in anime face. To translate an anime face to human face, the model needs to handle both high-level attributes and local textures. As shown in Figure 15, the FID score of our model with 3 dilated convolutional layers is 24.47, which is lower than other variants with less dilated convolutional layers.

#### B. Effect of the Input-output Drawer

To further evaluate the effect of our proposed drawer structure, we perform another experiment on a dataset containing painting collections from multiple artists, including Van-Gogh,

TABLE IV: Network architecture for  $256 \times 256$  painting style dataset. Conv(d,k,s) and DeConv(d,k,s) denote the convolutional layer and transposed convolutional layer with d as dimension, k as kernel size and s as stride. IN is instance normalization. We do not use batch normalization or instance normalization in the discriminator.

| Encoder                        | Decoder                          | Discriminator              | Classifier           |
|--------------------------------|----------------------------------|----------------------------|----------------------|
| Conv(64,4,2), IN, Leaky ReLU   | DeConv(1024,4,2), IN, ReLU       | Conv(64,4,2), Leaky ReLU   |                      |
| ResNet(64,3,1)                 | DeConv(512,4,2), IN, ReLU        | Conv(128,4,2), Leaky ReLU  |                      |
| Conv(128,4,2), IN, Leaky ReLU  | DeConv(256,4,2), IN, ReLU        | Conv(64,4,2), Leaky ReLU   |                      |
| ResNet(128,3,1)                | DeConv(128,4,2), IN, ReLU        | Conv(256,4,2), Leaky ReLU  |                      |
| Conv(256,4,2), IN, Leaky ReLU  | DeConv(64,4,2), IN, ReLU         | Conv(512,4,2), Leaky ReLU  |                      |
| ResNet(128,3,1)                | DeConv( $3 \times N$ ,3,1), Tanh | Conv(1024,4,2), Leaky ReLU |                      |
| Conv(512,4,2), IN, Leaky ReLU  |                                  | FC(1024), Leaky ReLU       | FC(1024), Leaky ReLU |
| Conv(1024,4,2), IN, Leaky ReLU |                                  | FC(1)                      | FC(13), Softmax      |

TABLE V: The cost of parameters and model capacity in the process of training. The advantage of the unified methods in retrenching calculating space and time is obvious.

| Method   | # Parameters | # Models with m=10              |
|----------|--------------|---------------------------------|
| MUNIT    | 54.06M       | $C_m^2 = \frac{m(m-1)}{2} = 45$ |
| CycleGAN | 52.6M        | $C_m^2 = \frac{m(m-1)}{2} = 45$ |
| DRIT     | 123.42M      | $C_m^2 = \frac{m(m-1)}{2} = 45$ |
| FUNIT    | 46.24M       | 1                               |
| StarGAN  | 53.28M       | 1                               |
| Ours     | 69.74M       | 1                               |

Monet, Cezanne, and Gauguin. This task is more challenging than the aforementioned collection style transfer task, since we attempt to generate stylized images in high-resolution. This requires the capacity of the model to preserve the style details such as the color, stroke size, and texture pattern. To demonstrate the advantage of the drawer in preserving local textures, we train one model with the input-output drawer and another model without the input-output drawer for reference. During the training stage, each image is resized to have a smaller edge of 512 pixels while its original aspect ratio remains unchanged. Then a  $512 \times 512$  image patch is randomly extracted as the input that is used to train the model. At the testing stage, the full image with a smaller edge of 512 pixels is fed to the model for style transfer. The results are shown in Figure 17 and 18. We observe that the model with the input-output drawer yields the best synthesized style images with well-reconstructed details. While the images created by the model without the drawers contain artifacts in local textures.

### C. Effect of the Perceptual Self-Regularization

Our model proposes to use a single generator and is trained with a self-regularization term that enforces perceptual similarity between the output and the input, together with an adversarial term that enforces the output to appear like drawn from the target domain. The translated image should preserve both high-level attributes and local texture from the target domain. In other words, the output and input need to share perceptual similarities, such as the color and the pose, while allowing changing the domain-specific features. In contrast, the cycle loss seeks for pixel-level consistency. The main issue of the cycle-consistency is that it assumes the translated

images contain all the information of the input images in order to reconstruct the input images. This assumption leads to the preservation of source domain features, such as traces of objects and texture, resulting in unrealistic results. Moreover, the cycle-consistency constrains shape changes of source images (e.g. results produced by CycleGAN in Figure 7). To verify, we train three model variants: (a) ours w/o  $\mathcal{L}_D$  refer to one model trained without the adversarial loss; (b) ours w/ cycle loss stands for one model trained with the cycle loss instead of the perceptual self-Regularization loss; (c) ours w/ perceptual loss represents our full model trained with perceptual self-Regularization loss. Quantitative results of the FID scores at the training stage are reported in Figure 16. Our w/o  $\mathcal{L}_D$  produces a higher FID score indicating that this model can not learn a mapping between different domains. This proves that the adversarial loss is the driven power to enforce the output to appear like drawn from the target domain. Even though ours w/ cycle loss model achieves a lower FID score, the generated human face contains all the information of the input anime images. However, our full model achieves the best FID score and creates realistic human images.

