electronics

Article

Research on High-Resolution Face Image Inpainting Method Based on StyleGAN

Libo He [1], Zhenping Qiang [2,*], Xiaofeng Shao [2], Hong Lin [2], Meijiao Wang [1] and Fei Dai [2]

[1] Information Security College, Yunnan Police College, Kunming 650221, China; LiboHe2022@gmail.com (L.H.); meijiaow@gmail.com (M.W.)
[2] College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China; shaoxf1996@gmail.com (X.S.); linh1226@swfu.edu.cn (H.L.); daifei@swfu.edu.cn (F.D.)
* Correspondence: qzp@swfu.edu.cn

Abstract: In face image recognition and other related applications, incomplete facial imagery due to obscuring factors during acquisition represents an issue that requires solving. Aimed at tackling this issue, the research surrounding face image completion has become an important topic in the field of image processing. Face image completion methods require the capability of capturing the semantics of facial expression. A deep learning network has been widely shown to bear this ability. However, for high-resolution face image completion, the network training of high-resolution image inpainting is difficult to converge, thus rendering high-resolution face image completion a difficult problem. Based on the study of the deep learning model of high-resolution face image generation, this paper proposes a high-resolution face inpainting method. First, our method extracts the latent vector of the face image to be repaired through ResNet, then inputs the latent vector to the pre-trained StyleGAN model to generate the face image. Next, it calculates the loss between the known part of the face image to be repaired and the corresponding part of the generated face imagery. Afterward, the latent vector is cut to generate a new face image iteratively until the number of iterations is reached. Finally, the Poisson fusion method is employed to process the last generated face image and the face image to be repaired in order to eliminate the difference in boundary color information of the repaired image. Through the comparison and analysis between two classical face completion methods in recent years on the CelebA-HQ data set, we discovered our method can achieve better completion results of $256 * 256$ resolution face image completion. For $1024 * 1024$ resolution face image restoration, we have also conducted a large number of experiments, which prove the effectiveness of our method. Our method can obtain a variety of repair results by editing the latent vector. In addition, our method can be successfully applied to face image editing, face image watermark clearing and other applications without the network training process of different masks in these applications.

Keywords: face completion; high-resolution face image completion; generative adversarial network; StyleGAN

check for updates

Citation: He, L.; Qiang, Z.; Shao, X.; Lin, H.; Wang, M.; Dai, F. Research on High-Resolution Face Image Inpainting Method Based on StyleGAN. *Electronics* **2022**, *11*, 1620. https://doi.org/10.3390/electronics11101620

Academic Editor: Adam Glowacz

Received: 31 March 2022
Accepted: 16 May 2022
Published: 19 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

Image completion, as an import image editing operation, aims to fill the missing area of an image in a semantically reasonable manner, bearing true fidelity to the original pixels of the real image. The traditional image completion methods which are based on diffusion [1] or based on matching [2] are effective in texture inpainting. However, they are inadequate for face completion with high semantics.

In recent years, with the rapid development of deep learning, the emergence of image inpainting methods based on feature learning have just managed to compensate for the shortcomings of traditional image inpainting methods. On the one hand, feature learning-based methods are based on Convolutional Neural Networks (CNN), which have been proven to bear the ability to extract features and semantic information from images,

making them effective tools for image inpainting. On the other hand, through extensive research on the Generative Adversarial Network (GAN) [3], it has been confirmed to be capable of generating certain types of imagery with rich detailed information, thus providing a new concept of the inpainting of specific types of images. For example, Zhao et al. [4] proposed a method based on GAN and achieved convincing results for 3D face reconstruction from a single color image under occluded scenes. Pathak et al. [5] first proposed an image inpainting algorithm based on unsupervised visual feature learning based on contextual pixel prediction using CNN. They used adversarial loss to maximize the realism of the completed image. Based on this algorithm, Yeh et al. [6] applied DCGAN (Deep Convolutional GAN) [7] to complete images, and their method can generate and fill the missing area of images effectively. In addition, their method is suitable for masks of any shape. Iizuka et al. [8] proposed a new creative generation model, which contains a generator, a local discriminator and a global discriminator to ensure the local consistency and global consistency of the in-painted image. Based on this model, Li et al. [9] proposed another generational model, which combined reconstruction loss, two adversarial losses and semantic parsing loss to further ensure the authenticity of synthetic pixels and the consistency of local and global content. Yu et al. [10] also proposed a coarse to fine network structure based on the model proposed by Iizuka et al. [8], in which a context attention layer was added to copy or obtain feature information about the background to generate real pixels. Zeng et al. [11] proposed a progressive inpainting structure, which is constructed on the basis of the U-Net [12] structure. The structure allowed for encoding the context semantics of full resolution input, and decoding the learned semantic features back to the image to realize image completion. Further, for the general convolution process, indiscriminate convolution is performed on the effective pixel area and the pixel area to be inpainted, which usually leads to artifacts such as chromatic aberration and blurring in the final result. In [13,14], they applied partial convolutions and gated convolution to complete images. Yang et al. [15] proposed a generative landmark-guided face inpainting method, which use landmarks of the face predicted by the information about the unobstructed area to inpaint face imagery.

