

Unsupervised Image-to-Image Translation via Pre-trained StyleGAN2 Network

Jialu Huang, Jing Liao, Sam Kwong, Fellow, IEEE

Abstract—Image-to-Image (I2I) translation is a heated topic in academia, and it also has been applied in real-world industry for tasks like image synthesis, super-resolution, and colorization. However, traditional I2I translation methods train data in two or more domains together. This requires lots of computation resources. Moreover, the results are of lower quality, and they contain many more artifacts. The training process could be unstable when the data in different domains are not balanced, and modal collapse is more likely to happen. We proposed a new I2I translation method that generates a new model in the target domain via a series of model transformations on a pre-trained StyleGAN2 model in the source domain. After that, we proposed an inversion method to achieve the conversion between an image and its latent vector. By feeding the latent vector into the generated model, we can perform I2I translation between the source domain and target domain. Both qualitative and quantitative evaluations were conducted to prove that the proposed method can achieve outstanding performance in terms of image quality, diversity and semantic similarity to the input and reference images compared to state-of-the-art works.

I. INTRODUCTION

Image-to-Image (I2I) translation has been a hot topic recently. With rapid development of deep learning, I2I translation has been greatly improved in the past few years. It also provides solutions to real-world applications like de-noise, de-rain, colorization, super-resolution and image editing.

Recent work has made huge progress in this field due to the development of the generative adversarial network (GAN) [10]. After the pioneer work of Pix2Pix [20], BicycleGAN [59] was developed to support I2I translation with multiple styles. Later, researchers proposed UNIT [30], CycleGAN [58] MUNIT [18], and DRIT/DRIT++ [25], [27] for unsupervised I2I, where MUNIT and DRIT/DRIT++ can perform multi-modal translation. Nevertheless, there are still challenges in the I2I translation task. First, paired data are limited for supervised training methods while quality of the results of unsupervised methods has been relatively low. In addition, most I2I networks must train at least two generators for source and target domains

together. This may lead to some problems: training resources including time and memory, would be high; uneven data in either source or target domains could cause instability in training process. Moreover, controllable generation and multi-domain translation [33], [9], [56], [54] attract lots attention in both academia and industry, but most of the proposed methods require additional training data or complex structural design, which makes the original translation much more resource-consuming.

We propose a new solution to the I2I translation task via a pre-trained network. We want to find models that can generate paired images based on the same latent representation. Specifically, for model G_X , which can generate image $X \in \mathcal{X}$ based on a given latent code z , we try to find the model G_Y that can generate image $Y \in \mathcal{Y}$, where \mathcal{X} and \mathcal{Y} are two image domains, and $G_Y(z)$ is semantically similar to $G_X(z)$. In our initial experiments, we found that with conservative learning methods and proper layer manipulation, fine-tuning model G_X on the \mathcal{Y} domain data can yield model G_Y , which satisfies the above assumption. We refer to this process as model transformation T in the following sections, where $G_Y = T_{X \rightarrow Y}(G_X)$. Previous inversion work including IdInvert [57] and Image2StyleGAN [1] proposed methods to obtain latent representation z of an image I given its pre-trained model G . With these method, the I2I problem can be solved by (1) inverting the given image X to obtain the latent representation $z = \text{Inv}(X, G_X)$ where Inv is an inversion method; (2) performing model transformation on the pre-trained model G_X to generate the target domain model G_Y , and (3) generating the paired image in \mathcal{Y} domain $Y = G_Y(z)$.

Different from traditional methods that train from the scratch, we choose pre-trained models like StyleGAN2 [24] to ensure image quality and fidelity of the output. In addition, our method requires training only in the model translation. This process contains conservative model fine-tuning and layer manipulation, which requires much less training resources in terms of both time and memory. As the model transformation focus only on the given domain, our method is unsupervised. It does not require any paired data across different domains. To achieve unsupervised I2I translation solutions, previous methods including CycleGAN [58], MUNIT [18], and DRIT/DRIT++ [25], [27] all used cycle loss. This means bidirectional translations $X \rightarrow Y$ and $Y \rightarrow X$ should be trained together. Model collapse can happen at one side if the training data are unbalanced. In other words, those methods cannot achieve proper results when data in one domain are limited. Our method that trains models in a single domain at a time can preserve the generation quality as much as possible, and it is not affect by unbalanced data in

This work was supported in part by the National Natural Science Foundation of China Grant 61672443, in part by Hong Kong GRF-RGC General Research Fund under Grant 9042322 (CityU 11200116), Grant 9042489 (CityU 11206317), and Grant 9042816 (CityU 11209819)

This work was supported in part by the Hong Kong Research Grants Council (RGC) Early Career Scheme under Grant 9048148 (CityU 21209119), and in part by the CityU of Hong Kong under APRC Grant 9610488.

Jialu Huang, Jingliao and Sam Kwong are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: jialu8-c@my.cityu.edu.hk, jingliao@cityu.edu.hk, cssamk@cityu.edu.hk) Sam Kwong is also with City University of Hong Kong Shenzhen Research Institute.

Implementation can be found at: https://github.com/HideUnderBush/UI2I_via_StyleGAN2

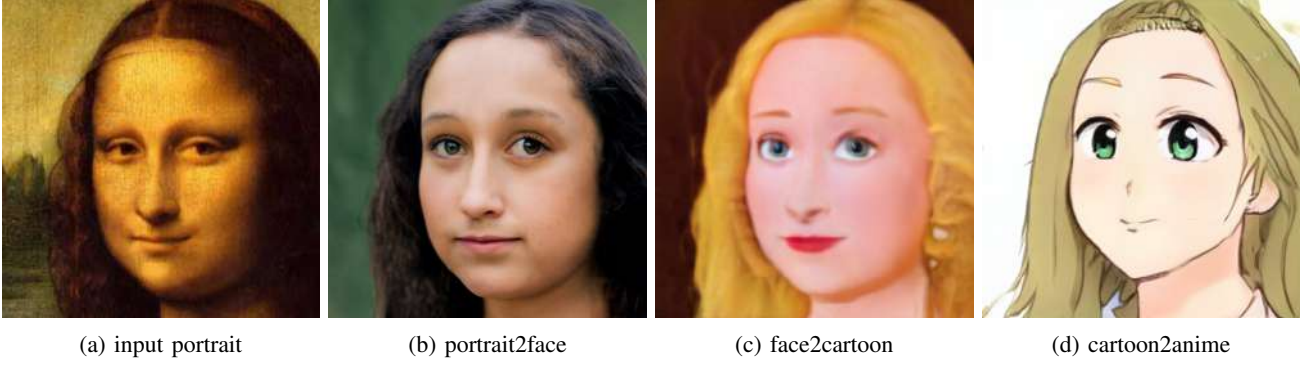


Fig. 1: An example of the proposed multi-domain I2I translation results. We train transformed models on four different face domains and then can perform I2I translation between any two domains, such as portrait2face (b), face2cartoon (c), and cartoon2anime (d). Compared to existing I2I translation methods, which need at least $C_4^2 = 6$ models to perform arbitrary I2I translation among the above 4 domains, our method only needs 4 models since any two of those models can be used for translation.

two domains.