## VI. MODEL COMPLEXITY ANALYSIS

We compared the overall model capacity with other baselines. The number of models and the number of model parameters on the dataset for different  $m$  image domains are shown in Table V. CycleGAN, DRIT, MUNIT are unsupervised methods, and they require  $C_m^2$  models to learn  $m$  image domains, but each of them contains two generators and two discriminators, while StarGAN and FUNIT only need to train one model to learn all the mappings of  $m$  domains. We also report the number of parameters on the face aging dataset, this dataset contains 10 different age periods, which means  $m = 10$ . Note that our model takes the drawers in the first and last layers in the generator, which brings additional parameters.

## VII. NETWORK STRUCTURE

The proposed model consists of one unified generator and discriminator for multi-domain image translation. Table IV shows the detailed network architectures of our model. The discriminator D is a stack of convolutional layers followed by fully connected layers, and the classifier C has a similar

architecture and shares all convolutional layers with D. The generator contains one pair of encoder and decoder. Following the Unet, we utilize the lateral connection between the encoder and decoder, which have been shown to produce high-quality results on the image translation task. The encoder is a stack of convolutional layers and the decoder is a stack of transposed convolutional layers. The Architectures for  $128 \times 128$  images are used in the comparisons with other baselines on the tasks, human2anime, photo2portrait, animal image translation, and face aging.  $256 \times 256$  images are used on the task of collection image transfer.  $256 \times 256$  images are shown in this experiment for better visual effect.

### VIII. CONCLUSION

We have presented a novel GAN model for image-to-image translation across multi-domains. Instead of taking the cycle-consistency constraint, the proposed model exploits a perceptual self-regularization constraint. Specifically, we propose the input-output drawers to preserve domain-specific features. By exploring multi-scale information, our model effectively handles large shape deformation. Extensive experiments demonstrate that the proposed method promotes the unified network to go beyond simple texture transfer and improves shape deformation across multiple domains. The system is capable of performing challenging translations such as from human to anime and from cat to dog. The generated images possess a higher quality compared with the baselines.