Although these methods have achieved good results in face image completion, the use of high-resolution face data sets makes network training difficult and the generation of credible high-resolution images challenging, leading to suboptimal repair of high-resolution face images. Karras et al. [16] proposed StyleGAN, which can generate $1024 * 1024$ high-resolution face images gradually.

Inspired by StyleGAN, we propose a high resolution face inpainting method. We first use ResNet(Deep Residual Network) [17] to predict the latent vector of the face image, then generate the face image through the StyleGAN model, and finally use the Poisson mixing method [18] to fuse the generated image with the original image. This method can repair masks of any shape and size. Figure 1 displays some examples of the repair results of our method.

The contributions of this paper are as follows:

1. We propose an effective high-resolution face image completion method based on StyleGAN, which solves the problem of the high-resolution image completion network model being difficult to converge in training.

2. In our method, a model is designed to complete face image completion, face image editing, face-image watermark clearing and other image processes without the network training process of different repair masks.

3. In our model, we adopt a new latent vector processing method to improve the quality of repaired images and obtain diverse repair results.

(a) Original images **(b) Masked images** **(c) Inpainted images**

Figure 1. Examples of the repair results of our method.

2. Related Work

2.1. ResNet

A convolutional neural network or a fully connected network may have problems such as information loss during information transmission and issues of gradient disappearance or gradient explosion, which render the deep network untrainable. ResNet addresses this issue to a certain extent by directly bypassing the input information to the output to protect information integrity and simplifying the learning objectives and difficulties. Therefore, ResNet can stack more layers to improve the performance of the network and is generally used for classification tasks. We use ResNet to predict latent vectors through migration learning and the shape size of our output meets the size requirement of the input shape of the StyleGAN.

2.2. StyleGAN

With the proposal of GAN, the authenticity of the generated pictures continues to grow. Karras et al. [19] proposed the PGGAN (Progressive Growing GAN) to improve the quality, stability and variation of the generated images. The training of this model begins from a generator (G) and discriminator (D) of $4 * 4$ pixel resolution. As training progresses, layers are gradually added to G and D to improve spatial resolution of the generated image until a $1024 * 1024$ high-resolution image has been generated. Based on PGGAN, and inspired by arbitrary style transfer [20], Karras et al. [16] proposed another generator architecture StyleGAN. This generator can automatically learn and unsupervised separate the random changes in advanced attributes (such as posture and identity of the training face) and generated images (such as freckles and hair). It can intuitively compose these changes in a specified proportion, thus the generator has excellent image generation ability. StyleGAN mainly makes the following changes to PGGAN: (1) Removes the traditional input and takes a learnable constant as the initial input of the generator. (2) A mapping network is added to encode the input vector into an intermediate vector, then different elements of the intermediate vector w can be used to control different visual features. (3) Adds style modules, the output w of the mapping network is transformed into translation and scaling factors through a learnable affine transformation A (which is a full connection layer), and then each normalized spatial feature map is scaled and translated through the AdaIN

(Adaptive Instance Normalization) module to control the different visual features. (4) Adds noise to generate random details for the generator to increase the realism of the generated images.

GAN can generate an image from a random vector, so the question arises whether the feature code corresponding to this image can be extracted from a picture. LIPTON Z. C. et al. [21] proposed a method to rebuild latent vectors. This method initializes a vector with the same shape as the input random vector and generates a picture through this vector. It then updates the vector with random gradient descent and random clipping, so that the generated picture gradually approaches the original picture and eventually obtains the latent vector of the picture. Then the latent vector is used to generate the image that is similar to the original image.

Based on the idea of this method, we use ResNet to predict the latent vector of the damaged image, and use StyleGAN to generate images similar to the damaged image to complete image inpainting.

3. The Approach

In this section, we describe the proposed method of high-resolution face image completion. Given a masked face image, our goal is to synthesize the missing content, which is not only semantically consistent with the whole image, but also visually reasonable and real. Figure 2 shows the procedure of our method.

