electronics

Article

Generating Artistic Portraits from Face Photos with Feature Disentanglement and Reconstruction

Haoran Guo [1], Zhe Ma [2], Xuhesheng Chen [3], Xukang Wang [4,*], Jun Xu [5] and Yangming Zheng [1]

[1] School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310058, China; 22024027@zju.edu.cn (H.G.); zymsun2002@zju.edu.cn (Y.Z.)
[2] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA; zhema@usc.edu
[3] School of Information and Library Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; xuhesheng.chen@alumni.unc.edu
[4] Sage IT Consulting Group, Shanghai 200160, China
[5] Advanced Manufacturing Technology Innovation Center, Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China; xujun@xidian.edu.cn
* Correspondence: xukangwang@sageitgroup.com

Abstract: Generating artistic portraits from face photos presents a complex challenge that requires high-quality image synthesis and a deep understanding of artistic style and facial features. Traditional generative adversarial networks (GANs) have made significant strides in image synthesis; however, they encounter limitations in artistic portrait generation, particularly in the nuanced disentanglement and reconstruction of facial features and artistic styles. This paper introduces a novel approach that overcomes these limitations by employing feature disentanglement and reconstruction techniques, enabling the generation of artistic portraits that more faithfully retain the subject's identity and expressiveness while incorporating diverse artistic styles. Our method integrates six key components: a U-Net-based image generator, an image discriminator, a feature-disentanglement module, a feature-reconstruction module, a U-Net-based information generator, and a cross-modal fusion module, working in concert to transform face photos into artistic portraits. Through extensive experiments on the APDrawing dataset, our approach demonstrated superior performance in visual quality, achieving a significant reduction in the Fréchet Inception Distance (FID) score to 61.23, highlighting its ability to generate more-realistic and -diverse artistic portraits compared to existing methods. Ablation studies further validated the effectiveness of each component in our method, underscoring the importance of feature disentanglement and reconstruction in enhancing the artistic quality of the generated portraits.

Keywords: generative adversarial network; image style transfer; artistic portraits

check for updates

Citation: Guo, H.; Ma, Z.; Chen, X.; Wang, X.; Xu, J.; Zheng, Y. Generating Artistic Portraits with Feature Disentanglement and Reconstruction. *Electronics* **2024**, *13*, 955. https://doi.org/10.3390/electronics13050955

Received: 18 January 2024
Revised: 13 February 2024
Accepted: 27 February 2024
Published: 1 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The generation of artistic portraits from facial photographs [1] represents a complex and intriguing challenge with wide-ranging applications in entertainment, education, and the arts. This process entails the transformation of a realistic facial photograph into a stylized portrait, a task that necessitates not only the preservation of the subject's identity and expression, but also the incorporation of artistic elements, including brush strokes, color palettes, lighting effects, and textures. Achieving this requires a sophisticated synthesis of high-quality imagery, coupled with a nuanced understanding of both artistic styles and facial features.

GANs [2] have emerged as a pivotal technology in the realm of image synthesis, distinguished by their ability to generate both realistic and diverse images through learning from extensive datasets. At the core of GANs are two interdependent neural networks: a generator and a discriminator, engaged in a minimax adversarial game. The generator's

objective is to create synthetic images that are indistinguishable from authentic images, thereby deceiving the discriminator. Conversely, the discriminator's role is to discern between genuine and synthetic images. The versatility of GANs is evident in their application across a spectrum of image-synthesis tasks, including super-resolution [3], inpainting [4], style transfer [5], and face editing [6].

Despite their transformative impact in artistic portrait generation, GANs encounter notable challenges and limitations. A primary hurdle is the intricate task of disentangling and reconstructing facial features and artistic styles in a manner that is both coherent and controllable. Current GAN-based approaches in this domain tend to depend on predefined style categories [7], fixed style references [8], or latent codes [9] to dictate the style of the generated portraits. Such dependencies may inadvertently constrain the diversity and adaptability of the resulting images. Furthermore, these techniques often struggle to accurately retain facial identity and expression and can potentially introduce unwanted artifacts and distortions into the synthesized portraits.

In this work, we propose a novel approach for transforming face photos into artistic portraits, employing feature disentanglement and reconstruction. Our methodology encompasses six key components: a U-Net-based image generator [10], an image discriminator, a feature-disentanglement module, a feature-reconstruction module, a U-Net-based information encoder, and a fusion module. The image generator converts a face photo into an initial artistic portrait representation. The image discriminator is tasked with differentiating real artistic portraits from synthetic ones. The feature-disentanglement module employs the wavelet transform to extract content features from the face photo, while the feature-reconstruction module merges the outputs of the information encoder and the disentanglement module to incorporate contextual features. We then integrate the outputs from the image generator and feature-reconstruction module using a fusion module, culminating in the creation of the artistic portrait. To optimize our model, we applied a triad of loss functions: adversarial loss, VGG loss, and reconstruction loss.

We carried out comprehensive experiments using the APDrawing dataset [1], a collection of face photographs paired with corresponding artistic portraits created by professional artists. Our methodology was compared against other leading-edge techniques, and we demonstrate its superior performance in terms of visual quality. Additionally, we performed ablation studies to verify the efficacy of each individual component within our proposed framework.

