Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.041479

ARTICLE

Image to Image Translation Based on Differential Image Pix2Pix Model

Xi Zhao[1], Haizheng Yu[1,*] and Hong Bian[2]

[1]College of Mathematics and System Sciences, Xinjiang University, Urumqi, 830017, China

[2]School of Mathematical Sciences, Xinjiang Normal University, Urumqi, 830017, China

*Corresponding Author: Haizheng Yu. Email: yuhaizheng@xju.edu.cn

Received: 24 April 2023 Accepted: 05 September 2023 Published: 31 October 2023

ABSTRACT

In recent years, Pix2Pix, a model within the domain of GANs, has found widespread application in the field of image-to-image translation. However, traditional Pix2Pix models suffer from significant drawbacks in image generation, such as the loss of important information features during the encoding and decoding processes, as well as a lack of constraints during the training process. To address these issues and improve the quality of Pix2Pix-generated images, this paper introduces two key enhancements. Firstly, to reduce information loss during encoding and decoding, we utilize the U-Net++ network as the generator for the Pix2Pix model, incorporating denser skip-connection to minimize information loss. Secondly, to enhance constraints during image generation, we introduce a specialized discriminator designed to distinguish differential images, further enhancing the quality of the generated images. We conducted experiments on the facades dataset and the sketch portrait dataset from the Chinese University of Hong Kong to validate our proposed model. The experimental results demonstrate that our improved Pix2Pix model significantly enhances image quality and outperforms other models in the selected metrics. Notably, the Pix2Pix model incorporating the differential image discriminator exhibits the most substantial improvements across all metrics. An analysis of the experimental results reveals that the use of the U-Net++ generator effectively reduces information feature loss, while the Pix2Pix model incorporating the differential image discriminator enhances the supervision of the generator during training. Both of these enhancements collectively improve the quality of Pix2Pix-generated images.

KEYWORDS

Image-to-image translation; generative adversarial networks; U-Net++; differential image; Pix2Pix

1 Introduction

Image-to-image translation [1] is one of the most prominent tasks in the field of computer vision, which has been widely applied in photography and engineering processing. The task aims to establish a mapping function between the source domain image and the target domain image, enabling the conversion between images, such as black-and-white photos to color photos, semantic segmentation images to real-world images, realistic images to oil painting style images, etc. Common image engineering tasks such as image denoising and image super-resolution reconstruction also

belong to the image-to-image translation task. Therefore, researchers have proposed various deep-learning models to accomplish the image translation task.

In 2017, Isola et al. proposed the Pix2Pix framework [1] based on conditional generative adversarial networks [2], which are designed from the perspective of image-to-image translation tasks. It is the first universal framework for image-to-image translation tasks and is almost compatible with all single-domain image translation tasks. The Pix2Pix model has many advantages, such as a simpler structure compared to other GANs, strong universality of the model, and a stable training process. However, Pix2Pix also has some significant drawbacks. For example, the traditional Pix2Pix model is prone to losing information features during the encoding and decoding process of image generation, which results in poor-quality generated images. Moreover, during the image translation process, it is easy to generate distorted images due to a lack of strong constraints.

Therefore, in this paper, we aim to address the above-mentioned shortcomings of the Pix2Pix model and propose the following improvements:

(1) To reduce information loss, we use a more densely connected U-Net++ as the generator instead of the original U-Net generator and improve the convolutional blocks of U-Net++ to make it more suitable for image translation tasks.

(2) We add a special discriminator to the original Pix2Pix network framework to enhance the constraints during the image translation process. This discriminator discriminates the different images between the source domain and target domain images. We name it the Difference Image Discriminator.

2 Related Work

2.1 Early Image-to-Image Translation Models

Early image-to-image translation models required the establishment of different domain models for different tasks. For example, in 2001, Efros proposed the Image Quilting model [3], which is suitable for generating texture details on images, Chen et al. proposed the Sketch2Photo model in 2009 [4], which can convert sketches into realistic images, and Laffont et al. proposed the Transient Attributes model in 2014 [5], which can perform seasonal transformations on outdoor images. These models are only suitable for specific tasks and specific datasets, which greatly hindered the development of image-to-image translation. However, with the introduction of generative adversarial networks (GANs) by Goodfellow et al. in 2014 [6], these obstacles were fundamentally alleviated.

2.2 The Development of GANs in the Field of General Image-to-Image Translation

In 2017, Isola et al. proposed a general framework for image-to-image translation called Pix2Pix [1]. Pix2Pix has shown superior performance on multiple image-to-image translation tasks, and its model structure is simpler than other image-to-image translation models, with stronger training reliability. The original Pix2Pix model was improved based on conditional GANs, and the overall framework still consists of a generator and a discriminator. The network architecture of Pix2Pix is shown in the following diagram.

Fig. 1 is drawn based on an image-to-image translation task of converting sketch portraits to realistic portraits. As shown in Fig. 1, the generator part consists of a U-Net [7], with feature fusion performed using Skip-Connect from ResNet in the intermediate layers. The discriminator part is composed of a Markov discriminator, which is different from the discriminator in other generative adversarial networks in that the typical discriminator outputs a single value (real or fake) to discriminate the entire input image, while the Markov discriminator segments the whole image into

several patches and discriminates the authenticity of each patch, outputting multiple values, each of which evaluates the authenticity of the corresponding patch.

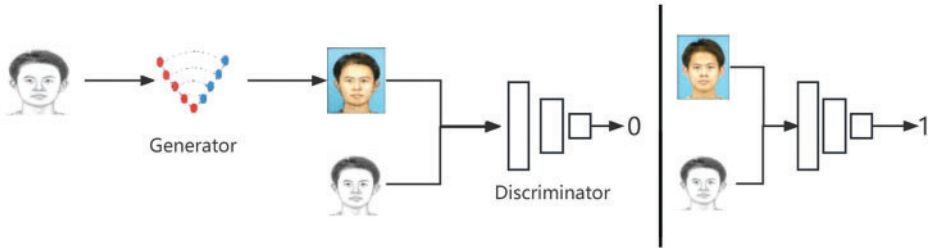

Figure 1: The structure diagram of the Pix2Pix

The optimization function of the Pix2Pix model is as follows:

$$G^* = arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{1}$$

where $\mathcal{L}_{cGAN}(G, D)$ is the loss function of the conditional GANs, and $\mathcal{L}_{L1}(G)$ is the added L1 norm to enhance the convergence of the objective function, improve the quality of images generated by Pix2Pix, and speed up the convergence rate. The specific expression is as follows:

$$\mathcal{L}_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_x[\log(1 - D(x, G(x)))] \tag{2}$$

$$\mathcal{L}_{L1}(G) = E_{x,y}[||y - G(x)||_1] \tag{3}$$

Although the Pix2Pix model has shown significant progress compared to previous models in the field of image-to-image translation, there is still significant room for improvement in terms of the quality of generated images.