### REFERENCES

- [1] Z. Chen, S. Nie, T. Wu, and C. G. Healey, “High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks,” *arXiv preprint arXiv:1801.07632*, 2018. [1](#)
- [2] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711. [1, 2, 5](#)
- [3] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666. [1, 7](#)
- [4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks.” *arxiv*, 2016. [1, 2, 6, 7, 8](#)
- [5] W. Shen and R. Liu, “Learning residual images for face attribute manipulation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4030–4038. [1](#)
- [6] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, “Few-shot unsupervised image-to-image translation,” *arXiv preprint arXiv:1905.01723*, 2019. [1, 3, 6](#)
- [7] 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. [1, 2, 6](#)
- [8] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C. J. Kuo, “Esther: Extremely simple image translation through self-regularization.” in *BMVC*, 2018, p. 110. [1](#)
- [9] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, “Transgaga: Geometry-aware unsupervised image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8012–8021. [1, 3](#)
- [10] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv preprint arXiv:1706.05098*, 2017. [1](#)
- [11] G. Zhou, Y. Zhao, F. Guo, and W. Xu, “A smart high accuracy silicon piezoresistive pressure sensor temperature compensation system,” *Sensors*, vol. 14, no. 7, pp. 12 174–12 190, 2014. [2](#)
- [12] W. Xu, Y. Wu, W. Ma, and G. Wang, “Adaptively denoising proposal collection for weakly supervised object localization,” *Neural Processing Letters*, vol. 51, no. 1, pp. 993–1006, 2020. [2](#)
- [13] F. Cen and G. Wang, “Boosting occluded image classification via subspace decomposition-based estimation of deep features,” *IEEE transactions on cybernetics*, vol. 50, no. 7, pp. 3409–3422, 2019. [2](#)
- [14] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, “Sosd-net: Joint semantic object segmentation and depth estimation from monocular images,” *Neurocomputing*, vol. 440, pp. 251–263, 2021. [2](#)
- [15] W. Xu, K. Shawn, and G. Wang, “Stacked wasserstein autoencoder,” *Neurocomputing*, vol. 363, pp. 195 – 204, 2019. [2](#)
- [16] W. Xu, S. Keshmiri, and G. Wang, “Adversarially approximated autoencoder for image generation and manipulation,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2387–2396, 2019. [2](#)
- [17] K. Li, M. I. Fathan, K. Patel, T. Zhang, C. Zhong, A. Bansal, A. Rastogi, J. S. Wang, and G. Wang, “Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations,” *arXiv preprint arXiv:2104.10824*, 2021. [2](#)
- [18] X. Dong, C. Long, W. Xu, and C. Xiao, “Dual graph convolutional networks with transformer and curriculum learning for image captioning,” *arXiv preprint arXiv:2108.02366*, 2021. [2](#)
- [19] W. Xu, G. Wang, A. Sullivan, and Z. Zhang, “Towards learning affine-invariant representations via data-efficient cnns,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. [2](#)
- [20] R. Huang, W. Xu, T.-Y. Lee, A. Cherian, Y. Wang, and T. Marks, “Fx-gan: Self-supervised gan learning via feature exchange,” in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 3194–3202. [2](#)
- [21] W. Xu, S. Keshmiri, and G. Wang, “Toward learning a unified many-to-many mapping for diverse image translation,” *Pattern Recognition*, vol. 93, pp. 570 – 580, 2019. [2](#)
- [22] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, “Adversarially learned inference,” *arXiv preprint arXiv:1606.00704*, 2016. [2](#)
- [23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680. [2](#)
- [24] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” *arXiv preprint arXiv:1711.01558*, 2017. [2](#)
- [25] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, “Toward multimodal image-to-image translation,” in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476. [2](#)
- [26] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *In the International Conference on Machine Learning-Volume 48*, 2016, pp. 1558–1566. [2](#)
- [27] W. Xu, C. Long, R. Wang, and G. Wang, “Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6383–6392. [2](#)
- [28] 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. [2](#)
- [29] L. Tran, X. Yin, and X. Liu, “Disentangled representation learning gan for pose-invariant face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. [2](#)
- [30] Z. Zhang, Y. Song, and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. [2, 6, 9](#)
- [31] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International Conference on Machine Learning*, 2017, pp. 214–223. [2, 4](#)
- [32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5769–5779. [2, 4](#)
- [33] 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. [2](#)
- [34] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Advances in Neural Information Processing Systems*, 2017, pp. 700–708. [2](#)
- [35] Z. Yi, H. Zhang, P. T. Gong *et al.*, “Dualgan: Unsupervised dual learning for image-to-image translation,” *arXiv preprint arXiv:1704.02510*, 2017. [2](#)
- [36] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” *arXiv preprint arXiv:1611.02200*, 2016. [2](#)

- [37] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” *arXiv preprint arXiv:1612.05424*, 2016. [2](#)
- [38] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189. [2, 6](#)
- [39] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Diverse image-to-image translation via disentangled representations,” *arXiv preprint arXiv:1808.00948*, 2018. [2, 6](#)
- [40] A. Gokaslan, V. Ramanujan, D. Ritchie, K. In Kim, and J. Tompkin, “Improving shape deformation in unsupervised image-to-image translation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 649–665. [3](#)
- [41] Y. Shi, D. Deb, and A. K. Jain, “Warpgan: Automatic caricature generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 762–10 771. [3](#)
- [42] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in *Advances in Neural Information Processing Systems*, 2017, pp. 406–416. [3](#)
- [43] K. Cao, J. Liao, and L. Yuan, “Carigans: unpaired photo-to-caricature translation,” *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–14, 2018. [3](#)
- [44] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [3](#)
- [45] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797. [3, 6](#)
- [46] A. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, “A unified feature disentangler for multi-domain image translation and manipulation,” *arXiv preprint arXiv:1809.01361*, 2018. [3](#)
- [47] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3673–3682. [3](#)
- [48] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a single image,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833. [3](#)
- [49] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, “Combogan: Unrestrained scalability for image domain translation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783–790. [3](#)
- [50] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li, “Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning,” in *Asian Conference on Computer Vision*. Springer, 2018, pp. 341–356. [3](#)
- [51] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. [5](#)
- [52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242. [7](#)
- [53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6629–6640. [7](#)
- [54] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. [7](#)
- [55] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008. [9](#)