Besides, the pre-trained model StyleGAN2 [24] in nature supports the image generation of multiple styles. By injecting various latent style code at different layers of StyleGAN2 [24], the final output can have different styles. This enables our method to support multi-modal I2I translation without any additional training or modification of the network structure. We also found that the proposed method can be extended to multi-domain I2I translation. Assuming that we have a pre-trained model trained on domain \mathcal{D}_0 , and data from various domains \mathcal{D}_i where $i \neq 0$, with model transformation $T_{\mathcal{D}_0 \rightarrow \mathcal{D}_i}$, we can obtain models for different domains $G_{\mathcal{D}_i} = T_{\mathcal{D}_0 \rightarrow \mathcal{D}_i}(G_{\mathcal{D}_0}, \mathcal{D}_i)$. According to that images generated by $G_{\mathcal{D}_0}$ would be semantically similar to images generated by $G_{\mathcal{D}_i}$, combined with the transitivity of semantic similarity, such similarities will be preserved between any pairs of two generated models $G_{\mathcal{D}_m}$ and $G_{\mathcal{D}_n}$ where $m \neq n$. Therefore, as shown in Fig.1, we can perform I2I translation between any two domains \mathcal{D}_m and \mathcal{D}_n without additional training.

Our major contributions are:

- We define the distance between two models to measure the semantic similarity between two images generated by two models, based on the same input latent vector. The model distance help us analyze the properties of I2I translation between the two domains.
- We propose an unsupervised I2I translation method via a pre-trained StyleGAN2 model, which can generate images in the target domain with high quality. Our proposed method can support multi-modal and multi-domain I2I translation. Compared to traditional I2I translation methods, it drastically improved the results and requires much less training resource.
- We also proposed an inversion method, which can perform conversion between an image and a latent vector based on a given StyleGAN2 model. Compared to previous algorithms, our inversion method is based on an embedded GAN space, which provides a boundary constraint for searching the latent code of the input image.
- We perform both qualitative and quantitative evaluations

on various challenging dataset, and we compare the proposed method with state-of-the-art works. This shows that the proposed method achieves outstanding output and better results in all the experiment.

II. RELATED WORK

A. Image-to-Image Translation

I2I translation aims to transfer an input image from the source domain to target domains. This can be applied to many applications including colorization, super-resolution, and face-editing. With the development of GAN [10], Isola et al. proposed the first supervised I2I translation network pix2pix [20], which they then extended to pix2pixHD [49] to support translation between images with higher resolution. As the paired training data are limited, an unsupervised method CycleGAN [58] was then proposed by Zhu et al. They used cycle loss to build up the constraints between the source and target images, which were then applied in other unsupervised I2I translation methods [26], [38], [40], [41].

Moreover, several multi-modal methods including BicycleGAN [59], MUNIT[18], and DRIT/DRIT++[25], [27] made it possible for an input image to have multiple counterparts with different styles. In those methods, a style latent code could either be randomly sampled or extracted from a given reference image. By injecting the combined content and style code into the generator, they could generate output with specified appearance. However, this type of style control is not accurate, and it can be applied only to images in a global way. Later works [47], [33], [15] used an example to semantically control the output images, while other works[56], [60], [16] tried to locally control attributes of the output such as color and pose, based on users' input. Most recently, motivated by the high quality of StyleGAN [22], some works begin to incorporate it into the I2I framework. Elad et al. [39] proposed to learn a mapping from images to StyleGAN latent codes via an end-to-end training method, which can achieve high-quality I2I translation between two domains. In addition, some researchers used Toonify [21] which is based on the distillation of a blended

StyleGAN [22] network into a pix2pixHD [49] I2I translation network. The translation part is similar to the pix2pixHD.

Current I2I translation methods require online training on at least two domains, and most of them have domain-distinguished generators and discriminator in an overall network to ensure the output image fits its own domain. However, this has a high demand on training resources in terms of both time and memory, and such demand would increase as the domains increase. Modal collapse also can happen when training data are not balanced in any domains. Therefore, we proposed the method that focus on training images in one domain at a time, and use model transformation to link models trained on different domains to achieve multi-modal and multi-domain I2I translation.

B. Inverse Mapping from Image to Latent Code

With the rapid development of GAN [10], more and more models can generate high quality images from a latent code. These include StyleGAN/StyleGAN2 [22], [24], BigGAN [6] and PCGAN [28]. Researchers further studied the GAN space and tried to find a way to map between the input latent code and the output image. Perarnau et al. proposed IcGAN [36], which trained an encoder in the cGAN framework to compress a real image X into a latent representation z . Similarly, AEGAN [31] was then proposed by Luo et al. It used an autoencoder to train an inverse model by minimizing the differences between noise-generated samples and related reconstructed images. Ma et al. [32] theoretically proved that a convolutional GAN is invertible that same inversion could be achieved even if only a subset of samples is observed.

Then researchers found that it was also possible to use optimization methods [7], [29], [1], [2] to search the latent representation iteratively, according to their experiments, the optimization-based methods could provide more accurate noise as well as better reconstructed images. Gu et al. [11] believed that one image could be determined by multiple priors. They proposed the Multi-code GAN prior for image inversion. They proposed that multiple latent representation could reconstruct the image with more details and higher precision. To speed up the inversion process, Zhu et al. [57] proposed the in-domain GAN inversion method. Combining the optimisation with a trained encoder, the inversion encoder can output a rough latent representation which could then be taken as an initial vector in the optimization. This could greatly diminish the optimisation time without losing the precision of the inversion.

Most inversion methods use the difference between the input image and the reconstructed images as major loss functions. Such difference includes general L1/L2 loss and perceptual loss. In addition to those loss functions, there are no extra constraints applied to the latent representation or latent space. Our proposed method adds an explicit boundary to the latent space so that the output latent code can always be a point in the pre-trained GAN space.

C. Transfer Learning

Transfer learning aims to solve problems using data that are not directly related to the task considered. This can be

classified as two major types: (1) similar data for different tasks or (2) different dataset for the same task. The second type tries to find a solution for the task with limited data. Our method can also be classified into the second type as it aims to perform I2I translation (same task) among images in multiple domains (different dataset). Transfer learning contains many typical learning methods such as conservative learning, multi-task training, domain adversarial learning and teacher-student systems. Domain adversarial learning and multi-task training focus on solving the first type of problem, while conservative learning is proposed to solve the limited-data learning problems. Conservative learning work including [4], [19], [53], [12], [42], [46], [48], [5] conducts fine-tuning with partial layers, while other works [14], [50], [8], [45] add a regularizator or loss functions to constraint the distance between the current output and the pre-trained output. Since the teacher-student structure is an essential subject in the transfer learning system, we do not discuss it here because it is out of the scope of this work.

III. METHOD

To describe the semantic similarities between images generated by different models, we define the model distance in a formal way as a metric to measure such similarities. Then we propose a series of transformations T to reduce the distance between models trained on different domains. We also propose an improved inversion method to obtain the latent code from images; This enables I2I translation between semantically similar models. As we use StyleGAN2 [24] as the base model, our method not only ensures high-quality results; it also supports multi-modal translation in its essence. The overall flow of the proposed method is shown in Fig. 2

A. Model distance and transformation

The I2I translation task is based on deep convolution neural network (DCNN). With an input image, the network can generate a related output in the target domain. Deep neural models are determined by model structures and parameters, since there are no unique criteria for structural design, and the amount of parameters in one model could be huge. Therefore, it is difficult to evaluate the relationship between two models directly. Although this seems like a black box problem, we can define the distance from another point of view. That is, we can use the distance between two outputs from the models given the same input as a representation of the model distance. Two models, G_{D_i} and G_{D_j} , can generate images from latent code z in two domains $D_i \in \mathcal{D}_i$ and $D_j \in \mathcal{D}_j$. The model distance between G_{D_i} and G_{D_j} can be defined as:

$$d(G_{D_i}, G_{D_j}) = \int dist(G_{D_i}(z), G_{D_j}(z)) dz, z \in \mathcal{N}(0, \mathcal{I}) \quad (1)$$

where $dist$ is the semantic similarity between two images. Specifically, we use LPIPS [55] distance in the proposed method.