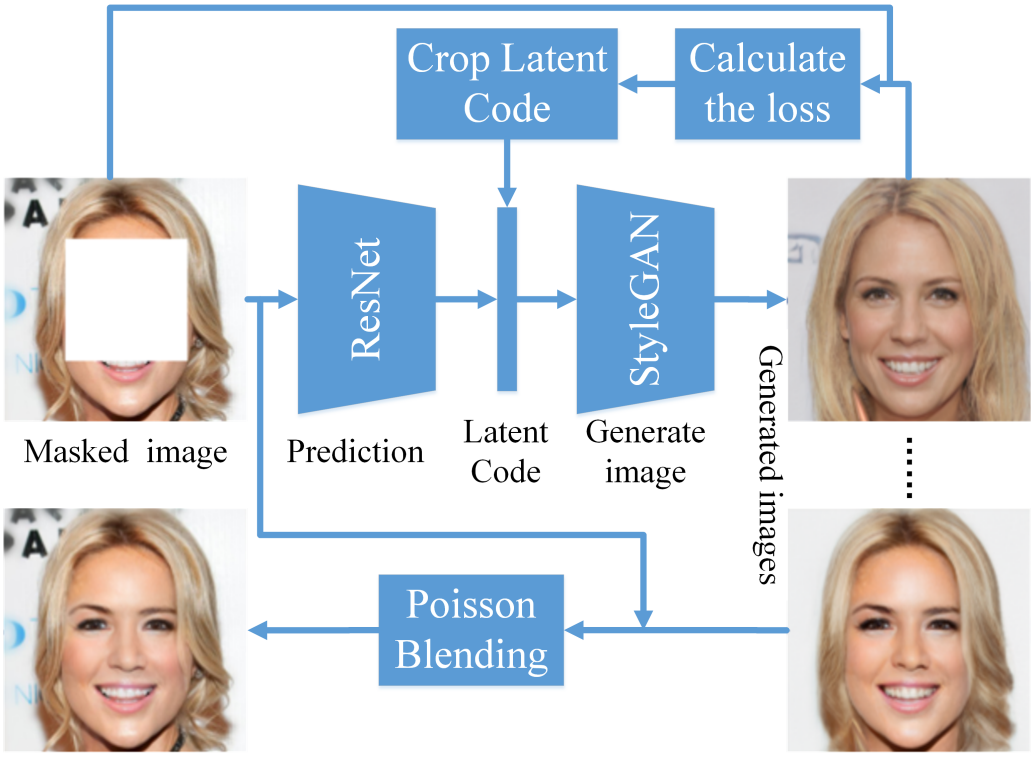


Figure 2. Network structure diagram of our method.

Each image generated by StyleGAN is generated by a random latent vector. We assume that each picture corresponds to a latent vector. If we find the latent vector corresponding to a picture, we can complete the original image through it. The method behind extraction of the latent vector of an image is an important issue to be resolved. Different from the method proposed by LIPTON Z. C. et al. [21], we first employ ResNet50 to predict the latent vector of the image through migration learning, then use the SytleGAN to generate a face image. Finally, we use the corresponding loss functions to generate the inpainting image iteratively. Throughout the entire process, the training data of ResNet50 are directly generated by StyleGAN. This accelerates the process of generating images and makes

the generated image more realistic than an image constructed using only StyleGAN. Our comparison results are shown in Figure 3. In the loss calculation, we only calculate the undamaged area of the input image, that is:

$$X = X \odot (1 - M). \tag{1}$$

where X is an input image, M is a mask image to label the missing area (the value of missing area is 0, the other area is 1), and $\odot$ means element by element multiplication.

(a) Masked images **(b) Without ResNet prediction** **(c) With ResNet prediction** **(d) Original images**

Figure 3. Comparison of completion results with or without ResNet prediction.

After determining the network structure, the loss function bears a great impact on the final results. We analyze the final inpainting results for different loss functions. Specifically, we choose the following loss functions to fine-tune the repaired image: VGG Loss, L_2 Loss, *Log-Cosh* Loss, $SSIM$ Loss, *MS-SSIM* Loss, $LPIPS$ Loss.

(1) VGG Loss

We use VGGNet [22] to extract the feature map of the original image, then use the L_1 distance between the feature map of the original image and the feature map of the generated image as VGG loss. The VGG loss is defined as:

$$Loss_{VGG} = \|V(G(X)) - V(X)\|_1 \tag{2}$$

where $V()$ represents StyleGAN generator, $V()$ represents extracting feature map by VGG.