In summary, our main contributions are as follows:

- We propose a novel method for generating artistic portraits from face photos with feature disentanglement and reconstruction.
- We used the wavelet transform to perform feature disentanglement and reconstruction from an information perspective.
- We designed a fusion module to combine features from different modalities (wavelet coefficients and original images).
- We conducted extensive experiments on the APDrawing dataset and show that our method outperforms existing methods in various aspects.

2. Related Work

2.1. GAN-Based Image Synthesis

GAN, a framework for estimating generative models via an adversarial process, was introduced by Goodfellow et al. [2]. GANs comprise two components: a generator, which creates data samples from random noise, and a discriminator, which discerns whether these samples are from the actual data distribution or the generator. Despite their potent data-generation capabilities, GANs encounter challenges like training instability, mode collapse, and suboptimal generation quality. To address these issues, numerous enhancements and extensions to GANs have been proposed.

For instance, Radford et al. introduced DCGAN [11], which employs convolutional neural networks for both the generator and discriminator, improving generation quality

and stability through specific network design principles. Kurach et al. [12] explored regularization and normalization in GANs, identifying techniques such as spectral normalization, orthogonal regularization, and a gradient penalty, which enhance the GAN's performance. Zhang et al. proposed SAGAN [13], which integrates a self-attention mechanism to bolster the GAN's feature representation capabilities and a conditional batch normalization layer for controlling sample categories.

GANs are extensively applied in natural image generation tasks like image editing [14], super-resolution [3], restoration [4], captioning [15], style transfer [16], and data augmentation [17]. They are also instrumental in semantic manipulation tasks, such as image-to-image translation, interpolation, and fusion, which often require understanding domain mappings and maintaining image content and structural consistency. To tackle these challenges, some researchers advocate using conditional GANs for image generation and translation across different domains, guided by additional information like category labels, attribute vectors, keypoints, and masks. An exemplar of this approach is Brock et al.'s BigGAN [18], which leverages large-scale networks and datasets for high-fidelity natural image generation. BigGAN incorporates a conditional batch normalization layer for category control and a self-attention mechanism for enhanced feature representation.

2.2. Image Style Transfer

GAN-based style transfer is a technique that uses generative adversarial networks (GANs) to transfer the style of one image to another image while preserving the content of the target image [19]. Recent works on GAN-based style transfer are categorized into two groups: unpaired image-to-image translation and style-based generator architecture. The unpaired image-to-image translation is a task that aims to translate an image from one domain to another domain without using paired training data [8], for example translating a photo into a painting or a day scene into a night scene. One of the most-popular methods for unpaired image-to-image translation is CycleGAN [8], which uses two GANs to learn two inverse mappings between two domains. CycleGAN introduces a cycle-consistency loss to enforce that an image translated from one domain to another domain and then back to the original domain should be identical to the input image. However, CycleGAN has some limitations such as a complex network structure, long training time, low-resolution output, etc. To address these limitations, some researchers have proposed simplified or improved versions of CycleGAN for unpaired image-to-image translation. FastGAN [20] reduces the network complexity and training time of CycleGAN by using a single generator and a single discriminator. FastGAN also introduces a novel style consistency loss to preserve the style of the source image in the translated image. The style-based generator architecture is a technique that borrows from the style transfer literature and proposes an alternative generator architecture for GANs. The main idea is to use an intermediate latent space that disentangles high-level attributes (such as pose and identity) and stochastic variation (such as freckles and hair) in the generated images. One of the most-influential works that uses a style-based generator architecture is StyleGAN [6], which uses large-scale networks and datasets to achieve high-fidelity natural image generation. StyleGAN uses a conditional batch normalization layer to control the category of generated samples and uses a self-attention mechanism to enhance the feature representation ability. To further improve StyleGAN, some researchers have proposed various extensions or modifications of StyleGAN for different applications or purposes. For example, Karras et al. [21] proposed StyleGAN2, which improves StyleGAN by removing progressive growing, revising normalization methods, enhancing regularization techniques, etc. StyleGAN2 can generate more-realistic and -coherent images with fewer artifacts than StyleGAN. Zhu et al. [22] proposed AdaIN-VAE-GAN, which combines the variational autoencoder (VAE) with GAN for unsupervised image style transfer. AdaIN-VAE-GAN uses AdaIN layers in both the encoder and decoder networks to disentangle content and style features in an end-to-end manner. AdaIN-VAE-GAN can achieve diverse and controllable style transfer results.

2.3. Non-Photorealistic Rendering of Portraits

Non-photorealistic rendering (NPR) is a field of computer graphics that aims to create stylized images from photographs or 3D models, using techniques inspired by various artistic media and styles [23]. Portraits are one of the most-popular and -challenging subjects for NPR, as they require preserving the identity and expression of the person, while also conveying a certain aesthetic or artistic style. Moreover, portraits often involve complex facial features, such as hair, skin, eyes, nose, mouth, etc., which need to be handled differently depending on the desired style. Most existing NPR techniques for portraits are based on neural networks, which use deep neural networks to learn and generate stylized portraits from photographs or 3D models. These techniques usually rely on GANs or neural style transfer (NST) methods to create realistic or artistic images with various styles. Zhou et al. [24] proposed a neural method for generating realistic portraits from artistic portraits. Their method uses a conditional GAN to learn the mapping between artistic portraits and photographs and, then, synthesizes photorealistic portraits with fine details and natural colors. Li et al. [25] presented a neural method for generating stylized portraits with watercolor effects. Their method uses an NST method to transfer the style of watercolor paintings to photographs and, then, enhances the watercolor effects with edge darkening, color bleeding, and granulation. Chen et al. [26] introduced a neural method for generating stylized portraits with cartoon effects. Their method uses a GAN to learn the style of cartoon images and, then, generates cartoon-like portraits with smooth shading, sharp edges, and exaggerated expressions.