Based on the previous discussion, the I2I translation can achieve reasonable results only if there is a small distance between the models in the source and target domains. Otherwise, the latent code obtained from the source image

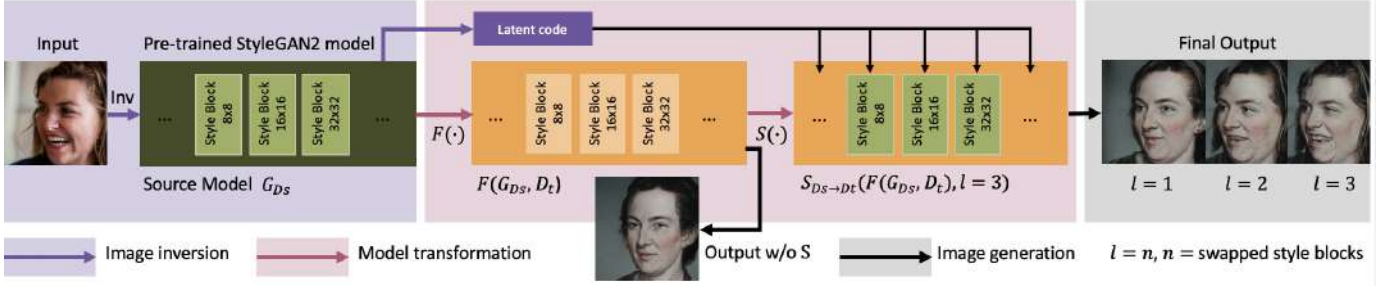


Fig. 2: The main structure of the proposed work. The left-most block aims to obtain a latent code from a given image via the inversion method. And the middle part is the model transformation, where F stands for fine-tuning process and S represents high-level layer-swap. Although results generated by the model after F process belongs to the target domain, later S transformation can achieve higher semantic similarities to the input images and such similarities increase with the increase of the number of swapped layers.

$z = \text{Inv}(D_i, G_{D_i})$ cannot generate a related output in the target domain. That is, $D_i \approx D_j$, where \approx indicates high semantic similarity between two images. In our experiments, we found fine-tuning can reduce the model distance. We use the StyleGAN2 [24] model trained on source domain \mathcal{D}_s as the base model, and we fine-tune it with data in target domain \mathcal{D}_t with the GAN loss:

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{D_t \sim \mathcal{D}_t} [\log D(D_t)] + \mathbb{E}_{z \sim \mathcal{N}} [\log(1 - D(G_{D_t}(z)))] \quad (2)$$

As we know, the StyleGAN2 network first maps the latent code z to an embedded space \mathcal{W} via an 8-layer MLP $w = f(z)$, $w \in \mathcal{W}$, where $f: \mathcal{N} \rightarrow \mathcal{W}$ is the feature mapping function. Features in \mathcal{W} space are then be sent to the generator to output the images. Inspired by FreezeD [35], we freeze the fully-connected (FC) layers in the fine-tuning process, in which the fine-tuned model can achieve smaller model distance from the base model. This is because, with a frozen FC layers, the fine-tuned model would have the same embedded space as the base model. This ensures that the input latent code is always initialized at the same point in the embedded space as that of the base model.

After that, given a latent code z , the fine-tuned model $G_{D_t} = F(G_{D_s}, D_t)$, where F indicate the fine-tuning process, can generate images $D_t = G_{D_t}(z)$, $D_t \in \mathcal{D}_t$ and $D_t \sim G_{D_s}(z)$. However, during fine-tuning, the semantic similarity decreases due to domain difference. The new model G_{D_t} must learn specific features in the new domain to reduce the GAN loss. This is similar to catastrophic forgetting [34] in transfer learning where the new model G_{D_t} would fail to generate images in the source domain after fine-tuning, and it forgets partial features previously learned from the source domain. Inspired by Justin et al. [37], they applied layer-swapping on image analogy task. And we found it is possible to swap higher convolutional layers of G_{D_s} to G_{D_t} . Then the target model can generate images and preserve more features from the source domain. This further diminishes the model distance between G_{D_s} and G_{D_t} . We define this process as $G_{D_t} = S_{D_s \rightarrow D_t}(G'_{D_t}, l)$, $l \in N$, where G'_{D_t} is the model after fine-tuning and l indicates the number of swapped layers. According to experiments, we found layer-swap cannot be used between two models with large model

distance (e.g. models without fine-tuning). This is because features in such models have large differences, they might even be two feature spaces. Therefore features outputted by swapped layers cannot be used by the following layers. However, we can control the final output of G_{D_t} by modifying the number of swapped layers l . With a larger l , the generated images are more semantically similar to the images in the source domain. With a smaller l , the generated images maintain more features in the target domain. Hence, compared to freezing the generator directly in the fine-tuning stage, layer-swap provides more flexibility in controlling the final output. Finally, We define the transformation for reducing the model distance between G_{D_s} and G_{D_t} as

$$G_{D_t} = T_{D_s \rightarrow D_t}(G_{D_s}) = S_{D_s \rightarrow D_t}(F(G_{D_s}, D_t), l) \quad (3)$$

With model transformation, models with short distances can be created in various domains.

B. Image-to-code inversion

So far, we have two models in different domains, where $G_{D_s}(z) \sim G_{D_t}(z)$. Specifically, StyleGAN has an MLP mapping the latent code z to an embedded space, that is $G_{D_s}(w) \sim G_{D_t}(w)$ where $w = f(z)$. However, in the I2I translation task, the input is an image in the source domain I_s , and the required output I_t should be corresponding images in the target domain, namely $I_s \sim I_t$. To achieve the I2I translation, we need to bridge the gap between the input image and the embedded latent vector w . Therefore, we propose an inversion method to find the latent code related to the input image $w = \text{Inv}(D_s, G_{D_s})$. The latent code will then be sent to the transformed model to generate output in the target domain $D_t = G_{D_t}(w)$. This process is defined as:

$$D_t = T_{D_s \rightarrow D_t}(G_{D_s})(\text{Inv}(D_s, G_{D_s})) \quad (4)$$

where Inv is the inversion algorithm taking an image and a related model trained in the same domain as the input.

An algorithm to project the image to latent space has been proposed in StyleGAN2, which obtains the latent code directly by optimizing loss functions:

$$w^* = \arg \min_{w \in \mathbb{R}^d} \|I - G(w)\|_1 + \|\Phi(I) - \Phi(G(w))\|_2 \quad (5)$$

where I is the input image, G represents the pre-trained network on the related domain and Φ indicates the pre-trained VGG [44] network, which is used to extract image features.

This method tries to find a specific w in the embedded space \mathcal{W} , but it cannot guarantee the found w is within the embedded space for the following reasons: (1) there is no explicit description about the \mathcal{W} space, as it is formed by fully connected layers. (2) there is no constraint directly applied to w , as the loss functions are based on images. To tackle this problem, we propose an improved method that optimises the w within an explicit defined embedded space. Shen et al. [43] proposed SeFa to find a semantic space for a pre-trained GAN model, where major directions in the space can control specific semantic attributes in generated images. In SeFa [43], the difference between two latent codes can be described as:

$$\Delta w = (Az + b) - (A(z + \alpha n) + b) = \alpha An \quad (6)$$

where $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^d$ are weight and bias in the fully-connect layers and α is a hyperparameter. SeFa states that A is a semantic selector of the model, and finding a specific semantic attribute is to optimize $n^* = \arg \max \|An\|_2^2$, and n^* is the direction for that attribute. This can then be extended to a problem finding the top k major semantic directions, so the optimization would be rewritten as:

$$N^* = \arg \max_{N \in \mathbb{R}^{d \times k}} \sum_{i=1}^k \|An_i\|_2^2 - \sum_{i=1}^k \lambda_i (n_i^T n_i - 1) \quad (7)$$

where $N = [n_1, n_2, \dots, n_k]$ and $\sum_{i=1}^k \lambda_i (n_i^T n_i - 1)$ is the Lagrange Multiplier. After taking the partial derivative on n_i , we have:

$$A^T A n_i - \lambda_i n_i = 0 \quad (8)$$

According to SeFa [43], the eigen vectors of $A^T A$ are semantic directions. Therefore we can use this metric as a formed embedded space $V = A^T A$. We rewrite the optimization in Eq.5 as follows:

$$v^* = \arg \min_{v \in \mathbb{R}^d} \|I - G(v^T V)\|_1 + \|\Phi(I) - \Phi(G(v^T V))\|_2 \quad (9)$$

where v is a projection point in space V , and the latent vector $w = v^T V$. Specifically, we first compute V_{D_s} based on a pre-trained model in the source domain G_{D_s} , and we obtain the vector v_{D_s} by optimizing Eq.9. Then we have the inverted latent code for a given input image D_s that $w = v_{D_s}^T V_{D_s} = \text{Inv}(D_s, G_{D_s})$. Compared to direct projection, our proposed method is more like selecting one point in a pre-defined embedded space, based on a given network. The constraints of latent space can improve the previous optimization.