After the proposal of the Pix2Pix model, GANs quickly advanced in the domain of image translation tasks. CycleGAN, introduced by Zhu et al. in 2017 [8], is a model capable of performing image-to-image translation without paired data. CycleGAN consists of two separate generative adversarial networks that respectively translate images from one domain to another. To enhance the accuracy of the translations, CycleGAN incorporates a cycle consistency loss into the objective function, ensuring the consistency of the transformations. The introduction of CycleGAN has significantly alleviated the challenge of acquiring paired datasets.

In 2017, Liu et al. proposed the Unsupervised Image-to-Image Translation Network (UNIT) [9]. UNIT assumes the existence of a latent factor space, where input images are encoded into latent factors and subsequently decoded into target domain images during the image-to-image translation process. Multiple cycle-consistency losses are incorporated into the objective function of the UNIT to improve the stability of the training process and enhance the quality of the generated images.

In 2018, Huang et al. proposed the Multimodal Unsupervised Image-to-Image Translation (MUNIT) [10], which enables translation across multiple target domains, whereas traditional image-to-image translation models only cater to single-modal image-to-image translation tasks. MUNIT assumes the existence of two distinct factor spaces: a content factor space and a style factor space. The content factor space is shared across different domains, while the style factor space is domain-specific and non-shareable. By combining a content factor with different style factors and going through the decoding process, images from different target domains can be translated.

Starting in 2018, the development of GANs in the domain of general image-to-image translation has slowed down. However, with the emergence of new models like GPT by OpenAI and Segment Anything by Meta in 2022, researchers have intensified their efforts in the study of Artificial General Intelligence (AGI). Consequently, we believe that further innovation and improvement in GANs for general image-to-image translation are highly meaningful in the current context.

2.3 The Development of GANs in the Field of Specific Image-to-Image Translation

Since 2018, GANs have exhibited remarkable progress in specific image translation tasks. Among these, StarGAN [11] stands out as a model specialized in facial transformation, facilitating translations across multiple target domains. StarGAN successfully overcomes the previous limitations of facial transformation models, which were confined to binary domain conversions, thereby enhancing the flexibility of translations to other domains. In 2019, StarGAN-V2 [12] was introduced, incorporating the concepts of style diversity and domain diversity to further enhance the quality of transformations.

PSGAN, introduced in 2021 [13], represents the pioneering attempt to employ GANs for generating high-quality pan-sharpened images. PSGAN accepts panchromatic and multispectral images as input and maps them to the desired high-resolution images while selecting the optimal solution from various architectures and designs.

In 2022, Amirkolaee et al. proposed a specialized GAN model for medical image translation at the image level. This model effectively combines local and global features, resulting in commendable performance [14].

2.4 The Development of GANs in the Field of Multimodal Image-to-Image Translation

Since 2019, multimodal tasks have become a mainstream direction in the development of deep learning models [15–18]. As a result, there have been several multimodal models [19–23] in image-to-image translation. For example, in 2019, Park et al. proposed GauGAN [19], which transforms an image by inputting a semantic mask and its corresponding real image, producing impressive results. In 2022, Yan et al. proposed MMTrans [20], which is based on the Swin Transformer and GAN for medical image translation.

Since 2022, large-scale models based on diffusion models [24–26] and CLIP [27] technology have gained significant attention in the field of text-to-image and image translation, causing a global sensation in AI art with generated images surpassing those generated by GANs in terms of image quality. However, in practical commercial applications, generative adversarial networks have matured significantly, and it is expected that they will remain one of the main models in the field of image generation for a considerable period. Therefore, researching generative adversarial networks in the field of image translation is still of great significance in the present and future.

3 Models and Methods

3.1 U-Net++ Generator

The original Pix2Pix uses a U-Net network as the generator to generate images. The U-Net network was proposed by Ronneberger in 2015 and plays a crucial role in medical image and semantic segmentation fields. As shown in Fig. 2, the U-Net network consists of two parts. The first part is the encoder, which is composed of multiple downsampling layers. The input image is encoded by the encoder to extract deep information. The second part is the decoder, which is composed of multiple

upsampling layers. The deep information extracted by the encoder is decoded by the decoder. To reduce information loss in this process, skip connections from ResNet are used.

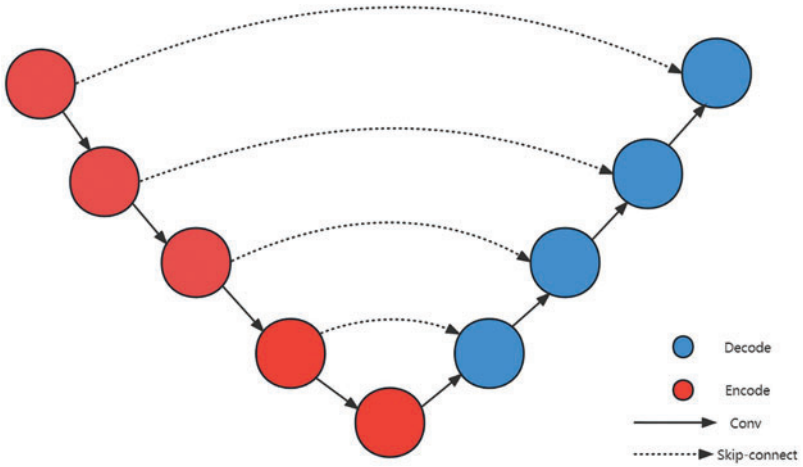

Figure 2: The structure diagram of the U-Net

Although the U-Net network performs remarkably well in image-to-image translation tasks such as medical image segmentation, it has some significant drawbacks. For instance, researchers cannot determine the optimal depth of the network, and the addition of skip connections for feature fusion in U-Net imposes an unnecessary constraint by only allowing skip connections between feature maps of the same scale. Due to this constraint, U-Net's feature fusion significantly affects the image quality generated by the network, leading to some unnecessary information loss. Therefore, if Pix2Pix continues to use U-Net as its generator, the quality of the generated images will inevitably be affected for the same reasons. Fortunately, effective solutions have been proposed to address these drawbacks.

In 2020, Zhou et al. proposed the U-Net++ [28], which is a very powerful model that improves upon the U-Net. It effectively overcomes the shortcomings of the original U-Net. As shown in Fig. 3 [28], the U-Net++ network uses denser and more clever skip connections, allowing for feature fusion of information extracted at different scales and depths. Multiple decoders can share or partially share the same encoder, which largely avoids unnecessary information loss.