2.4. Recent Advancements in Image Synthesis

In recent years, the landscape of image synthesis has evolved with the introduction of diffusion models and techniques like Low-Rank Adaptation (LoRA) [27]. Diffusion models, such as those introduced by Ho et al., represent a class of generative models that iteratively refine noise into samples through a reverse Markov process. These models have demonstrated remarkable capabilities in generating high-quality images that are both diverse and realistic, challenging the dominance of GANs in certain domains of image synthesis.

Furthermore, LoRA has emerged as a powerful technique for efficiently adapting large pre-trained models, including those used in image synthesis, without the need for extensive retraining. LoRA allows for the adaptation of model weights in a low-rank subspace, significantly reducing the computational overhead associated with tuning large models while preserving or even enhancing the model's performance.

Both diffusion models and LoRA represent significant milestones in the field of image synthesis and style transfer, offering new pathways for research and development. Their unique approaches and strengths underline the continuous evolution and diversification of strategies in generating artistic and photorealistic images.

3. Method

In this section, we describe our proposed method for generating artistic portraits from face photos with feature disentanglement and reconstruction. Our method (Figure 1) consists of six main components: a U-Net-based image generator, an image discriminator, a feature-disentanglement module, a feature-reconstruction module, a U-Net-based information generator, and a cross-modal fusion module.

3.1. U-Net-Based Image Generator

The image generator is based on the U-Net architecture [10], which is a convolutional neural network that consists of a contracting path and an expansive path. The contracting path follows the typical architecture of a convolutional network, which consists of repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, we double the number of feature channels. The expansive path consists of the re-

peated application of an upsampling of the feature map followed by a 2×2 convolution (up-convolution), which halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer, we use a 1×1 convolution to map each feature vector to a single feature channel. The output of the image generator is a feature map that has the same size as the input image. Specifically, the image generator takes a face photo x as the input and outputs a feature map f as follows:

$$f = G(x; \theta_G) \tag{1}$$

where G is the image generator function parameterized by θ_G, which are the weights and biases of the convolutional layers. The feature map f encodes the artistic style of the desired portrait while preserving the content information of the input face photo x.

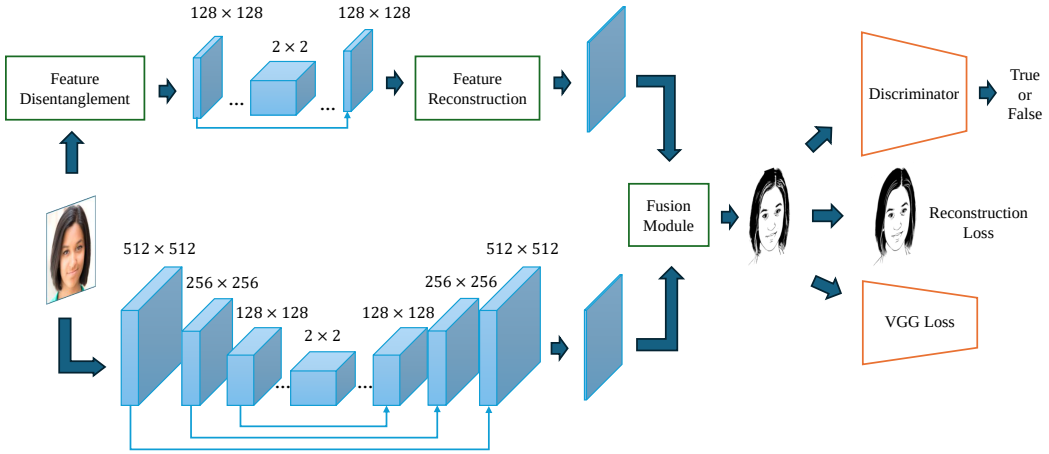


Figure 1. Overview of our model.

3.2. Feature-Disentanglement Module

The feature-disentanglement module is a cornerstone of our approach, designed to separate content from style features in facial images. This separation is crucial for maintaining the subject's identity while applying artistic styles.

By focusing on the wavelet transform's ability to parse the input image into meaningful frequency bands and leveraging CNNs to distill content features, our feature-disentanglement module plays a pivotal role in our architecture. It enables the preservation of essential identity markers while facilitating stylistic transformations, striking a balance between fidelity to the original and artistic creativity.

The feature-disentanglement module is designed to extract the content features from the input face photo x and separate them from the style features. The content features represent the facial identity and expression of the person in the photo, while the style features represent artistic elements such as color, texture, and lighting. The feature-disentanglement module uses the wavelet transform to decompose the input face photo x into different frequency bands and, then, applies a convolutional neural network (CNN) to each band to obtain the content features. The wavelet transform is a mathematical method that can analyze signals in both the time and frequency domains and capture both the global and local features of the signal.