(2) L_2 Loss

L_2 Loss is defined as the L_2 distance between the original image and the generated image:

$$Loss_{L_2} = \|G(X) - X\|_2. \tag{3}$$

(3) *Log-Cosh* Loss

Log-Cosh Loss is defined as logarithm of hyperbolic cosine of prediction error of original image and generated image:

$$Loss_{Log\text{-}Cosh} = \sum_{i=1}^{n} \sum_{j=1}^{m} log(cosh(X_{i,j} - G(X)_{i,j})) \tag{4}$$

where n and m represent the width and height of the image to be inpainted, respectively.

(4) $SSIM$ and $MSSIM$ Loss

$SSIM$ (Structural SIMilarity) [23], as a measure of similarity between two images, is compared with three aspects: brightness, contrast and structure similarity. *MS-SSIM*

(MultiScale Structure SIMilarity) [24] is more flexible than the single-scale method under different observation conditions. $SSIM$ is defined as:

$$SSIM(X, G(X)) = [l(X, G(X)]^{\alpha} \cdot [c(X, G(X)]^{\beta} \cdot [s(X, G(X)]^{\gamma}. \quad (5)$$

where α, β, γ are used to adjust the weight of each component and they are generally set as 1; $l(X, G(X)) = \frac{2\mu_X\mu_{G(X)}+C_1}{\mu_X^2+\mu_{G(X)}^2+C_1}$ uses the means (μ_X and $\mu_{G(X)}$) of X and $G(X)$ to estimate the similarity of luminance; $c(X, G(X)) = \frac{2\sigma_X\sigma_{G(X)}+C_2}{\sigma_X^2+\sigma_{G(X)}^2+C_1}$ uses the variances (σ_X and $\sigma_{G(X)}$) of X and $G(X)$ to estimate the similarity of contrast; $s(X, G(X)) = \frac{\sigma_{X,G(X)}+C_3}{\sigma_X\sigma_{G(X)}}$ uses the co-variance ($\sigma_{X,G(X)}$) to estimate the similarity of structure. $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$ and $C_3 = C_2/2$ are constants used to maintain stability, where L is the dynamic range of the pixel value in the images. According to past experience, k_1 is set as 0.01 and k_2 is set as 0.03.

The calculation of $MS\text{-}SSIM$ is to take the original image X and the generated image $G(X)$ as input, then iteratively employ a low-pass filter and 1/2 down-sampling to build image sets of different scales for both X and $G(X)$. Assuming that the original image scale is 1, and the highest scale is M, which is obtained after M-1 iterations. For the *j-th* scale, only the similarity between contrast $c(X, G(X))$ and the similarity of structure $s(X, G(X))$ are calculated. The similarity of luminance $l(X, G(X))$ is calculated only at scale M. $MS\text{-}SSIM$ is defined as:

$$MS\text{-}SSIM(X, G(X)) = [l(X, G(X))]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(X, G(X)]^{\beta_j} \cdot [s_j(X, G(X)]^{\gamma_j} \quad (6)$$

where α_M, β_j and γ_j are weight parameters (usually: $\alpha_M = \beta_j = \gamma_j$, and $\sum_{j=1}^{M} \gamma_j = 1$).

The closer the values of $SSIM$ and $MS\text{-}SSIM$ are to 1, the more similar the two images. On the contrary, the closer they are to 0, the greater the difference between the two images. So based on the definitions of $SSIM$ and $MS\text{-}SSIM$, the $SSIM$ and $MS\text{-}SSIM$ losses are defined as:

$$Loss_{SSIM} = 1 - SSIM(X, G(X)). \quad (7)$$

$$Loss_{MS\text{-}SSIM} = 1 - MS\text{-}SSIM(X, G(X)). \quad (8)$$

(5) $LPIPS$ Loss

Zhang R et al. [25] proposed that the traditional similarity measurement methods (e.g., L_2 and $SSIM$) are often inconsistent with human judgment. They find that $LPIPS$ (Learned Perceptual Image Patch Similarity) outperforms all previous metrics by large margins of their dataset.

For calculating the perceptual similarity between two images x_1, x_2, given a base network F (e.g., SqueezeNet, AlexNet, VGG), we first calculate the deep embeddings of x_1 and x_2 on network F, respectively, normalize the activations in each channel dimension, use the vector w_l to scale the size of each channel, and calculate the L_2 distance of the scaled features of each channel. Finally, we average across spatial dimension and across all layers to obtain the perceptual similarity between the two patches x_1, x_2.

$$d(x_1, x_2) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}^l_{1,hw} - \hat{y}^l_{2,hw}) \right\|_2^2. \quad (9)$$

where l represents the *l-th* layer of the network F, the result of *l-th* layer $\hat{y}^l_1$ is obtained by inputting x_1 into the network F, and H, W are the height and width of the feature map of the *l-th* layer.