Our improved Pix2Pix model, in terms of the generator, abandons the traditional U-Net and adopts the U-Net++ for image generation. However, during the initial experimental verification, it was found that a large amount of noise interference appeared in the generated images after adversarial training using the original U-Net++ generator. After analysis and consideration, it was believed that this was because U-Net++ was proposed for the special task of medical image segmentation, which can also be seen as a classification problem. Therefore, the U-Net++ used VGG block, which is suitable for classification tasks, but not for other image-to-image translation tasks. To make the generator more adaptable to a wider range of image-to-image translation tasks, the block used in this paper's U-Net++ consists of a stridden convolution layer, a channel normalization layer, and a Dropout layer, which performs better than the VGG block on non-image-segmentation-tasks.

It is evident from Fig. 4 that the face generated by the U-Net++ before improvement has some facial features that are not only more blurred but also have some light blue noise interference. The improved U-Net++ generates higher-quality images in the task of generating facial images, resulting in more natural and brighter facial expressions with no noise interference on the face.

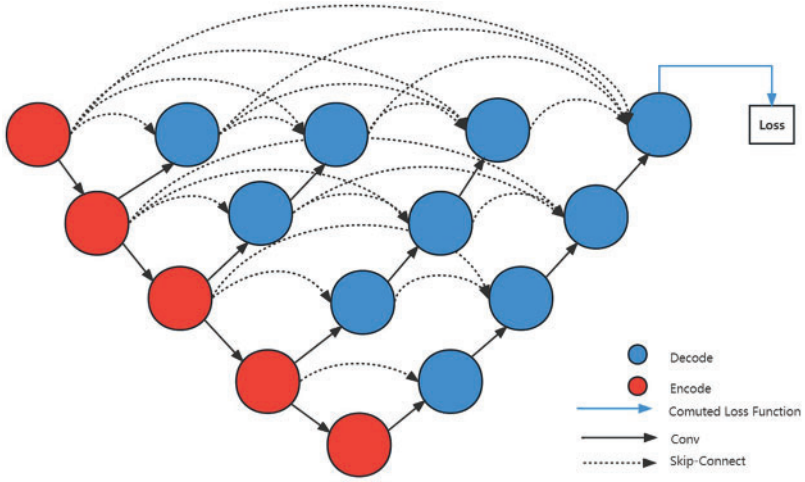

Figure 3: The structure diagram of the U-Net++

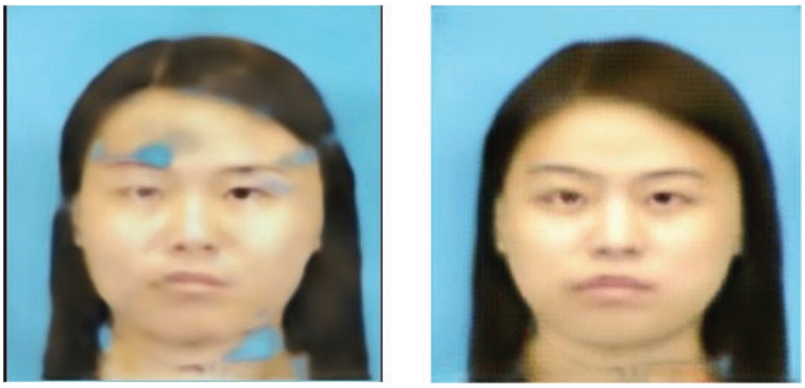

Figure 4: Comparison of generated images before and after improvement of U-Net++. (a) Images generated by the original U-Net++ generator. (b) Images generated by the improved U-Net++ generator

3.2 Differential Image Discriminator

To further strengthen the constraints on the generated images, we have added a differential image discriminator to the traditional Pix2Pix network structure. As shown in Fig. 5, the differential image here is obtained by taking the difference between the target domain image and the source domain image. This discriminator only discriminates between the two types of differential images generated. If the discriminator receives the differential image between the GroundTruth in the target domain and the input of the generator, it is judged as true; if it receives the differential image between the fake target domain image generated by the generator and its input, it is judged as false.

The network structure of Pix2Pix with the addition of a differential image discriminator is shown in the following Fig. 6.

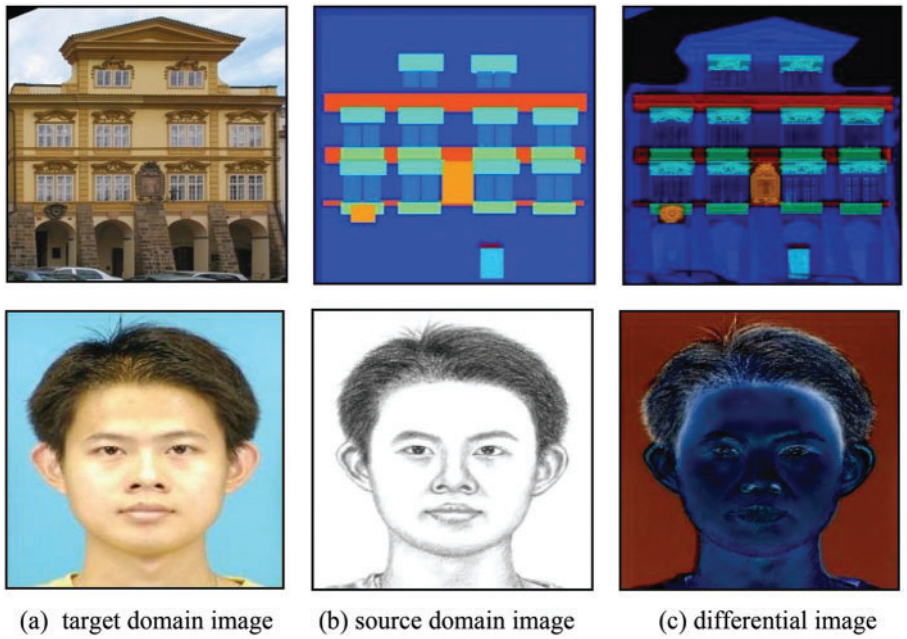

Figure 5: Example of differential image. (a) Real target domain images. (b) Source domain image inputted to the generator. (c) The difference image is obtained by subtracting the source domain image from the target domain image