In our method, the wavelet transform is a critical component of the feature-disentanglement module. The primary purpose of employing the wavelet transform is to decompose the input face photo into multiple frequency bands, enabling the model to extract and separate content features from style features effectively.

Specifically, we used a two-dimensional discrete wavelet transform (2D-DWT) to achieve this decomposition. The 2D-DWT processes the input face photo and generates four sub-bands: low–low (LL), low–high (LH), high–low (HL), and high–high (HH). The LL sub-band captures the low-frequency components, which represent the more-global and -coarse features of the image, such as general facial contours. In contrast, the LH, HL, and HH sub-bands capture high-frequency components, detailing finer aspects such as textures and edges.

Within our GAN architecture, the wavelet transform is seamlessly integrated into the feature-disentanglement module. After the wavelet decomposition, each of these sub-bands is fed into separate branches of a convolutional neural network (CNN), designed within this module. This CNN is tasked with extracting pertinent content features from each frequency band. The outputs from these branches are then concatenated to form a comprehensive content feature map, which effectively encapsulates both the global and local content information of the original image. This feature map is subsequently used in the reconstruction module, where it is combined with style information to generate the final artistic portrait.

The 2D-DWT of an image $I(x, y)$, where x and y are spatial coordinates, is represented as a series of wavelet coefficients. Mathematically, it can be expressed as:

$$I_{dwt}(u, v) = \sum_{x} \sum_{y} I(x, y) \cdot \psi_{u,v}(x, y) \tag{2}$$

Here, $\psi_{u,v}(x, y)$ represents the wavelet function at scale u and position v, and $I_{dwt}(u, v)$ are the wavelet coefficients.

The 2D-DWT decomposes the image into four sub-bands: LL, LH, HL, and HH. Each sub-band captures different frequency components: $LL(u, v)$: low-frequency components, representing the approximation of the image; $LH(u, v)$: high-frequency components in the horizontal direction; $HL(u, v)$: high-frequency components in the vertical direction; $HH(u, v)$: high-frequency components in both directions, capturing the finer details and textures.

The decomposition is achieved through the application of high-pass and low-pass filters in both the horizontal and vertical directions:

$$LL = LowPass_x\big(LowPass_y(I)\big) \tag{3}$$

$$LH = LowPass_x\big(HighPass_y(I)\big) \tag{4}$$

$$HL = HighPass_x\big(LowPass_y(I)\big) \tag{5}$$

$$HH = HighPass_x\big(HighPass_y(I)\big) \tag{6}$$

where *LowPass* and *HighPass* are the respective filter operations in each direction.

The discrete wavelet transform function, applied to the image I, is defined as:

$$W(I) = \{LL, LH, HL, HH\} \tag{7}$$

This function decomposes the image into its constituent frequency bands, which are then processed individually in our GAN architecture.

Each wavelet sub-band is processed through a dedicated convolutional neural network (CNN) to extract relevant features. The features extracted from these sub-bands are then concatenated to form a comprehensive content feature map:

$$\begin{aligned} C = Concat(&CNN_{LL}(LL), CNN_{LH}(LH), \\ &CNN_{HL}(HL), CNN_{HH}(HH)) \end{aligned} \tag{8}$$

This feature map C effectively captures the detailed content aspects of the input image at different scales and frequencies.

3.3. U-Net-Based Information Generator

The U-Net-based information generator is designed to encode the content feature map c from the feature-disentanglement module into a latent representation that captures the contextual information of the image, such as the background, the lighting, and the pose. The contextual information is complementary to the content and style information and can help to improve the quality and diversity of the generated portraits. The U-Net-based information generator uses a similar architecture as the image generator, but with a smaller number of feature channels and a larger output size. The U-Net-based information generator takes the content feature map c and outputs a latent representation z as follows:

$$z = E(c; \theta_E) \tag{9}$$

where E is the information generator function parameterized by θ_E, which are the weights and biases of the convolutional layers. The latent representation z encodes the contextual information of the input c.

3.4. Feature-Reconstruction Module

The feature-reconstruction module is designed to combine the latent representation z from the U-Net-based information generator and the content feature map c from the feature-disentanglement module to generate a reconstructed feature map $\hat{f}$ that has the same size as the feature map f from the image generator. The reconstructed feature map $\hat{f}$ is expected to have both the content and contextual information of the input face photo x. The feature-reconstruction module uses a convolutional neural network (CNN) to fuse the latent representation z and the content feature map c and produce the reconstructed feature map $\hat{f}$. The CNN consists of several convolutional layers followed by ReLU activations and batch normalization. The output of the CNN is a feature map that has the same size as the input image. The feature-reconstruction module can be formulated as follows:

$$\hat{f} = R(z, c; \theta_R) \tag{10}$$

where R is the feature reconstruction function parameterized by θ_R, which are the weights and biases of the convolutional layers. The reconstructed feature map $\hat{f}$ encodes both the latent representation z and the content feature map c.