The $LPIPS$ perception distance needs to be trained to predict the perception judgment from a distance calculated by Equation (9). Then the network of computer distance needs

to be trained. The training methods include lin method, tune method and scratch method, which are described in the literature [25].

The lower the value of $LPIPS$, the more similar the two images are, and vice versa, the greater the difference. The $LPIPS$ loss is represented by the distance between X and $G(X)$ calculated by Equation (10).

$$Loss_{LPIPS} = d(X, G(X)). \tag{10}$$

We analyze the impact on different loss functions of the final inpainting results through two types of experiments. The first type of experiment is to complete the face image completion through only one loss function. The experimental results of the two groups are shown in Figure 4. Column a is the images that need to be inpainted. Column b, c, d and e are inpainted images by using only L_2, Log-$Cosh$, $SSIM$, MS-$SSIM$ losses, respectively. Column f is the original images. It can be found that the images generated by only using L_2 loss are too smooth and losing more details; the images generated by only using Log-$Cosh$ loss are better, but they also have a problem of excessive smoothness in the hair area; the generated images only using $MMIS$ loss are not ideal and only have the structural integrity of the faces; the inpainted images by only using MS-$MMIS$ loss also have the structural integrity, but they have obvious brightness deviations.

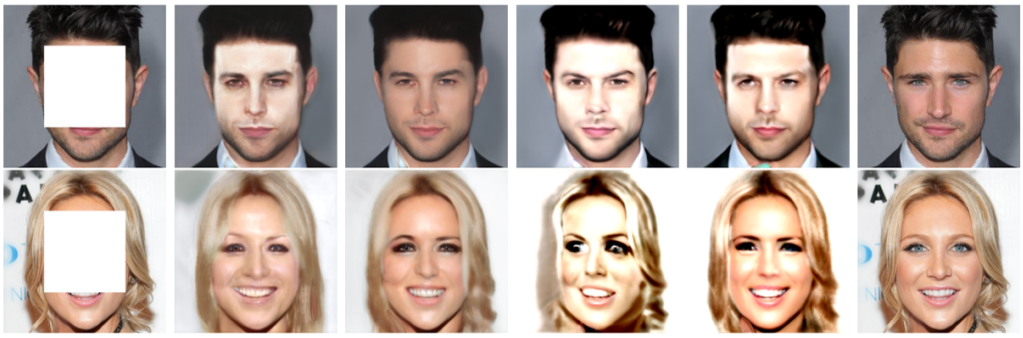


Figure 4. Comparison of completion results with different loss functions. (**a**) Masked images, (**b**–**e**) Inpainted results using L_2, Log-$Cosh$, $SSIM$, MS-$SSIM$ losses, respectively. (**f**) Original images.

The second type of experiment describes the inpaint of face images by combining different types of loss functions. The experimental results of two groups are shown in Figure 5. The amazing inpainting results can be clearly seen, and are obtained through the combination of various loss functions (as shown in Figure 5g). After generating high-resolution face images by using various loss functions, the generated images must be processed to complete the entire inpainting work. A simple processing method is to directly paste the to-be-repaired area corresponding to the generated image into the to-be-repaired image. This processing method can cause boundary marks to appear in the final inpainting results. In experiments, we process the inpainting boundary via Poisson mixing [18] to make the boundary of the final inpainting results more natural (Figure 5h).

Specifically, in the follow-up experiments, we completed the inpainting of high-resolution face images through a combination of four losses: $Loss_{VGG}$, $Loss_{Log\text{-}Cosh}$, $Loss_{MS\text{-}SSIM}$ and $Loss_{LPIPS}$. The overall loss function is shown in Equation (11).

$$Loss_{total} = \lambda_1 Loss_{VGG} + \lambda_2 Loss_{Log\text{-}Cosh} + \lambda_3 Loss_{MS\text{-}SSIM} + \lambda_4 Loss_{LPIPS}. \tag{11}$$

where λ_1, λ_2, λ_3 and λ_4 are the weights of VGG Loss, Log-$Cosh$ Loss, MS-$SSIM$ Loss and $LPIPS$ Loss.