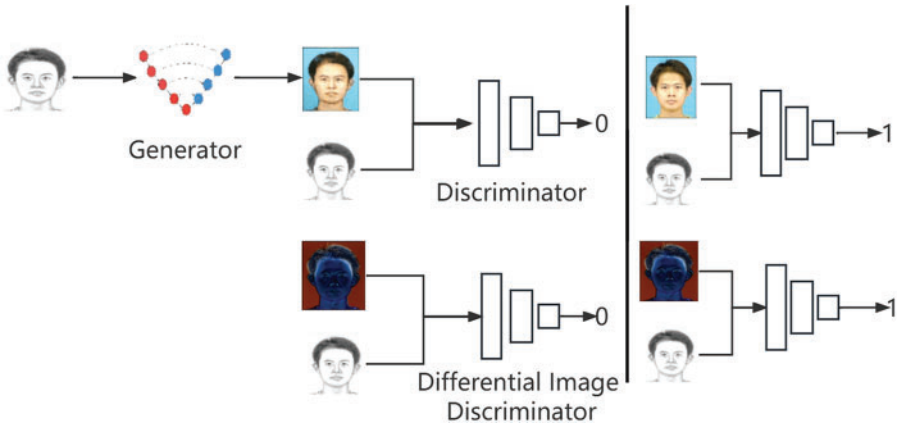

Figure 6: The network structure of the Pix2Pix with differential image discriminator

If we consider the image generation process of the generator in the Pix2Pix network as a simple addition model:

$$x + \hat{g} = G(x) \tag{4}$$

In this context, x represents the input image to the generator, $G(x)$ is the output image generated by the generator, and $\hat{g}$ is the additional data that the generator needs to simulate. The differential image discriminator proposed in this article is used to discriminate between the generated additional data $\hat{g}$ and the real additional data g, thereby providing stronger guidance to the generator during the

training process. Correspondingly, the optimization function used during the training of the Pix2Pix network has also changed. The optimization function of the Pix2Pix network in this article is:

$$G^* = arg \min_{G} \max_{D,D_{differ}} \mathcal{L}_{cGAN}\left(G, D, D_{differ}, \hat{g}, g\right) + \lambda \mathcal{L}_{L1}(G) \tag{5}$$

The specific formula for $\mathcal{L}_{cGAN}\left(G, D, D_{differ}, \hat{g}, g\right)$ and $\mathcal{L}_{L1}(G)$is:

$$\begin{aligned}\mathcal{L}_{cGAN}\left(G, D, D_{differ}, \hat{g}, g\right) &= E_{x,y}[\log D(x, y)] + E_x[\log(1 - D(x, G(x)))] \\ &+ E_{x,g}\left[\log D_{differ}(x, g)\right] + E_{x,\hat{g}}\left[\log\left(1 - D_{differ}\left(x, \hat{g}\right)\right)\right]\end{aligned} \tag{6}$$

$$\mathcal{L}_{L1}(G) = E_{x,y,z}[y - G(x)] \tag{7}$$

In this case, G represents the generator, D represents the regular discriminator, D_{differ} represents the differential image discriminator proposed in this chapter. x is the input source domain image, y is the real target domain image, G(x) is the fake target domain image generated by the generator, g is the differential image between the real target domain image and the source domain image and $\hat{g}$ is the differential image between the fake target domain image and the source domain image.

From the above formula, it can be seen that the objective optimization function of Pix2Pix based on the differential image discriminator has a significant change compared to the original Pix2Pix objective optimization function at Eq. (1). This can be considered in adding a strong constraint to the entire Pix2Pix network on the original basis, which can more effectively guide the training of the generator, allowing the generator to more smoothly generate the required additional data.

4 Results

4.1 Experimental Environment and Model Training Steps

We selected the open-source facades dataset and the CUHK Sketch Portrait Dataset for model training and validation. As shown in Table 1, the former consists of 606 images of buildings and their corresponding label maps, with 400 images used for training, 100 images used for testing, and 106 images used for validation. The latter consists of 188 facial images of CUHK students and their corresponding face sketch images, with 168 images used for training, 10 images used for testing, and 10 images used for validation.

Table 1: Dataset distribution

Dataset	Training dataset	Testing dataset	Validation dataset
Facades	400	100	106
CUHK	168	10	10

The model was trained on a Linux system with a 4-core Intel(R) Xeon(R) Gold 6330 CPU and one NVIDIA GeForce RTX 3090 GPU. The programming language used was Python 3.8, and the deep learning framework employed was PyTorch 1.8 with CUDA version 11.1. The training process consisted of 100 epochs with a batch size of 8.

The pseudo-codes for training the original Pix2Pix and the Pix2Pix with the U-Net++ generator are shown below:

Algorithm 1: Pix2Pix training algorithm

Input: $y \sim P_{data}(y), x \sim P_{data}(x)$

Output: $G(x), D(y, x)$

1: for 100 **do**

2: Sampling from the distribution of input images, with a sample size of 10, results in the set $X = (x^1, x^2, x^3, \ldots, x^{10})$.

3: Sampling from the distribution of target domain images, with a sample size of 10, results in the set $Y = (y^1, y^2, y^3, \ldots, y^{10})$.

4: To update the parameter weights of the discriminator using the following gradient:

$$\nabla_{\theta_d} \frac{1}{10} \sum_{i=1}^{10} [\log D(y^i, x^i) + \log(1 - D(G(x^i), x^i))]$$

5: To update the parameter weights of the generator using the following gradient:

$$\nabla_{\theta_g} \frac{1}{10} \sum_{i=1}^{10} [\log(1 - D(G(x^i), x^i)) + \| y^i - G(x^i) \|_1]$$

6: end for

7: return $G(x), D(y, x)$

The pseudo-code for training our joint model is presented as follows:

Algorithm 2: Our joint model training algorithm

Input: $y \sim P_{data}(y), x \sim P_{data}(x)$

Output: $G(x), D(y, x), D_d(g, x)$

1: for 100 **do**

2: Sampling from the distribution of input images, with a sample size of 10, results in the set $X = (x^1, x^2, x^3, \ldots, x^{10})$.

3: Sampling from the distribution of target domain images, with a sample size of 10, results in the set $Y = (y^1, y^2, y^3, \ldots, y^{10})$.

4: Obtaining two types of differential image samples: $g = \left(g^1, g^2, g^3, \ldots, g^{10}\right)$, $\hat{g} = (\hat{g}^1, \hat{g}^2, \hat{g}^3, \ldots, \hat{g}^{10})$.