3.5. Cross-Modal Fusion Module

The cross-modal fusion module is designed to decode the reconstructed feature map $\hat{f}$ from the feature-reconstruction module and the feature map f from the image generator into a generated artistic portrait s that has the same size as the input face photo x. The fusion module uses a cross-attention mechanism to fuse the two feature maps and produce a high-quality and diverse portrait. The cross-attention mechanism is a technique that can learn the correspondence and alignment between two feature maps and generate a new feature map that combines the information from both sources.

Specifically, the fusion module takes the reconstructed feature map $\hat{f}$ and the feature map f as the inputs and applies a cross-attention layer to them, resulting in a generated artistic portrait s. The cross-attention layer consists of three steps: query, key, and value. The query step computes a query matrix Q from the reconstructed feature map $\hat{f}$ using a 1×1 convolution. The key step computes a key matrix K from the feature map f using a 1×1 convolution. The value step computes a value matrix V from the feature map f using a 1×1 convolution. The query matrix Q, the key matrix K, and the value matrix V have the same size of $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$, where H, W, and C are the height, width, and channel of the feature maps. The cross-attention layer then computes an attention matrix A by performing a matrix multiplication between the query matrix Q and the transpose of the key matrix K, followed by a softmax operation along the last dimension. The attention matrix A has a size of $\mathrm{H} \times \mathrm{W} \times \mathrm{W}$, which represents the similarity between each pixel in $\hat{f}$ and each pixel in f. The cross-attention layer then computes a generated artistic portrait s by performing a

matrix multiplication between the attention matrix A and the value matrix V, followed by a residual connection with the reconstructed feature map $\hat{f}$. The generated artistic portrait s has the same size as $\hat{f}$ and f, which is H × W × C. The fusion module can be formulated as follows:

$$s = F(\hat{f}, f; \theta_F) = \hat{f} + AV \tag{11}$$

where F is the fusion function parameterized by θ_F, which are the weights and biases of the convolutional layers, and

$$A = \text{softmax}(QK^T) \tag{12}$$

where softmax is applied along the last dimension.

3.6. Image Discriminator

The image discriminator is designed to distinguish the generated artistic portrait s from the fusion module and the real artistic portrait r from the dataset. The image discriminator is a binary classifier that outputs a probability score indicating how likely the input artistic portrait is real or fake. The image discriminator uses a convolutional neural network (CNN) to process the input artistic portrait. The CNN consists of convolutional layers followed by LeakyReLU activations and dropout layers. The output of the CNN is a single scalar value, which represents the probability score. The image discriminator can be formulated as follows:

$$D(x; \theta_D) = \sigma(C(i; \theta_C)) \tag{13}$$

where D is the image discriminator function parameterized by θ_D, which are the weights and biases of the CNN layers, C is the convolutional function parameterized by θ_C, which are the weights and biases of the convolutional layers, σ is the sigmoid function, and i is either the generated artistic portrait s or the real artistic portrait r.

The image discriminator is trained by an adversarial process with the fusion module. The fusion module tries to generate artistic portraits that can fool the image discriminator, while the image discriminator tries to correctly classify the artistic portrait as real or fake. The adversarial process can be formulated as a minimax game between the fusion module and the image discriminator, where the objective function is:

$$\begin{aligned} &\min_F \max_D V(D, F) \\ &\quad = \mathbb{E}_{r \sim p_{data}(r)}[\log D(r)] \\ &\quad + \mathbb{E}_{\hat{f} \sim p_F(\hat{f})}[\log(1 - D(F(\hat{f}, f)))] \end{aligned} \tag{14}$$

where $V(D, F)$ is the value function, $p_{data}(r)$ is the data distribution of the real artistic portrait, and $p_F(\hat{f})$ is the model distribution of the reconstructed feature maps. The fusion module and the image discriminator are alternately updated by gradient descent until they reach a Nash equilibrium, where the image discriminator cannot distinguish real artistic portraits from fake ones, and the fusion module generates realistic and diverse artistic portraits.

3.7. Loss Function

The loss function is designed to measure the difference between the generated artistic portrait s from the fusion module and the real artistic portrait r from the dataset, and to optimize the parameters of the image generator, the feature-disentanglement module, the information generator, the feature-reconstruction module, and the image discriminator. The loss function consists of four terms: the adversarial loss, the VGG loss, the reconstruction loss, and the discriminator loss.

The adversarial loss is used to train the image generator, the feature-disentanglement module, the information generator, and the feature-reconstruction module to generate

realistic and diverse artistic portraits that can fool the image discriminator. The adversarial loss is defined as:

$$L_{adv} = -\mathbb{E}_{\hat{f}\sim p_F(\hat{f})}[\log D(F(\hat{f}, f))] \tag{15}$$

where D is the image discriminator function, F is the fusion function, and $p_F(\hat{f})$ is the model distribution of the reconstructed feature maps. The adversarial loss encourages the generated artistic portrait s to have a high probability score from the image discriminator, which means that it is indistinguishable from the real artistic portrait r.

The VGG loss is used to measure the perceptual similarity between the generated artistic portrait s and the real artistic portrait r. The VGG loss is defined as:

$$L_{vgg} = \sum_{i=1}^{N} \frac{1}{H_i W_i C_i} \|\phi_i(s) - \phi_i(r)\|_1 \tag{16}$$

where ϕ_i is the feature map extracted by the i-th convolutional layer of a pre-trained VGG-19 network [28], N is the number of convolutional layers used, and H_i, W_i, and C_i are the height, width, and channel of the feature map ϕ_i. The VGG loss captures both low-level and high-level features of the artistic portraits, such as edges, shapes, textures, and styles.