(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)**

Figure 5. Comparison of completion results with the combination of different types of loss functions. (**a**) Original images, (**b**) Masked images, (**c**) Generated images by using $Loss_{VGG}$ loss, (**d**) Generated images by using $Loss_{VGG}$ and $Loss_{L_2}$ losses, (**e**) Generated images by using $Loss_{VGG}$ and $Loss_{Log\text{-}Cosh}$ losses, (**f**) Generated images by using $Loss_{VGG}$, $Loss_{Log\text{-}Cosh}$ and $Loss_{MS\text{-}SSIM}$ losses, (**g**) Generated images by using $Loss_{VGG}$, $Loss_{Log\text{-}Cosh}$, $Loss_{MS\text{-}SSIM}$ and $Loss\text{-}LPIPS$ losses, (**h**) are the inpainted images obtained by Poisson fusion on the basis of (**f**).

4. Experimental Results and Analysis

To objectively and comprehensively evaluate the performance of our method, this section compares the completed results of various models through qualitative evaluation and quantitative evaluation. We evaluate the proposed method on the public dataset *CelebA-HQ*. *CelebA-HQ* is a high-resolution face image dataset and contains 30,000 $1024 * 1024$ high-definition frontal face images. In our experiments, the weight values of Equation (11) are set as $\lambda_1 = 0.4$, $\lambda_2 = 1.5$, $\lambda_3 = 100$, $\lambda_4 = 100$, respectively. In this paper, we use TensorFlow framework as the development environment. The configuration of the experimental platform is an Intel(R) Core(TM) i7-9750H 2.60GHz CPU, and an NVIDIA GeForce GTX 1650 GPU. The images of our intermediate generation process are shown in Figure 6. It can be seen that as the iteration progresses, the details of the generated face images become increasingly rich, and the visual consistency gradually improves in quality.

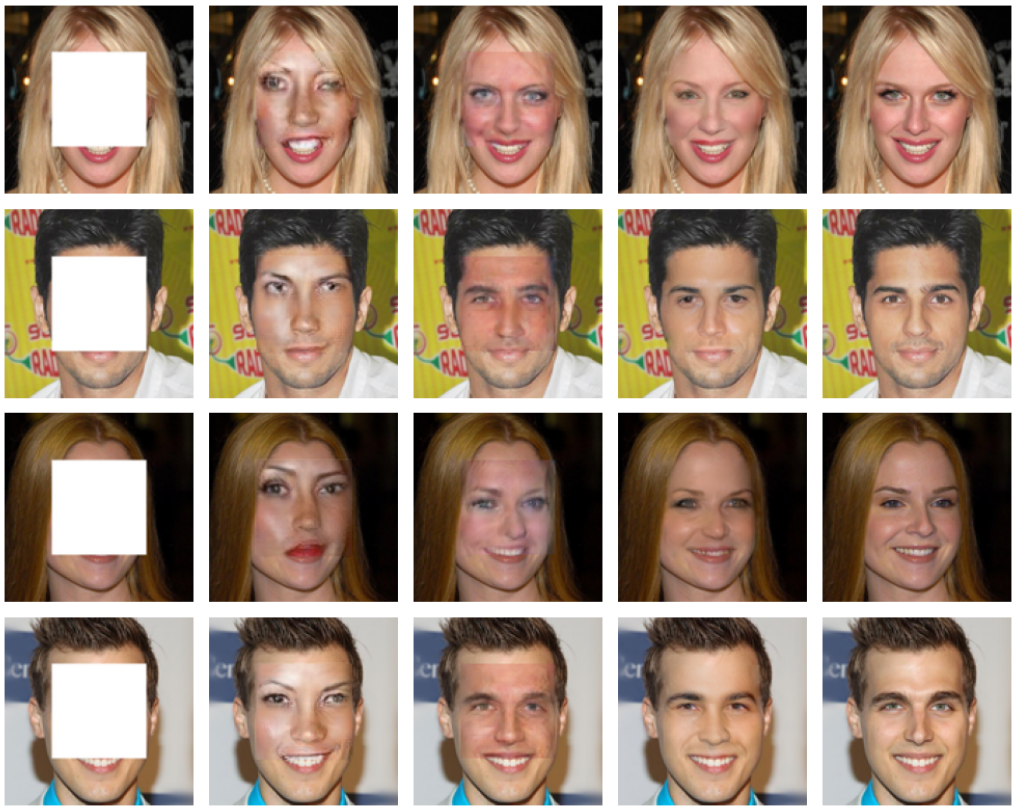
Figure 6. The images generated by the intermediate process of our method when the images are completed.