5: Updating the parameter weights of the regular discriminator using the following gradient:

$$\nabla_{\theta_d} \frac{1}{10} \sum_{i=1}^{10} [\log D(y^i, x^i) + \log(1 - D(G(x^i), x^i))]$$

6: Updating the parameter weights of the differential image discriminator using the following gradient:

$$\nabla_{\theta_d} \frac{1}{10} \sum_{i=1}^{10} \left[\log D_d(g^i, x^i) + \log\left(1 - D_d\left(\hat{g}^i, x^i\right)\right)\right]$$

7: To update the parameter weights of the generator using the following gradient:

$$\nabla_{\theta_g} \frac{1}{10} \sum_{i=1}^{10} \left[\log(1 - D(G(x^i), x^i)) + \log\left(1 - D_d\left(\hat{g}^i, x^i\right)\right) + \| y^i - G(x^i) \|_1\right]$$

8: **end for**

9: return $G(x), D(y, x), D_d(g, x)$

4.2 Evaluation Metrics

In terms of selecting evaluation metrics, the quality of the images generated by the image generation model was evaluated. Besides directly showing the generated images, there are no very

comprehensive objective metrics both domestically and abroad. Metrics such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) only calculate the pixel value differences between the real image and the generated image, and cannot reasonably compare the overall structure, texture, brightness, color, and other features between the images. Therefore, for the images generated by generative adversarial networks, researchers generally use the Inception Score (IS) [29] and Fréchet Inception Distance (FID) [30] to evaluate the generated images, which simulate subjective evaluation by the human eye. Therefore, this paper selects Inception Score, Fréchet Inception Distance, and Structural Similarity (SSIM) [31] as evaluation metrics.

The formula for the Inception Score is as follows:

$$IS\left(G\right)=\exp\left(E_{x\sim pg}D_{KL}\left(p\left(y|x\right)||p\left(y\right)\right)\right) \tag{8}$$

where x is the generated image, $p\left(y|x\right)$ is the class probability distribution obtained by inputting x into the Inception V3 classification neural network, and $p\left(y\right)$ is the marginal distribution probability obtained by averaging the predicted class probabilities for each generated image. D_{KL} is the KL divergence from $p\left(y|x\right)$ to $p\left(y\right)$. From Eq. (8), it can be seen that IS only measures the distance between input images and the ImageNet dataset.

The specific formula for Fréchet Inception Distance (FID) is as follows:

$$FID\left(y,x\right)=||\mu_y-\mu_x||_2^2+\mathrm{Tr}\left(\sum\nolimits_x+\sum\nolimits_y-2\left(\sum\nolimits_x\sum\nolimits_y\right)^{\frac{1}{2}}\right) \tag{9}$$

where x represents the generated image, y represents the real image, and assuming that both follow high-dimensional distributions, the statistics are denoted as $(\mu_x, \sum_x)$ and $(\mu_y, \sum_y)$. From the above formula, it can be seen that FID measures the distance between two distributions. Therefore, the smaller the value of FID, the closer the generated image data distribution is to the real image data distribution. In addition, because FID is calculated based on the features that appear in the image, it cannot calculate the spatial position relationship between features, so FID has some controversy.

The specific formula for Structural Similarity (SSIM) is as follows:

$$SSIM\left(x,y\right)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(2\Sigma_{xy}+c_2\right)}{\left(\mu_x{}^2+\mu_y{}^2+c_1\right)\left(\Sigma_x{}^2+\Sigma_y{}^2+c_2\right)} \tag{10}$$

Here, x and y represent individual sampled images from the generated and real images, respectively, μ_x and μ_y are the estimated pixel value means for x and y, Σ_x and Σ_y are the estimated variances, $c_1=(k_1L)^2$, $c_2=(k_2L)^2$, where L is the range of pixel values, $k_1=0.01$, $k_2=0.03$. Therefore, the SSIM metric can only measure a single sampled image at a time. In this paper, the SSIM was computed for 100 generated and corresponding real images, and the mean value was taken.

4.3 Experimental Analysis

4.3.1 Optimal Depth Exploration of U-Net++ Generator

U-Net++ performs well when the depth is constrained. For example, for a certain task, the optimal depth of the U-Net model is L4, but even when the depth of the U-Net++ model exceeds 4 (such as L6 or L8), its performance can still be comparable to or even better than the U-Net model at L4. However, when the optimal depth of the U-Net model is greater than the maximum depth of the U-Net++ model, the performance of U-Net++ will be worse than that of U-Net. Therefore, before conducting empirical analysis based on the U-Net++ generator in Pix2Pix, it is necessary to explore the maximum depth of U-Net++. We conduct exploration experiments using the facades datasets.

Based on Fig. 7, it can be visually perceived that as the depth of the U-Net++ generator increases, the quality of the generated images gradually improves. Among them, it is difficult to conclude the quality of the images generated by the U-Net++ generator with a depth of 6 and the one with a depth of 8, but it can be observed that the difference between them is very small and the improvement range is also very small.

Figure 7: Empirical results of Pix2Pix with U-Net++ generators at different depths. (a) Real target domain images. (b) Images generated by U-Net++ generator with a depth of 2 in Pix2Pix. (c) Images generated by U-Net++ generator with a depth of 4 in Pix2Pix. (d) Images generated by U-Net++ generator with a depth of 6 in Pix2Pix. (e) Images generated by U-Net++ generator with a depth of 8 in Pix2Pix

Based on Table 2, it can be seen that the three selected evaluation metrics generally improve with the increase of U-Net++ generator depth. Among them, the images generated by the U-Net++ generator with a depth of 8 achieved the best performance in the IS and SSIM evaluation metrics, while the images generated by the U-Net++ generator with a depth of 6 achieved the best performance in the FID evaluation metric.

Table 2: Performance metrics of U-Net++ at different depths

Model	FLOPS	IS↑	FID↓	SSIM↑
L2	11.21 G	1.315	231.4476	0.16713
L4	19.89 G	1.611	220.1723	0.19332
L6	28.12 G	1.729	**200.1185**	0.20892
L8	36.37 G	**1.747**	201.0625	**0.20989**

In addition, by observing the changes in the indicator data in the table, it can be understood that the greatest improvement in each indicator occurred when the depth increased from 2 to 6, while the improvement in each indicator was very small when the depth increased from 6 to 8, and entered a stagnant state. The U-Net++ generators with depths of 6 and 8 are very close in all three evaluation metrics. Therefore, considering all the evaluation metrics, this study determined the depth of the U-Net++ generator to be 8.

4.3.2 Experimental Results

For the task of semantic label-to-building image translation on the facades dataset, we select CycleGAN and the original Pix2Pix as the base models and compare them with Pix2Pix using only U-Net++ as the generator, Pix2Pix with only the differential image discriminator added, and Pix2Pix using both U-Net++ generator and differential image discriminator. The comparison results are shown in Fig. 8.