C. Multi-modal and Multi-domain I2I translation

Similar to MUNIT [18] and DRIT/DRIT++ [25], [27], the proposed method can support multi-modal I2I translation as well. First, we revisit the StyleGAN2 [24] structure that latent code will be mapped to layer-wise style code and then sent to the major generator via AdaIn [17] layers. StyleGAN [22] states that style code injected at the higher layers can change a

major structure of the output, such as identity, while the style code applied at lower layers can modify only minor features such as color, light conditions and other micro structures.

In our method, we separate the w into two parts, namely content code w_c and appearance code w_a , and $w = w_c, w_a$. To ensure high semantic similarities between the source and the target image, w_c is obtained via the inversion method from the source image $w_c = \text{Inv}(D_s, G_{D_s})$, while the appearance code w_a can be sampled from normal distribution $z_a \in \mathcal{N}(0, \mathcal{I})$, $w_a = f(z_a)$. To perform multi-modal I2I translation, w_c is inserted into most generator layers, while w_a is used at only the last few layers, the details of implementation are in Section IV. Therefore, multiple output in the target domain can be generated based on the input:

$$D_t^{(i)} = G_{D_t}(\text{Inv}(D_s, G_{D_s}), w_a^{(i)}) \quad i = \{1, 2, \dots, n\} \quad (10)$$

Moreover, the appearance code can also be obtained from a reference image from the target domain $w_a = \text{Inv}(D_r, G_{D_t})$, $D_r \in \mathcal{D}_t$. Similar to layer-swap mentioned above, we can choose different numbers of layers to inject the appearance code to control the similarities between the generated output and the reference.

In addition, the proposed method can support multi-domain translation. In traditional methods, multi-domain translation is limited because it requires that multiple generators are trained together. This places a high demand on both resources and data quality. All the training in our method focuses one domain at a time, and multi-domain translation can be performed off-line with pre-trained models. Assuming that we have an initial source domain data \mathcal{D}_s and a bunch of target domain data $\{\mathcal{D}_{t1}, \mathcal{D}_{t2}, \dots, \mathcal{D}_{tn}\}$, according to the proposed transformation method $T_{D_s \rightarrow D_{ti}}$, we can obtain models $G_{D_{ti}}$ in such domains. Let $\delta = d(G_{D_s}, G_{D_{ti}})$, where δ is a small value and the semantic similarity has a certain degree of transitivity. For instance, if A is semantically similar to both B and C , it is impossible that B is totally irrelevant to C . Therefore, we have $d(G_{D_{ti}}, G_{D_{tj}}) \simeq \delta$. In that way, models in different domains D_{ti} , and D_{tj} satisfy that $G_{D_{ti}}(z) \sim G_{D_{tj}}(z)$ and we can perform I2I translation between any pair of domains in the given target domains.

In the proposed method, the training process is required only in the model transformation, and it focuses on a single domain. This greatly reduces computational cost and the amount of overall memory used. Our experiments show that even with a small dataset, and conservative tuning could still achieve satisfactory results and the fine-tuned model could generate images with high quality and fidelity.

IV. EXPERIMENT

To evaluate the proposed I2I translation method, we conducted experiments on various datasets for the I2I translation task. We also compared the proposed methods with several state-of-the-art work, including CycleGAN [58], MUNIT [18] and DRIT++ [25]. Both the qualitative and quantitative results show that the proposed method can generate images in the target domains with much higher quality. Moreover, the proposed method also preserves the similarity between the input and

output images to a very large extent. Besides, experiments show that the proposed method can perform the I2I translation task with multi-modal results in multiple domains. The evaluation of the proposed inversion algorithm and the ablation studies are discussed in detail in the following section.

A. Baseline and Dataset

Several state-of-the-art I2I translation work were chosen as baselines for evaluation. CycleGAN is the first unsupervised I2I translation network that proposed cycle consistency loss to ensure identity invariance between the input and output images. Although it does not require any paired data, it could support only single modal translation. MUNIT [18] and DRIT++ [27] are proposed later to support multi-modal translation with better quality. Both methods assumed that images could be encoded into content and appearance latent space and that recombining the content code with various style codes in the target domain could achieve I2I translation.

We evaluated the proposed methods in different I2I translation scenarios including face2portrait, face2cartoon, face2anime, cat2dog and cat2wild. We used FFHQ [23] for all the face cases and AFHQ for cats, dogs and wild animals. As for the portrait, anime and cartoon, we used dataset [52], Danbooru2018 [3] and [21], respectively. Among those tasks, the face2anime I2I translation was more relatively challenging as the domain distance between the human face and anime data is quite large.

B. Implementation Details

For all the A2B cases, we froze the FC part of the A model, and we fine-tuned model A with data in the B domain. The fine-tuning contained around 12 000 (low resolution cases) - 20 000 (high resolution cases) iterations. We kept the training strategy and loss functions the same as those in the original StyleGAN2. Specifically, we used 1024x1024 images to fine-tune the pre-trained face model on the FFHQ dataset in face2portrait and face2cartoon. The face2anime model and the other animal related cases were fine-tuned with 256x256 images. In the layer-swap process, we found there are several convolution layers (8x8 - 32x32) in the styleGAN2 network to control the major content features of images in a domain. All of them have 512 channels. As we mentioned before, $G_{D_t} = S_{D_s \rightarrow D_t}(G'_{D_t}, l)$. We used $l = 1, 2, 3$ to indicate the layers with size 8x8, 16x16 and 32x32, respectively. Moreover, Adam optimizer was adopted in the proposed inversion method and 1000 iterations were required for all the inversion tasks. The fine-tuning process was conducted on a 2080Ti GPU with 11G memory. The fine-tuning took around 2 days. In addition, it would take around 0.3 ms for layer-swap and model-apply to generate an images in the target domain, and another 0.8 - 1 s for the inversion process.

C. Qualitative Results

The proposed method was compared with several state-of-the-art work, and the qualitative results are shown in Fig. 3. It shows that the proposed method generated images with much better quality, and our results were of higher semantic similarity

to the input images. Although the CycleGAN could maintain relatively fine quality, it could generate only a unique image based on an input. Both MUNIT [18] and DRIT++ [27] are unsupervised multi-modal methods. However, the generated was not very close to the input semantically, and we found artifacts in the results. Our method can also perform multi-modal and multi-domain I2I translation, which is shown in Fig. 4. A latent vector can be obtained via the proposed inversion method from the image in the source domain. In this case it is the human face domain. We used this latent vector as the content code and randomly sampled five style codes to generate images in anime domain (the first row), portrait domain (the second row) and cartoon domain (the third row) with different styles. Because those models can share the same content code, the translation between every two of those domains can be achieved as well. Fig. 4 shows that all the images are of high quality and semantic similarity to the source image. In addition, Fig. 5 shows that the proposed method can also generate images based on a reference image provided by a user. In that case, the latent code of the source image is used as the content code, and the inverse latent code from the reference is then adopted as the style code for generation. The generated images then inherit the style from the reference image, and they also maintain the semantic structure of the source image.