4.1. The Qualitative Evaluation

The center rectangle mask is the most common comparison method in image completion. CE [5] uses a method based on context coding to complete the missing region. GLCIC [8] uses global and local discriminators to complete images. Because each method has different requirements for the input size of the masked images, each method has different resolution of output images. The default resolution of output images by using GLCIC method is $128 * 128$ and the default resolution of output images by using CE is $256 * 256$. The method proposed in this paper is for high-resolution face image completion, and the resolution of output images is $1024 * 1024$. For uniform comparison, we re-scaled the resolution of all inpainted images to $128 * 128$. The central rectangular mask is a common damaged shape in face completion tasks. For a fair comparison, we ensure that the mask area occupies the same proportion of the entire image area in different methods. In the experiment, we set the proportion to be one-fourth, that is, for CE, GLCIC and our proposed

method, we use $64 * 64$, $128 * 128$ and $512 * 512$ center rectangle mask inputs to achieve face image inpainting.

As shown in Figure 7, we show four groups of the result images inpainted by these three methods. Regarding the experimental results, there are unnatural contents in the results of face completion by the GLCIC method, and there are conspicuous color differences on the edges inpainted by CE method. However, the results of our proposed method contain more detailed information while ensuring the consistency of visual effects. In addition, by using the Poisson fusion method, traces of inpainting boundaries are invisible in the completion imagery.



(a) **(b)** **(c)** **(d)** **(e)**

Figure 7. Completion results on the center rectangle mask. (**a**) Masked images. (**b**) GLCIC results [8]. (**c**) CE results [5]. (**d**) Our results. (**e**) Original images.

To compare the inpainting effect with GLICI more comprehensively, we also compare the inpainted images by using the large-area rectangular mask and the irregular mask. Figure 8 shows the inpainting results by using the large-area rectangular mask. Figure 9 shows the inpainting results by using the large-area irregular mask. The experimental results in Figure 8 show that the inpainting results of the method proposed in this paper have richer face structures and more coherent edges. In addition, the inpainting results are already very similar to the original image, especially in the case of symmetrical information loss. In the case of using arbitrary shape masks (Figure 9, the repair marks are very obvious in the repair results of the CE method, and some results even exist face distortion. The method in this paper ensures that the face structure is reasonable, and the color and facial expression are basically the same as the original image.

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8. Completion results on the large-area rectangular masks. (**a**) Masked images. (**b**) GLCIC results [8]. (**c**) Our results. (**d**) Original images. (**e**) Masked images. (**f**) GLCIC results [8]. (**g**) Our results. (**h**) Original images.

(a) (b) (c) (d)

Figure 9. Completion results on the large-area irregular masks. (**a**) Masked images. (**b**) GLCIC results [8]. (**c**) Ours results. (**d**) Original images.

The method proposed in this paper does not need to retrain the model for different repair areas. Thus, it can repair images of any mask easily. In our experiments, we repair the images under different damage conditions, with the experimental results shown in Figures 10 and 11. Figure 10 shows the completed results of images damaged by 20%, 30%, 40% and 50% random noise masks. We can see that in the case of 20% and 30% noise masks,

the completed results are not visually different from the original images, and in the case of 40% and 50% noise, the completed results are only slightly blurred. Figure 11 shows the completed results of images damaged by free-form brushes accounting for 10–20%, 20–30%, 30–40% and 40–50% of image pixels, the conclusion is still consistent with Figure 10. We also try to repair the image damaged by 90% random noise masks, our method can still achieve good repair effect, as shown in Figure 12. Therefore, our method can effectively repair face images with random damage.

Figure 10. Inpainting results of masked images with different proportion noise. The first to fourth rows are 20%, 30%, 40% and 50% noise masks, respectively.

Figure 11. Inpainting results of masked images with different proportion free-form brush. The first to fourth rows are 10–20%, 20–30%, 30–40% and 40–50% random free-form brush masks, respectively.

Figure 12. Inpainting results for 90% random noise masked face images.

4.2. The Quantitative Evaluation

In order to evaluate the image inpainting effect of these methods objectively, in this section we select the experimental results of using the rectangular center mask for quantitative comparison. PSNR(Peak Signal-to-noise Ratio) and SSIM(Structural similarity) are measures to evaluate the quality of completed images. In the image evaluation measures, PSNR is used most, but its value can not well reflect the subjective feeling of human eyes. Table 1 is the performance comparison on the CelebA-HQ dataset. In this table, the best value of each measure are indicated in bonds. From this table we can infer that the SSIM of our results is better than other compared methods and the highest PSNR is obtained by CE method.