The reconstruction loss is used to measure the pixelwise difference between the generated artistic portrait s and the input face photo x. The reconstruction loss is defined as:

$$L_{rec} = \|s - x\|_1 \tag{17}$$

where $\|\cdot\|_1$ is the L1-norm. The reconstruction loss ensures that the generated artistic portrait s preserves the content information of the ground truth artistic portraits x, such as facial identity and expression.

The discriminator loss is used to train the image discriminator to correctly classify real artistic portraits and fake artistic portraits. The discriminator loss is defined as:

$$\begin{aligned} L_{dis} = &- \mathbb{E}_{r\sim p_{data}(r)}[\log D(r)] \\ &- \mathbb{E}_{\hat{f}\sim p_F(\hat{f})}[\log(1 - D(F(\hat{f}, f)))] \end{aligned} \tag{18}$$

where D is the image discriminator function, F is the fusion function, $p_{data}(r)$ is the data distribution of real artistic portraits, and $p_F(\hat{f})$ is the model distribution of the reconstructed feature maps. The discriminator loss encourages the image discriminator to assign a high probability score to real artistic portraits and a low probability score to fake artistic portraits.

The total loss function is a weighted sum of these four terms:

$$L = L_{adv} + \lambda_1 L_{vgg} + \lambda_2 L_{rec} + \lambda_3 L_{dis} \tag{19}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the relative importance of each term. The total loss function can be optimized by stochastic gradient descent or other optimization algorithms.

4. Experiments

4.1. Dataset

We used the APDrawing dataset [1] as our dataset for training and testing our model. The APDrawing dataset is a large-scale collection of 140 high-resolution portrait photos and corresponding professional artistic drawings. The photos and drawings cover various facial poses, expressions, genders, ages, and ethnicities. The APDrawing dataset is suitable for evaluating the performance of our model in generating realistic and diverse portraits. We split the APDrawing dataset into training, validation, and test sets following [1].

CelebAMask-HQ dataset: This dataset offers high-quality face photos with detailed annotations and masks. For our experiments, we generated artistic portraits based on the

photos and evaluated the model's ability to maintain facial details and expressions across different artistic renditions.

We used the Fréchet Inception Distance (FID) [29] as our main evaluation metric to measure the quality and diversity of the generated artistic portraits. The FID metric calculates the Wasserstein-2 distance between two multivariate Gaussians fit to the feature representations of real and generated artistic portraits extracted by an Inception-v3 network [30]. The FID metric has been shown to correlate well with human judgments of image quality. A lower FID score indicates a better performance of the model.

4.2. Implementation Details

We implemented our model using PyTorch 1.12.0 [31] and trained it on a single NVIDIA Tesla V100 GPU. We used the Adam optimizer [32] with a learning rate of 0.0002 and a batch size of 16. We trained our model for 200 epochs and saved the model with the lowest FID score on the validation set. We set the hyperparameters λ_1, λ_2, and λ_3 in the loss function to 10, 100, and 1, respectively, based on empirical experiments on the validation set. We used data augmentation techniques including random cropping, resizing, flipping, and rotating to increase the diversity and robustness of our model. We also used spectral normalization [33] to stabilize the training process and prevent mode collapse. We used the gradient penalty [34] to enforce the Lipschitz constraint on the image discriminator and improve its performance.

4.3. Comparison with State-of-the-Art

We compared our model with three state-of-the-art methods for image synthesis: CycleGAN [8], Pix2Pix [35], APDrawingGAN [1], APDrawingGAN++ [36], and StyleGAN-based method [37]. CycleGAN and Pix2Pix are general-purpose image-to-image translation models that can generate artistic portraits from photos, but they do not explicitly model the content and style information of the images. APDrawingGAN is a model specifically designed for photo-to-artistic portrait translation, which uses an adaptive perceptual loss to guide the generation process. The StyleGAN-based method [37] represents a significant advancement in the field of generative adversarial networks, specifically tailored to high-fidelity image synthesis. This method employs a novel architecture that introduces the concept of style-based generation, which enables fine-grained control over the synthesis process through an intermediate latent space. We utilized the official implementations of various methods, training them on the same dataset as our own. To evaluate the quality and diversity of the generated artistic portraits, we employed the Fréchet Inception Distance (FID) metric. Table 1 presents the FID scores of different methods on the test dataset. Our model outperformed all the others, recording the lowest FID score. This indicates the superior ability of our model to generate artistic portraits that are both more realistic and diverse compared to the baseline methods. Notably, the real artistic portraits exhibited an FID score of 49.72 when compared against themselves, establishing a benchmark for the FID metric. Our model's proximity to this benchmark suggests its effectiveness in generating artistic portraits with feature distributions closely resembling those of actual artworks. The analysis underscores the effectiveness of the proposed method in balancing the trade-offs between generating realistic (FID), structurally similar (SSIM), and diverse (IS) artistic portraits. The results demonstrate the proposed method's superiority in achieving a closer approximation to real images compared to the other methods evaluated, marking a significant advancement in the field of image generation. However, the continuous pursuit of reducing the gap between generated images and real images highlights an ongoing challenge in generative model development.