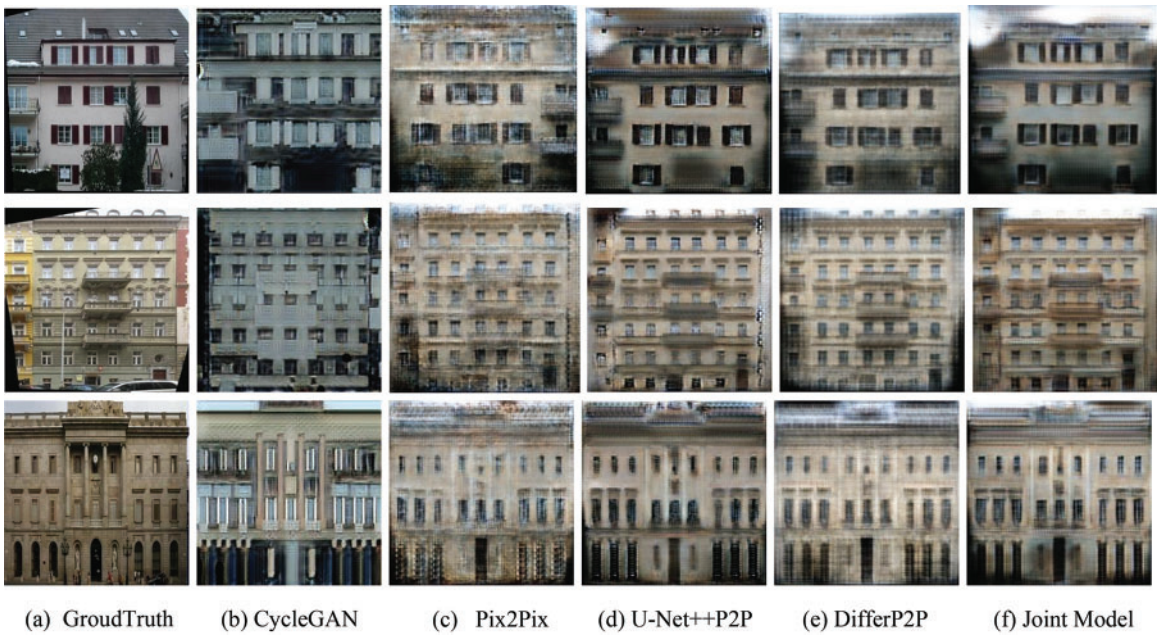

Figure 8: The validation results for the facades dataset. (a) GroudTruth: Real target domain images. (b) Validation results generated by CycleGAN. (c) Validation results generated by Pix2Pix. (d) U-Net++P2P: Validation results generated by Pix2Pix using U-Net++ generator. (e) Validation results generated by Pix2Pix using differential image discriminator. (f) Validation results generated by Pix2Pix using both U-Net++ generator and differential image discriminator

As shown in Fig. 8, compared with CycleGAN and the original Pix2Pix, the improved model in this paper generates images with significant improvements in terms of image quality and clarity, whether it is Pix2Pix using only U-Net++ as the generator (d), Pix2Pix with only the differential image discriminator added (e), or the joint model in this paper. Among them, the results of the CycleGAN network are the worst, which is because CycleGAN is an unsupervised model. In terms of image details, the images generated by Pix2Pix using only U-Net++ as the generator and the joint model in this paper have more delicate and clear details.

In the image-to-image translation task of transforming sketches into realistic portraits on the Sketch Portrait Dataset at the Chinese University of Hong Kong, the same base models were selected as in the previous experiment, and they were compared with the improved Pix2Pix model.

Based on the results shown in Fig. 9, on the sketch portrait dataset, the unsupervised CycleGAN model performs the worst, with significant color differences and uneven color distribution compared to the real images. Meanwhile, compared to the face images generated by the original Pix2Pix, the face

images generated by the proposed model in this paper are closer to real images in terms of skin color and facial pose. The generated face images are also brighter and more realistic.

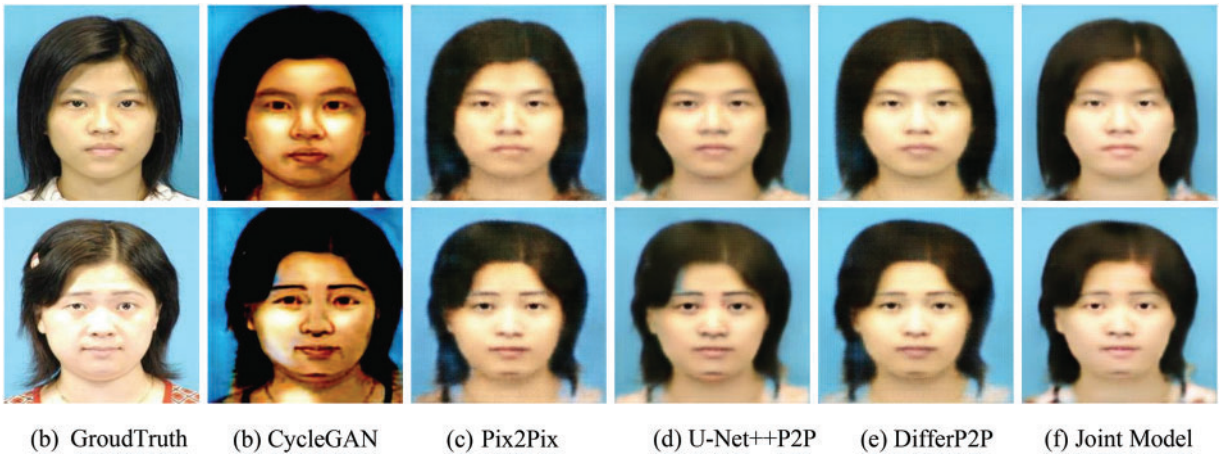

Figure 9: The validation results for the Sketch Portrait Dataset at the Chinese University of Hong Kong

Using the two datasets, validation images were generated for each model, and corresponding evaluation metrics were selected for evaluation.

According to Table 3, Our improved Pix2Pix models outperformed the original model. The Pix2Pix model with only U-Net++ as the generator had the largest improvement in the FID metric, decreasing by 43 points, but had a smaller improvement in the IS and SSIM metrics. The Pix2Pix model with only the differential image discriminator had the largest improvement in the IS and SSIM metrics, increasing by 17.1% and 3.9%, respectively, but had a smaller improvement in the FID metric. The joint model had the most evenly distributed improvement in all three metrics, with the largest improvement in the FID and SSIM metrics, decreasing by 38 points and increasing by 2.5%, respectively.

Table 3: The index of each model on facades

Models	FLOPS	IS↑	FID↓	SSIM↑
CycleGAN	68.73 G	1.112	262.6379	0.03346
Pix2Pix	19.07 G	1.714	244.3743	0.19911
U-Net++P2P	36.37 G	1.747	**201.0652**	0.20989
Differ P2P	20.79 G	**1.885**	244.3423	**0.238559**
Joint model	38.09 G	1.737	206.3197	0.224668

In terms of computational complexity, the CycleGAN model has the highest complexity, reaching 68.73 G, due to its complex architecture consisting of two generators and two discriminators. The next highest complexity is found in our proposed joint model, which reaches 38.09 G. This increase is mainly attributed to the utilization of the U-Net++ generator. Both the Pix2Pix model with the differential image discriminator and the original Pix2Pix model have relatively lower computational complexities, with only a minor difference between them.