D. Ablation study and Analysis

Ablation Study of Fine-tuning First, we found that model fine-tuning gave the fine-tuned model a GAN space that was similar to the base model. Thus, we used the semantic space introduced by [43] and found the directions controlling ages in the base model (styleGAN2 model trained on FFHQ [23]). Fig. 6 shows that, the fine-tuned model presents similar changes in terms of age at those directions. Moreover, we conducted an ablation study of the proposed freeze-FC method. (We refer to the fine-tuning process with freeze-FC as FC, fine-tune as FT and LS as layer-swap.) In Fig. 7, generated images with FC could have more similarities to the source domain. Freezing the FC layers preserves the mapping relation from a latent code to an input feature vector. Therefore, some semantic information like age and facial expression can be preserved. For example, in the first row of Fig. 7, the generated images from FC-FT are more like a child than the results from FT. Similarly, results from FC-FT/FC-FT-LS match the facial expression of the source domain with an open mouth while the results from FT can maintain only the rough pose.

Ablation Study of Layer-swap Then we analyzed the results with different levels of layer-swap. This is shown in Fig. 8, where Direct-LS represents layer-swap without model fine-tuning, while $LS = n$ presents the number of swapped layers. $LS = 0$ indicates model transformation without layer-swap. Fig. 8 shows that all the images generated with Direct-LS are corrupted. That is because that feature cannot be shared between two models with relatively large model distance. As mentioned before, fine-tuning could reduce the model distance, and those high-level features can then be re-used by similar models. The results also indicate that with more layer-swapping, the generated images could be more like the source domain

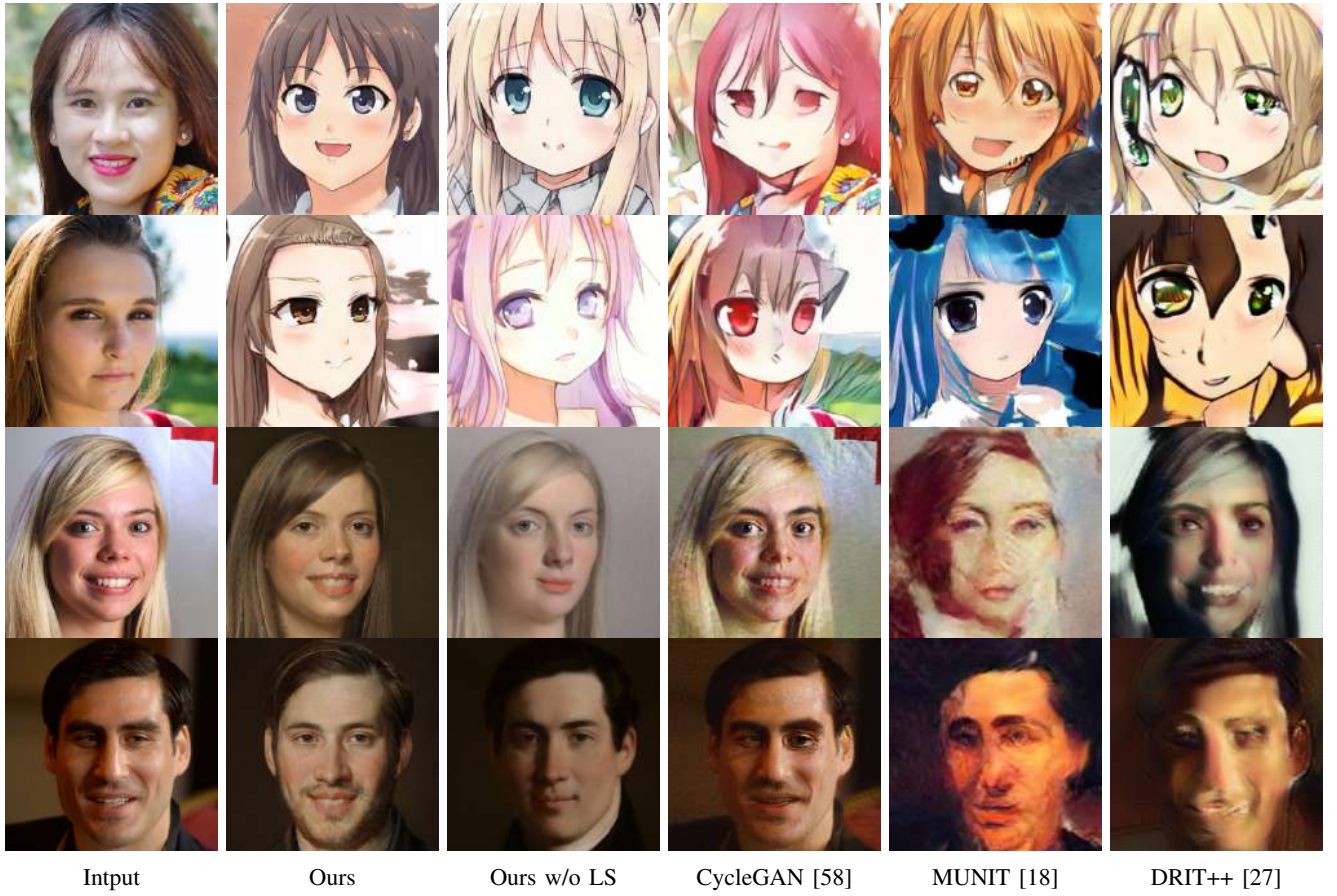


Fig. 3: Comparison with state-of-the-art work. The proposed method generates images with much higher quality, and it also preserves the semantic characteristics of the input image. It can be observed that the output with LS (the second row) is more semantically similar to the input than the output without LS (the third row)