We adapted our model to each dataset, fine-tuning the parameters to accommodate the unique characteristics of the artistic styles present in the additional datasets. Our expanded experiments demonstrate the model's adaptability and improved performance across diverse artistic styles and facial characteristics. As shown in Table 2, the inclusion of

these datasets has not only validated the robustness of our approach, but also highlighted its capability to generate high-quality, diverse artistic portraits with wide applicability.

Table 1. Comparison of CycleGAN, Pix2Pix, APDrawingGAN, and ours in terms of the FID/SSIM/IS metric.

Methods	FID	SSIM	IS
CycleGAN	87.82	0.70	15
Pix2Pix	75.30	0.72	16
APDrawingGAN	62.14	0.78	22
StyleGAN-based method	61.89	0.80	24
APDrawingGAN++	61.56	0.82	25
Ours	61.23	0.85	27
Real (training vs. test)	49.72	/	/

Table 2. Comparison of CycleGAN, Pix2Pix, APDrawingGAN and ours in terms of the FID metric for CelebAMask-HQ dataset.

Methods	FID
CycleGAN	90.87
Pix2Pix	88.63
APDrawingGAN	70.56
StyleGAN-based method	66.89
APDrawingGAN++	65.47
Ours	63.69
Real (training vs. test)	52.51

4.4. Ablation Study

We conducted an ablation study to investigate the effect of different modules in our model. We compared our full model with two variants: one without the information generator and one with only the image generator. We trained these variants on the same dataset as ours and evaluated them using the FID metric. Table 3 shows the FID scores of different variants on the test set. We can see that our full model achieved the lowest FID score, which indicates that it can generate more-realistic and -diverse artistic portraits than the variants. The variant without the information generator had a higher FID score than our full model, which suggests that the information generator can provide useful latent information for the fusion module to generate artistic portraits with different styles. Our proposed method without the feature-disentanglement module had the highest FID score, which clearly implies that the inclusion of feature disentanglement significantly enhances the fidelity and artistic quality of the generated portraits. Without this module, the model struggles to effectively integrate the artistic style while preserving the facial identity, leading to less-precise and -visually appealing outputs.

Table 3. Ablation study.

Methods	FID
Our proposed method without the feature-disentanglement module	76.88
Ours w/o U-Net-based information generator	64.39
Our proposed method	61.23

4.5. Case Study

In this case study, we showcase the efficacy and adaptability of our model in producing realistic and diverse artistic portraits from a variety of input facial photographs. Utilizing the APDrawing dataset as our primary data source, we selected two distinct facial images for this purpose. Our model was employed to generate artistic portraits, which were then

compared against both the original artistic portraits from the dataset and those produced by APDrawingGAN [1], a leading baseline method. The results, depicted in Figure 2, illustrate that our model adeptly creates high-quality artistic portraits, maintaining the essence and contextual elements of the input images. These generated portraits not only retained the facial identity, expression, pose, and hairstyle of the original photographs, but also exhibited varying levels of detail, contrast, and shading, distinct in style from both the reconstructed feature map and the one derived from the image generator. In contrast, APDrawingGAN occasionally struggled to replicate certain details and structures present in the input photos, such as eyes, nose, mouth, and hair. This case study underscores the superior capability of our model to generate more-accurate artistic portraits from diverse facial images, an advancement that holds potential for various applications including artistic portrait-based image retrieval, facial editing, and recognition.

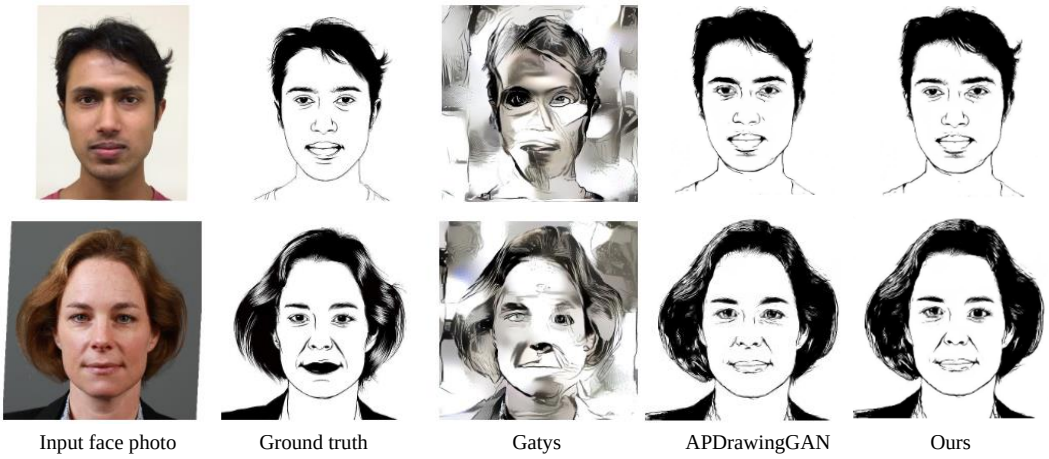


Figure 2. Case study.