According to Table 4, we can see that the three improved Pix2Pix models have all shown improvements over the original model in both evaluation metrics. Among them, using only U-Net++ as the generator in Pix2Pix has shown the greatest improvement in the IS metric, with an increase of 28.3%. In the SSIM metric, the joint model has shown the greatest improvement, with an increase of 3.84%. In the comparison between U-Net++ Pix2Pix, Differ Pix2Pix, and the joint model, it can be seen that the joint model has shown the most balanced improvement in both metrics.

Table 4: The index of each model on the Sketch Portrait Dataset

Models	IS↑	SSIM↑
CycleGAN	1.2379	0.1947
Pix2Pix	1.3223	0.6190
U-Net++P2P	**1.6036**	0.6476
Differ P2P	1.4225	0.6374
Joint model	1.5700	**0.6574**

4.3.3 Experimental Comparison of Two Types of Discriminators

From Section 3.2, it can be seen that the differential image discriminator does indeed improve the performance of Pix2Pix. However, empirical analysis in Section 3.2 alone is not sufficient to determine whether the differential image works as expected. If an ordinary discriminator is added to the original Pix2Pix network framework, can it also achieve the same level of performance improvement as the differential image discriminator-based Pix2Pix?

To verify this issue, we set up an additional Pix2Pix model with a regular discriminator called the dual-discriminator Pix2Pix, and compared it with the Pix2Pix model based on the differential image discriminator and the original Pix2Pix.

Fig. 10 shows the results of the discriminator comparison experiments on the facades dataset and the sketch-to-portrait dataset. It can be seen intuitively that the dual-discriminator Pix2Pix has indeed improved the quality of the generated images to some extent compared to the original Pix2Pix, and the images are also clearer, but the improvement is limited. The Pix2Pix based on the differential image discriminator generates the best image quality. Compared with the images generated by the dual-discriminator Pix2Pix, the brightness and darkness distribution of the images are more uniform, and the picture is also more delicate.

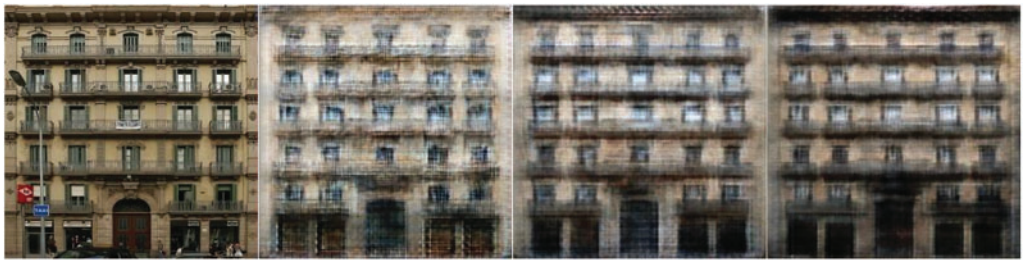

Figure 10: (Continued)

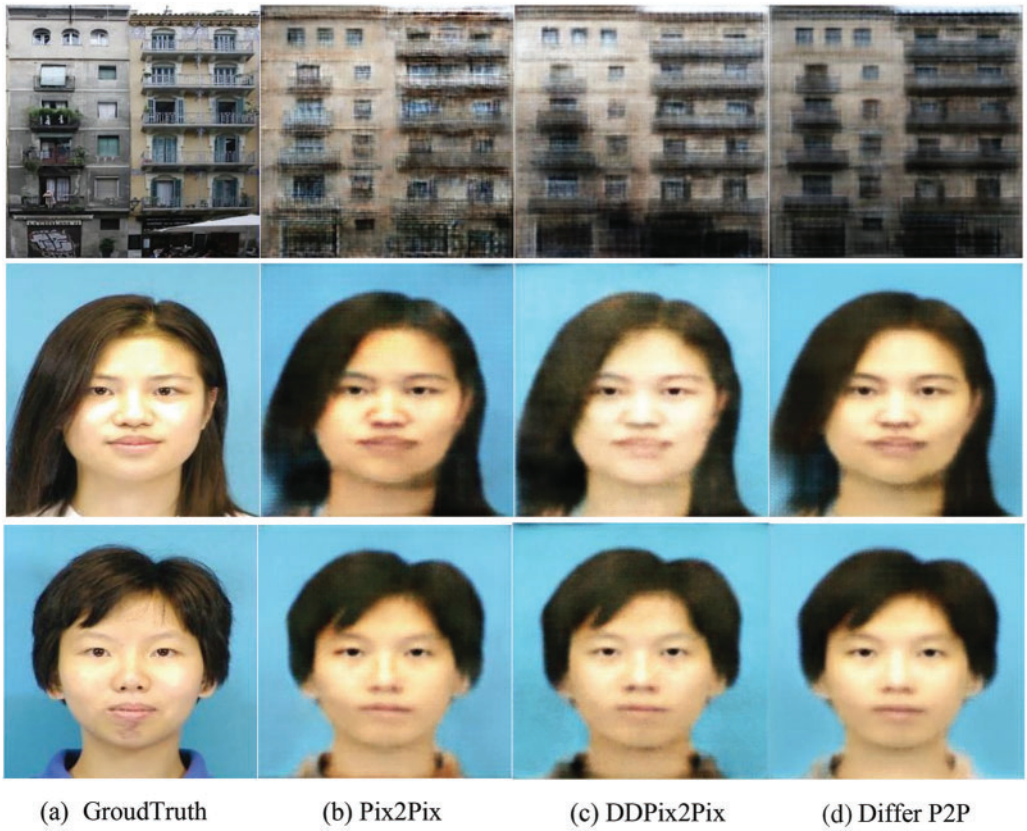

Figure 10: Example of verification results for comparative experiments. (a) GroudTruth: Real target domain images. (b) Validation results generated by Pix2Pix. (c) DDPix2Pix: dual-discriminator Pix2Pix. (d) Validation results generated by Pix2Pix using differential image discriminators

According to Table 5, based on the IS and SSIM metrics, the performance of the Pix2Pix model with a differential image discriminator exceeds that of the original Pix2Pix and the dual-discriminator Pix2Pix, achieving the best performance among the three. While the dual-discriminator Pix2Pix also outperforms the original Pix2Pix in all aspects, the improvement is significantly less than that of the differential image discriminator-based Pix2Pix. For example, on the facades dataset, compared to the original Pix2Pix, the dual-discriminator Pix2Pix improves by only 3.3% on the IS metric, far less than the 9.9% improvement of the differential image discriminator-based Pix2Pix. Similarly, on the SSIM metric, the dual-discriminator Pix2Pix improves by only 5.5%, which is also smaller than the 19.8% improvement of the differential image discriminator-based Pix2Pix. Therefore, it can be concluded that the performance gain brought by adding a differential image discriminator far exceeds that brought by adding a conventional discriminator. Based on the additive model assumption in Section 2.3, differential images play a critical role in guiding the discriminator to identify the generated content in the Pix2Pix model with a differential image discriminator.