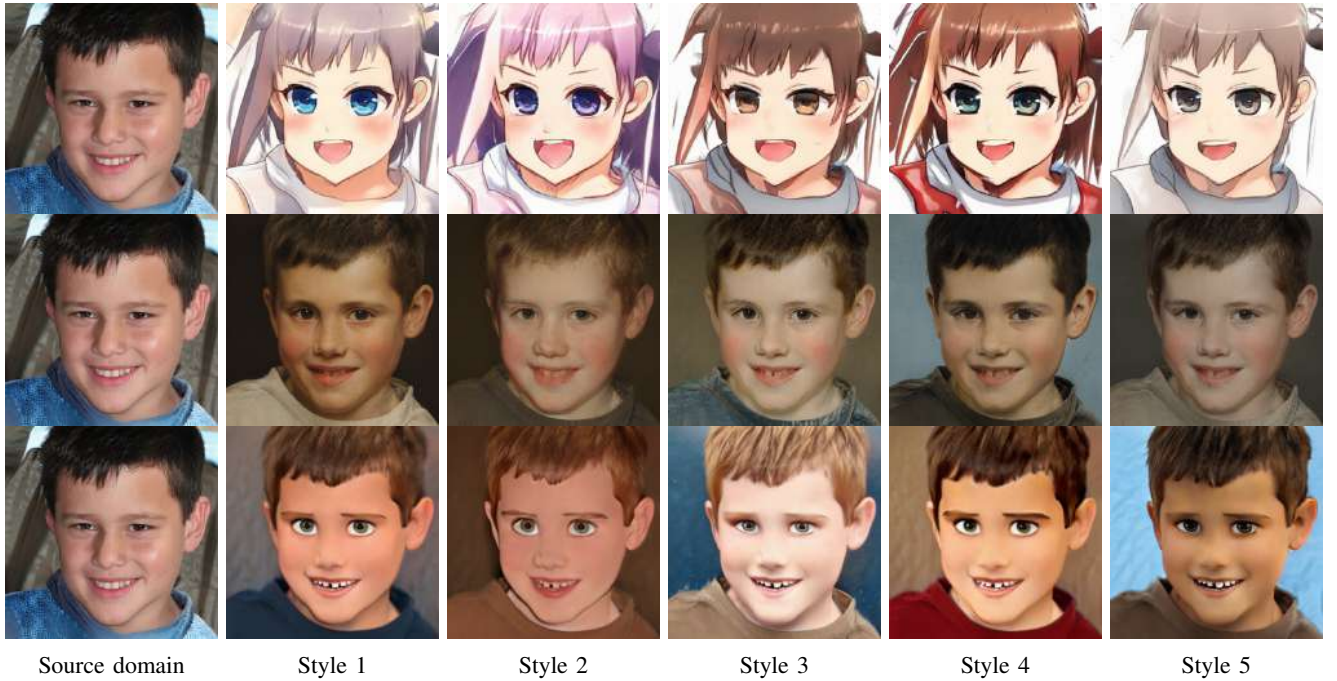


Fig. 4: Multi-modal and Multi-domain translation. Based on the same source, the transformed models can generate corresponding images with different styles in the target domains. I2I translation can be performed between any two of those domains.



Fig. 5: Reference guided image generation. The output not only preserves the style of the reference images in terms of color, light condition, texture and so forth, but it also maintains semantic similarity to the source images.

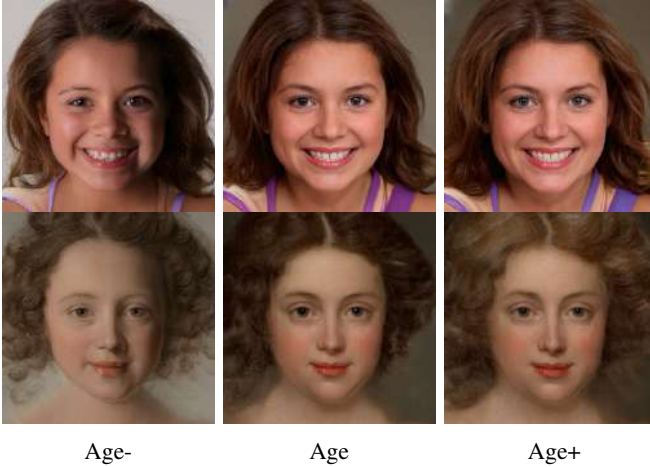


Fig. 6: GAN space invariance. After transformation, the GAN space of the base model can be inherited by the generated model. We found the age direction in the base model. By moving along this direction, the generated model (portrait model in the second row) can generate images at different ages as well.

images. However, this property cannot hold if there is a large distance between the domains themselves. For example, the anime domain is quite different from the face domain in terms of the shape of the face and mouth, and the anime does not have obvious human-like features such as noses. In such cases, with a large amount of layer-swapping, the generated output would contain artifacts (the first two rows, $LS = 5$), and they would even be out of the target domain (the last row, $LS = 5$).

Inversion Comparison. We also compared the proposed inversion method with several state-of-the-art work including the

projection method proposed by styleGAN2 [24], In-domain invert encoding and inversion method [57] and Image2StyleGAN [1]. The results are listed in Table. III, which shows that the proposed method has better reconstruction quality compared to In-domain invert methods and the Image2StyleGAN [1]. The proposed method has some superiority compared to the styleGAN2 [24] projection method. This is proved in the quantitative results.

E. Quantitative Results

The model distance was defined in Eq. 1, specifically we used LPIPS [55] to measure the distance between two images generated by two models. To perform the quantitative evaluation, we sampled 2000 latent vector z and fed them into the compared models to generate corresponding output. Then we computed the average LPIPS scores of the 2000 pairs of images. We found the proposed model transformation could drastically reduce the model distances as shown in Table. I. It is shown that the model distance is relatively large when they are trained independently (the first row), and FC-FT can achieve closer distance compared to direct FT. Moreover, layer-swap can also diminish the distance, and the more layers that are swapped, the smaller model distance will be. Furthermore, the model distances shown in Table I indicate that the original domain gap between the different dataset are different. For example, the distance between face and anime is still larger than that between face and portrait. This matches our perceptual understanding that portrait looks more like the images of face compared to anime. Besides, we found that the distance between the cartoon model and the portrait model transformed from the face model was around 0.434. Though this is a little bit higher than the results of direct transformation (0.398), it is much smaller than the original distance (0.692). This

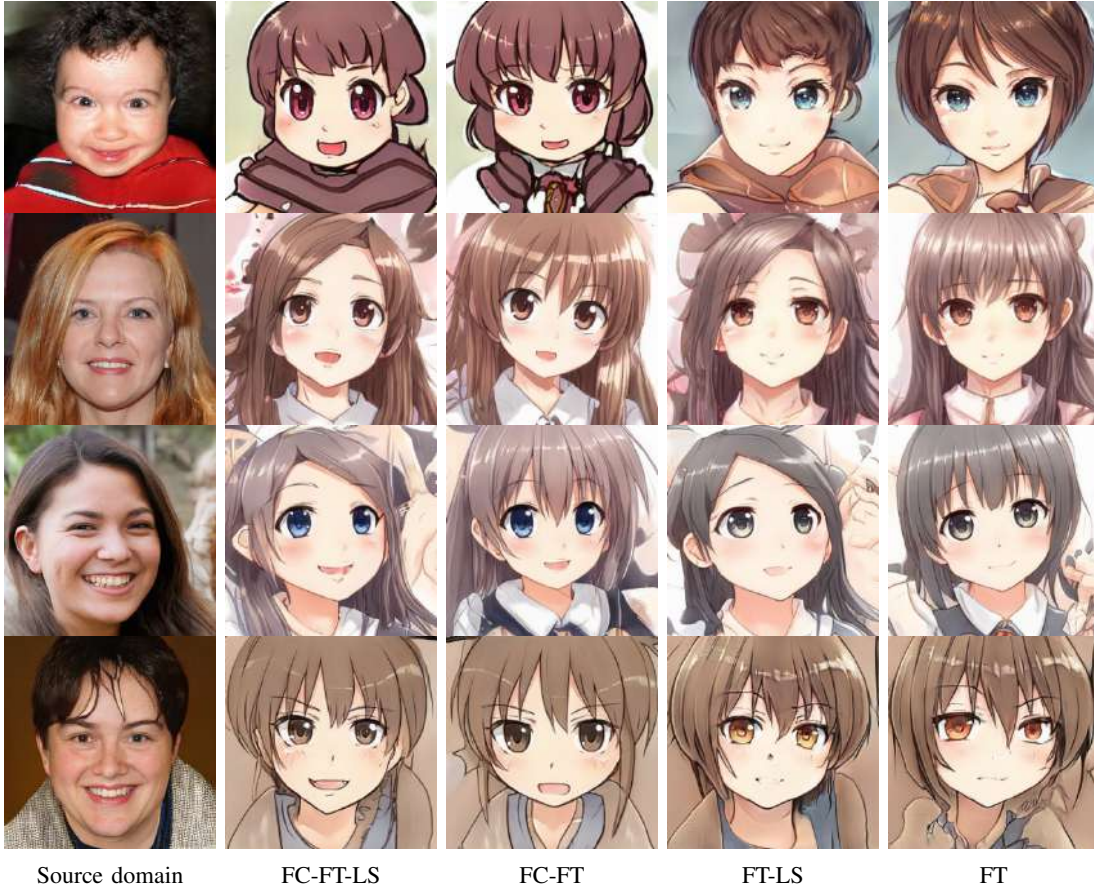


Fig. 7: Ablation study of different transformation set-ups. The combination of FC-FT-LS achieves the best results (the second column). It presents the highest semantic similarity to the source images including features like age, expression, hair and shape of the face, compared to other results.

proves that the proposed method can support multi-domain I2I translation.