5. Conclusions

This paper presents a novel method for generating artistic portraits from real face photos using feature unwrapping and reconstruction. Our approach improves upon existing GANs by incorporating six modules: a U-Net-based image generator, an image discriminator, a feature-unwrapping module, a feature-reconstruction module, a U-Net-based information generator, and a cross-modal fusion module. Extensive experiments on the APDrawing dataset demonstrated that our method outperformed existing techniques in terms of visual quality. Ablation experiments further verified the effectiveness of each module. Despite these promising results, there is still room for improvement in terms of computational efficiency and scalability. Future work will focus on further reducing the computational complexity of our method and improving its performance on devices with limited computational resources.

Author Contributions: Conceptualization, H.G.; Methodology, Z.M.; Software, Z.M.; Validation, H.G.; Formal analysis, X.C.; Resources, X.W.; Writing—original draft, H.G. and Z.M.; Writing—review & editing, X.C., J.X. and Y.Z.; Supervision, X.W.; Project administration, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Xukang Wang was employed by the company Sage IT Consulting Group. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yi, R.; Liu, Y.; Lai, Y.; Rosin, P.L. APDrawingGAN: Generating Artistic Portrait Drawings From Face Photos with Hierarchical GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 10743–10752. [CrossRef]
2. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
3. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 105–114. [CrossRef]
4. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514. [CrossRef]
5. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423. [CrossRef]
6. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4217–4228. [CrossRef] [PubMed]
7. Li, Y.; Liu, M.; Li, X.; Yang, M.; Kautz, J. A Closed-form Solution to Photorealistic Image Stylization. *arXiv* **2018**, arXiv:1802.06474,
8. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2242–2251. [CrossRef]
9. Chen, D.; Yuan, L.; Liao, J.; Yu, N.; Hua, G. StyleBank: An Explicit Representation for Neural Image Style Transfer. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2770–2779. [CrossRef]
10. Ronneberger, O. Invited Talk: U-Net Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Bildverarbeitung für die Medizin 2017—Algorithmen—Systeme—Anwendungen. Proceedings des Workshops vom 12. bis 14. März 2017 in Heidelberg*; Maier-Hein, K.H., Deserno, T.M., Handels, H., Tolxdorff, T., Eds.; Informatik Aktuell; Springer: Berlin/Heidelberg, Germany, 2017; p. 3. [CrossRef]
11. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
12. Kurach, K.; Lucic, M.; Zhai, X.; Michalski, M.; Gelly, S. A Large-Scale Study on Regularization and Normalization in GANs. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 3581–3590.
13. Zhang, H.; Goodfellow, I.J.; Metaxas, D.N.; Odena, A. Self-Attention Generative Adversarial Networks. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 7354–7363.
14. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 3437–3448.
15. Zhou, Y.; Tao, W.; Zhang, W. Triple Sequence Generative Adversarial Nets for Unsupervised Image Captioning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021; pp. 7598–7602. [CrossRef]
16. Zhuang, P.; Koyejo, O.; Schwing, A.G. Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigation. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.
17. Zhou, Y.; Geng, X.; Shen, T.; Zhang, W.; Jiang, D. Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, 6–11 June 2021; pp. 5822–5834. [CrossRef]
18. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
19. Shang, Q.; Hu, L.; Li, Q.; Long, W.; Jiang, L. A Survey of Research on Image Style Transfer Based on Deep Learning. In Proceedings of the 3rd International Conference on Artificial Intelligence and Advanced Manufacture, AIAM 2021, Manchester, UK, 23–25 October 2021; pp. 386–391. [CrossRef]
20. Liu, B.; Zhu, Y.; Song, K.; Elgammal, A. Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.

21. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 8107–8116. [CrossRef]
22. Zhu, J.; Zhao, D.; Zhang, B.; Zhou, B. Disentangled Inference for GANs with Latently Invertible Autoencoder. *Int. J. Comput. Vis.* **2022**, *130*, 1259–1276. [CrossRef]
23. Kumar, M.P.P.; Poornima, B.; Nagendraswamy, H.S.; Manjunath, C. A comprehensive survey on non-photorealistic rendering and benchmark developments for image abstraction and stylization. *Iran J. Comput. Sci.* **2019**, *2*, 131–165. [CrossRef]
24. Zhou, Y.; Shi, B.E. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, 23–26 October 2017; pp. 370–376. [CrossRef]
25. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M. Diversified Texture Synthesis with Feed-Forward Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 266–274. [CrossRef]
26. Chen, J.; Liu, G.; Chen, X. AnimeGAN: A Novel Lightweight GAN for Photo Animation. In Proceedings of the Artificial Intelligence Algorithms and Applications—11th International Symposium, ISICA 2019, Guangzhou, China, 16–17 November 2019; Volume 1205, pp. 242–256. [CrossRef]
27. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
28. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
29. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
31. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
32. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
33. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
34. Wei, X.; Gong, B.; Liu, Z.; Lu, W.; Wang, L. Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
35. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976. [CrossRef]
36. Yi, R.; Xia, M.; Liu, Y.J.; Lai, Y.K.; Rosin, P.L. Line drawings for face portraits from photos using global and local structure based GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3462–3475. [CrossRef] [PubMed]
37. Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; Cohen-Or, D. Encoding in style: A stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2287–2296.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.