Table 5: The index of comparative experiments

Models	Datasets	IS↑	FID↓	SSIM
Pix2Pix	facades	1.714	244.3743	0.19911
DD Pix2Pix	facades	1.772	244.3516	0.21011
DifferPix2Pix	facades	**1.885**	**244.3423**	**0.23856**
Pix2Pix	sketch2portrait	1.3223	/	0.6190
DD Pix2Pix	sketch2portrait	1.3913	/	0.6210
DifferPix2Pix	sketch2portrait	**1.4225**	/	**0.6374**

5 Conclusion

The structure of the original Pix2Pix model was improved to address some limitations in image-to-image translation tasks. The U-Net++ was adopted as the generator, and a differential image discriminator was added. Experimental results demonstrated that the proposed improvements effectively enhanced the image generation quality of the Pix2Pix model, resulting in clearer and more detailed facial features. Furthermore, the experiments confirmed the crucial role of the differential image in the performance improvement of the differential image discriminator. However, our proposed joint model did not outperform the models with a single improvement in certain metrics, which may be attributed to unknown interactions introduced by the combination of the two improvements. Therefore, investigating how to mitigate these unknown effects can be considered in future research work.

Acknowledgement: Authors gratefully acknowledge technical and financial support from the College of Mathematics and System Sciences, Xinjiang University, and Xinjiang Natural Science Foundation of China. We acknowledge the data resources from "Kaggle Inc" (https://www.kaggle.com/).

Funding Statement: This work is supported in part by the Xinjiang Natural Science Foundation of China (2021D01C078).

Author Contributions: Study conception and design: Xi Zhao, Haizheng Yu; data collection: Xi Zhao; analysis and interpretation of results: Xi Zhao, Haizheng Yu, Hong Bian; draft manuscript preparation: Xi Zhao, Haizheng Yu; All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are all open-source and can be downloaded by readers from https://www.kaggle.com/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Munich, Germany, pp. 1125–1134, 2017.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Science*, ArXiv preprint arXiv:1411.1784, 2014.
[3] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless and D. H. Salesin, "Image analogies," in *Proc. of the 28th Annual Conf. on Computer Graphics and Interactive Techniques*, New York, NY, USA, pp. 327–340, 2001.
[4] T. Chen, M. M. Cheng, P. Tan, A. Shamir and S. M. Hu, "Sketch2photo: Internet image montage," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 1–10, 2009.
[5] P. Y. Laffont, Z. Ren, X. Tao, C. Qian and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 149.1–149.11, 2014.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu and D. Warde-Farley, "Generative adversarial nets," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, USA, pp. 2672–2680, 2014.
[7] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional net works for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, pp. 234–241, 2015.
[8] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2223–2232, 2017.
[9] M. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. of Advances in Neural Information Processing Systems*, ArXiv preprint arXiv:1703.00848, 2017.
[10] X. Huang, M. Y. Liu, S. Belongie and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 172–189, 2018.
[11] Y. Choi, M. Choi, M. Kim, J. W. Hua, S. Kim *et al.,* "Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Munich, Germany, pp. 8789–8797, 2018.
[12] Y. Choi, M. Choi, M. Kim, J. W. Hua, S. Kim *et al.,* "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Munich, Germany, pp. 8188–8197, 2018.
[13] Q. Liu, H. Zhou and Q. Xu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10227–10242, 2021.
[14] H. A. Amirkolaee and H. A. Amirkolaee, "Medical image translation using an edge-guided generative adversarial network with global-to-local feature fusion," *The Journal of Biomedical Research*, vol. 36, no. 6, pp. 409–422, 2022.
[15] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–842, 2021.
[16] Z. Parekh, J. Baldridge, D. Cer, A. Waters and Y. F. Yang, "Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO," in *Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 2855–2870, 2021.
[17] T. H. Zhang, H. Y. Tseng, J. Lu, W. L. Yang, H. Lee *et al.,* "Text as neural operator: Image manipulation by text instruction," in *Proc. of the 29th ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 1893–1902, 2021.
[18] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert *et al.,* "BC-Z: Zero-shot task generalization with robotic imitation learning," in *Proc. of the 5th Conf. on Robot Learning*, vol. 164, pp. 991–1002, 2022.
[19] T. Park, M. Y. Liu, T. C. Wang and J. Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, pp. 2337–2346, 2019.
[20] S. Yan, C. Y. Wang, W. B. Chen and J. Lyu, "Swin transformer-based GAN for multi-modal medical image translation," *Frontiers in Oncology*, vol. 12, pp. 17, 2022. https://doi.org/10.3389/fonc.2022.942511
[21] B. Zhan, D. Li, X. Wu, J. Zhou and Y. Wang, "Multi-modal MRI image synthesis via GAN with multi-scale gate mergence," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 17–26, 2022.

[22] J. H. Liu, X. M. Song, Z. M. Chen and J. Ma, "MGCM: Multi-modal generative compatibility modeling for clothing matching," *Neurocomputing*, vol. 414, pp. 215–224, 2020.
[23] R. R. Nair, T. Singh, R. Sankar and K. Gunndu, "Multi-modal medical image fusion using LMF-GAN—A maximum parameter infusion technique," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 5, pp. 5375–5386, 2021.
[24] J. S. Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguil, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. of the 32nd Int. Conf. on Machine Learning*, vol. 37, pp. 2256–2265, 2015.
[25] J. Hoo, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
[26] Y. Takaigi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 14453–14463, 2023.
[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.,* "Learning transferable visual models from natural language supervision," in *Proc. of the 38th Int. Conf. on Machine Learning*, vol. 139, pp. 8748–8763, 2021.
[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: Rede signing skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
[29] T. Salimans, I. Goodfellow, W. Zareba, V. Cheung, A. Radford *et al.,* "Improved techniques for training GANs," in *Proc. of Advances in Neural Information Processing Systems (NIPS 2016)*, ArXiv preprint arXiv:1606.03498, 2016.
[30] 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 *Proc. of Advances in Neural Information Processing Systems (NIPS 2017)*, ArXiv preprint arXiv:1706.08500, 2017.
[31] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.