We performed quantitative evaluation on the proposed method and the base line models. The results are shown in Table II. FID [13] was employed to evaluate the quality of the generated results, where 2000 test images were randomly chosen as the input. We calculated the average FID between the output and images in the target dataset. To evaluate the diversity of generated images, we computed the LPIPS distance between two randomly selected images in the generated results, which is referred as $LPIPS_d$. And we computed LPIPS between the generated image and the input to assess the semantic similarity, which is represented by $LPIPS_s$. The proposed method, including FC-FT and FC-FT-LS, can achieve the best scores on both portrait and anime datasets. For FID and diversity, the proposed method without layer-swap could obtain a slightly better score. That is because the generated image might not be as similar as the results without swapping to the images in the target domain. As for $LPIPS_s$, our method with layer-swap achieved the best scores.

We also compared the proposed method with several state-of-the-art work quantitatively, where SSIM [51] was adopted as the metric to evaluate reconstruction accuracy. The results in Table III show that our method obtains a much higher score than

In-domain inversion and Image2StyleGAN [1]. Compared to the projection method (0.655), the proposed inversion method also achieved a slightly higher score (0.667).

V. CONCLUSION

We proposed an unsupervised I2I translation method via pre-trained StyleGAN2 network, which can generate images in target domains with high quality. We first defined the model distance and tried to generate a model from a base StyleGAN2 model in the target domain through a series of model transformations. The transformed model has a relatively small distance from the base model. Then we proposed an inversion method to obtain the latent vector from an image, based on a specific StyleGAN2 model. This was then injected to the transformed StyleGAN2 network to achieve I2I translation. Our method supports multi-modal and multi-domain I2I translation with less computation resources including training time and memory. In addition, it achieves much higher performance in terms of image fidelity, diversity and semantic similarity to the input or reference images than traditional I2I translation methods. This is because the proposed method focuses on one domain at a time, while traditional I2I methods train models with data in various domains, which may cause lower resolution, artifacts and modal collapse. Moreover, both qualitative and quantitative

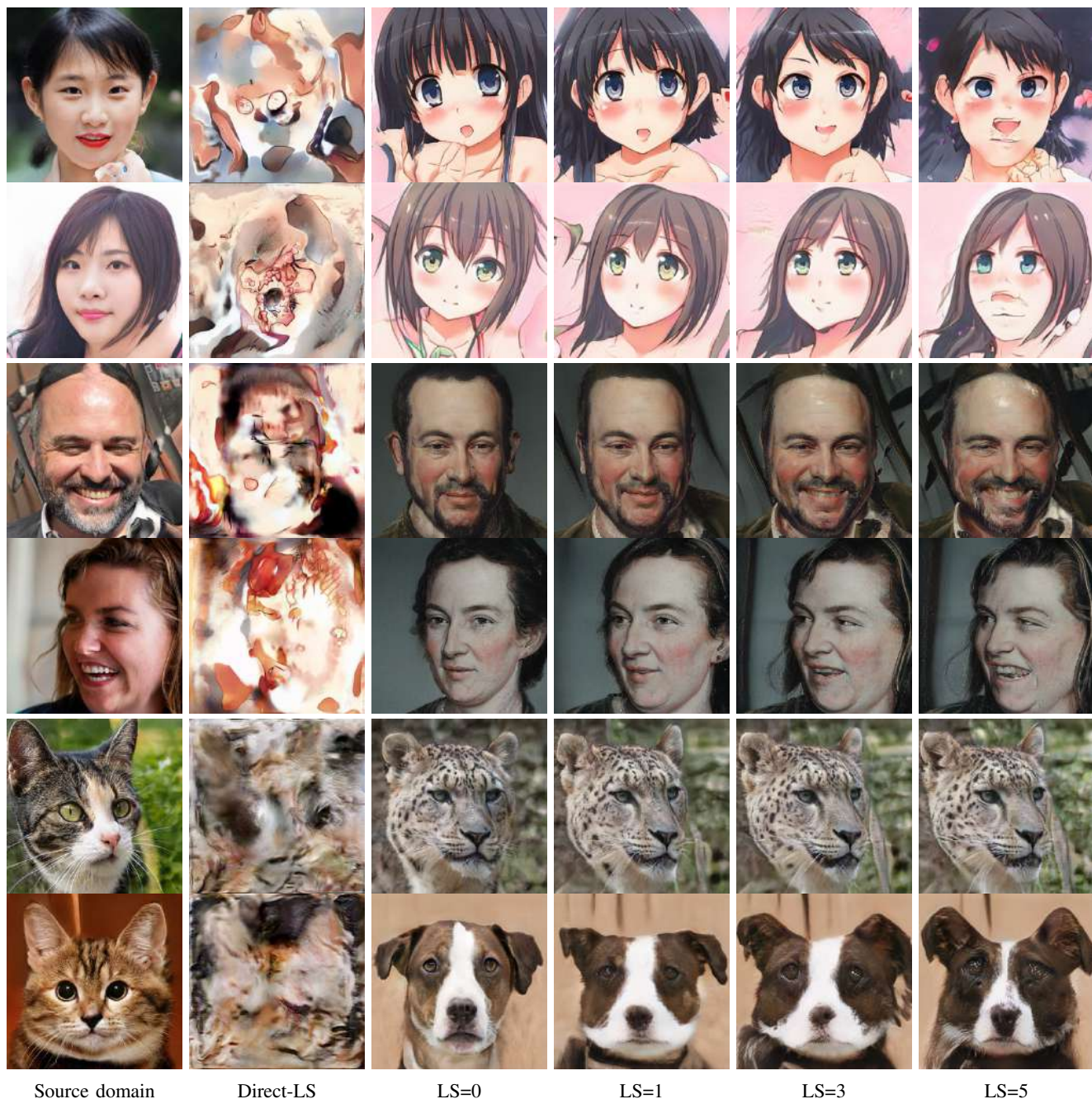


Fig. 8: Layer-swap analysis. We found direct layer-swap between two models without model transformation would not work (the second column). With more layers used in layer-swap, the generated output is more like the source image. However, as layer-swap tries to add features in one domain to images in another domain, the generated output might be corrupted since they are beyond the target domain when there is a large distance between the two domains such as a small intersection space. This is proved in face2anime (first two rows) and cat2dog (last two rows). When $LS = 5$, the output seems unnatural as it has many artifacts.

results showed that the proposed method has obvious superiority than state-of-the-art work. Although the proposed method can generate outstanding results, it does require relatively long time in test stage as the optimization based inversion method cost relatively long time then feed-forward methods. We aim to improve the performance in future work.

REFERENCES

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019.
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.
- [3] Anonymous, the Danbooru community, Gwern Branwen, and Aaron

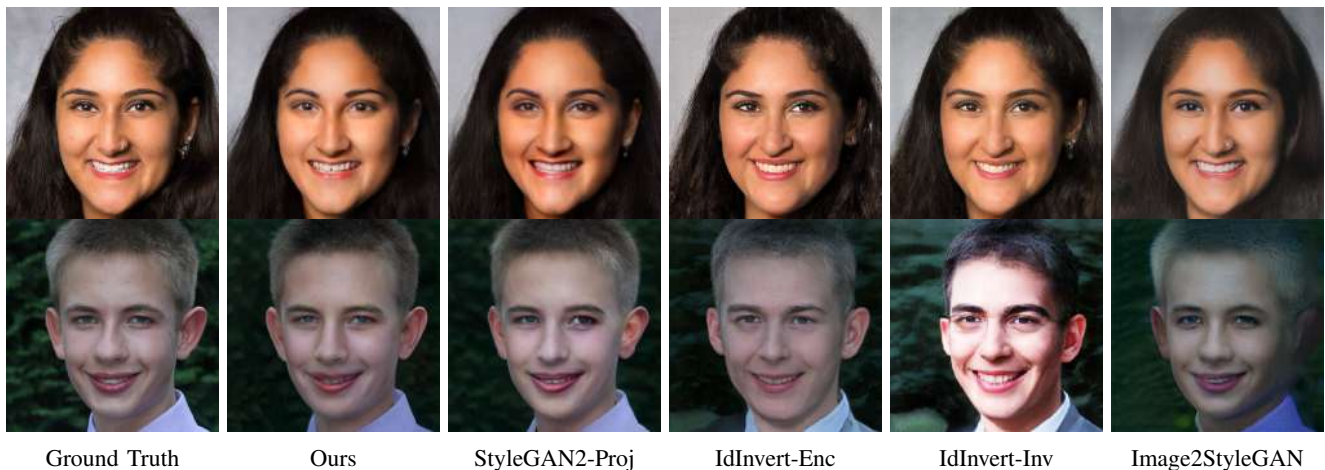


Fig. 9: Comparison of inversion methods. The proposed method is compared with other state-of-the-art inversion methods. It is obvious that our method and the StyleGAN2 projection method achieve better reconstruction results than the others, while our method also presents a slightly greater similarity to the ground truth image than the StyleGAN2 projection method. This is also proved by SSIM score in Table III

TABLE I: Model Distance between two models on different datasets. We refer to the fine-tuning process with freeze-FC as FC, fine-tune as FT and LS as layer-swap, where the number following LS indicates the number of swapped layers.

	face2anime	face2cartoon	face2portrait	cartoon2metface	cat2dog	cat2wild
w/o FT	0.747	0.632	0.759	0.692	0.672	0.593
FT	0.650	0.551	0.636	0.627	0.625	0.570
FC-FT	0.639	0.483	0.612	0.587	0.599	0.568
FC-FT-LS1	0.610	0.450	0.515	0.537	0.598	0.567
FC-FT-LS3	0.571	0.398	0.455	0.478	0.595	0.561
FC-FT-LS5	0.548	0.315	0.379	0.398	0.581	0.557

TABLE II: FID [13] and LPIPS [55] scores for the proposed method and other state-of-the-art work. We refer to the fine-tuning process with freeze-FC as FC, fine-tune as FT and LS as layer-swap. We choose $swap - layer = 3$ for anime data and $swap - layer = 5$ for portrait data as the anime data has relatively larger distance to the face data.

	face2portrait			face2anime		
	FID ↓	LPIPS _d ↑	LPIPS _s ↓	FID ↓	LPIPS _d ↑	LPIPS _s ↓
CycleGAN[58]	54.472	0.649	0.587	78.452	0.645	0.598
MUNIT[18]	62.683	0.650	0.677	66.261	0.749	0.770
DRIT++[27]	76.271	0.706	0.667	72.293	0.647	0.751
Ours (FC-FT)	32.480	0.743	0.489	52.930	0.767	0.591
Ours (FC-FT-LS)	34.147	0.742	0.468	53.737	0.767	0.573

TABLE III: SSIM [51] scores for different inversion methods

Method	SSIM ↑
Image2StyleGAN [1]	0.613
IdInvert-inv [57]	0.623
IdInvert-enc [57]	0.587
StyleGan2-proj [24]	0.655
Ours	0.667

illustration dataset. <https://www.gwern.net/Danbooru2018>, January 2019.

- [4] Imon Banerjee, Alexis Crawley, Mythili Bhethanabotla, Heike E Daldrop-Link, and Daniel L Rubin. Transfer learning on fused multiparametric mr images for classifying histopathological subtypes of rhabdomyosarcoma. *Computerized Medical Imaging and Graphics*, 65:167–175, 2018.
- [5] Suprit Bansod and Abhijeet Nandedkar. Transfer learning for video anomaly detection. *Journal of Intelligent & Fuzzy Systems*, 36(3):1967–1975, 2019.
- [6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [7] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018.
- [8] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer

- learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*, pages 243–250. Springer, 2017.
- [9] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, pages 1287–1298, 2018.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. in *advances in neural information processing systems*. 2014.
- [11] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020.
- [12] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814, 2019.
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [14] Junlin Hu. Discriminative transfer learning with sparsity regularization for single-sample face recognition. *Image and vision computing*, 60:48–57, 2017.
- [15] Jialu Huang, Jing Liao, and Sam Kwong. Semantic example guided image-to-image translation. *IEEE Transactions on Multimedia*, 2020.
- [16] Jialu Huang, Jing Liao, Zhifeng Tan, and Sam Kwong. Multi-density sketch-to-image translation network. *arXiv preprint arXiv:2006.10649*, 2020.
- [17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.
- [18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.
- [19] Zhen Huang, Sabato Marco Siniscalchi, and Chin-Hui Lee. A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition. *Neurocomputing*, 218:448–459, 2016.
- [20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [21] Doron Adler Justin Pinkney. Toonify yourself.
- [22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. *CoRR*, abs/1912.04958, 2019.
- [25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.
- [26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.
- [27] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Dri++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.
- [28] Dong Liang, Rui Wang, Xiaowei Tian, and Cong Zou. Pcgan: Partition-controlled human image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8698–8705, 2019.
- [29] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017.
- [30] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
- [31] Junyu Luo, Yong Xu, Chenwei Tang, and Jiancheng Lv. Learning inverse mapping by autoencoder based generative adversarial nets. In *International Conference on Neural Information Processing*, pages 207–216. Springer, 2017.
- [32] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In *Advances in Neural Information Processing Systems*, pages 9628–9637, 2018.
- [33] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation. *arXiv preprint arXiv:1805.11145*, 2018.
- [34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [35] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.
- [36] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.
- [37] Justin NM Pinkney and Doron Adler. Resolution dependant gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020.
- [38] Ori Press, Tomer Galanti, Sagie Benaïm, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. *arXiv preprint arXiv:2001.05017*, 2020.
- [39] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.
- [40] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [41] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*, pages 33–49. Springer, 2020.
- [42] Siyu Shao, Stephen McAleer, Ruqiang Yan, and Pierre Baldi. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 15(4):2446–2455, 2018.
- [43] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020.
- [44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [45] Masaaki Takada and Hironori Fujisawa. Transfer learning via l1 regularization. *arXiv preprint arXiv:2006.14845*, 2020.
- [46] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–5, 2018.
- [47] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019.
- [48] Tianyang Wang, Jun Huan, and Michelle Zhu. Instance-based deep transfer learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 367–375. IEEE, 2019.
- [49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. pix2pixhd: High-resolution image synthesis and semantic manipulation with conditional gans.
- [50] Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. Dual transfer learning for neural machine translation with marginal distribution regularization. In *AAAI*, pages 5553–5560, 2018.
- [51] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, volume 2, pages 1398–1402. Ieee, 2003.
- [52] Wikiart. Wikiart: portrait dataset.
- [53] Markus Wulfmeier, Ingmar Posner, and Pieter Abbeel. Mutual alignment transfer learning. *arXiv preprint arXiv:1707.07907*, 2017.
- [54] Sheng You, Ning You, and Minxue Pan. PI-REC: progressive image reconstruction network with edge and color domain. *CoRR*, abs/1903.10146, 2019.

- [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [56] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.
- [57] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [59] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.
- [60] Zhen Zhu, Tengting Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.