Table 1. Comparison of the center rectangle masked completion results on the CelebA-HQ dataset.

Methods	SSIM	PSNR
GLCIC	0.798	21.87
CE	0.887	25.93
Ours	0.903	23.35

In addition, the flexibility of CE and GLCIC is insufficient. For masks of different shapes, they all need to train a new inpainting model. However, our method can complete face images with arbitrary shape masks by training only one model.

In order to further prove the effectiveness of our method, we select 1000 face images in CelebA-HQ database randomly, destroy them with random noise masks and random free-form brush masks, respectively, and then repair the damaged images by our method. The experimental results are shown in Figures 10 and 11. Furthermore, we calculated the average values of PSNR and SSIM shown in Table 2. It can be seen, the values in Table 2 are higher than the values in Table 1, which shows that our method is universal in repairing arbitrarily damaged face images.

Table 2. Qualitative analysis of inpainting results of random noise and random free-form brush masks with different percentage.

Random Noise Masks			Random Free-Form Masks		
Percentage	SSIM	PSNR	Percentage	SSIM	PSNR
20%	0.960	38.79	10-20%	0.964	33.86
30%	0.890	35.29	20-30%	0.918	29.06
40%	0.820	32.29	30-40%	0.909	27.57
50%	0.807	31.90	40-50%	0.861	25.62

4.3. Network Complexity

The networks used in our method include StyleGAN and ResNet. We use StyleGAN network pre-trained on CelebA data-set and ResNet50 network pre-trained on ImageNet

data-set to obtain latent vector prediction network. Stylegan network includes 4.93$E7$ parameters and ResNet50 network includes 2.55$E6$ parameters. Therefore, the whole network include 5.12$E7$ parameters.

In Figure 13, we show the average value of the total weighted loss values in 100 times inpainting process about 1000 face images with random noise masks and random free-form brush masks. It can be seen that our method converges after 80 rounds of iteration for different types of masks.

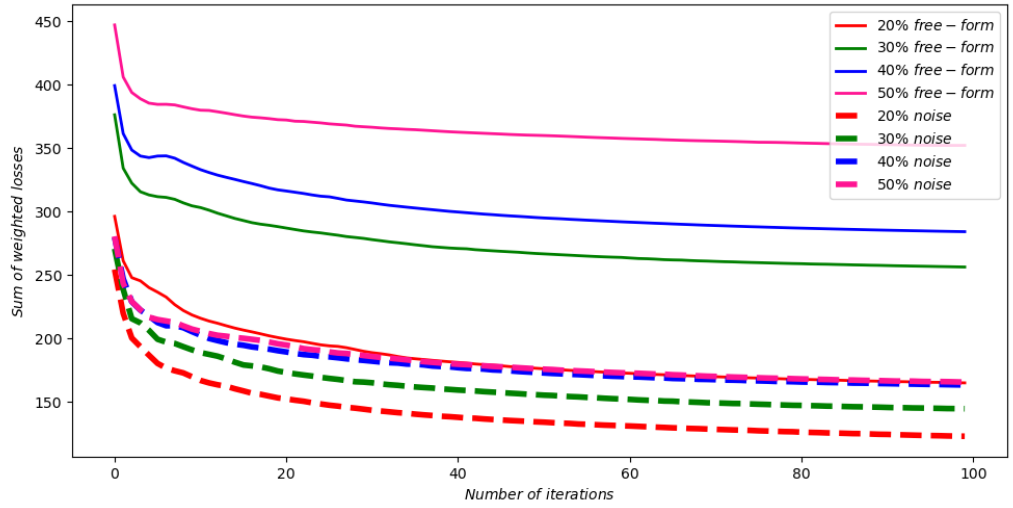


Figure 13. The graph of the total weighted loss changes during the iterative inpainting process.

In Figure 14, we show the average completion time of 1000 face images with random noise masks and random free-form brush masks. The completion time of each image includes the time of repair phase and fusion phase. It can be seen, the difference in completion time is caused by the fusion stage. As shown in Figure 14, the completion times of inpainting images with random noise masks are longer than that of the random free form brush masks, which is mainly due to the images of random noise masks need more fusion time because random noise is more dispersed. The average completion time of repairing a $1024 * 1024$ face image with 20% free-form brush masks is 122 s. The average completion time of repairing a $1024 * 1024$ face image with 20% random noise masks is 143 s. Nevertheless, the cost of these time can still meet the requirements of practical application.

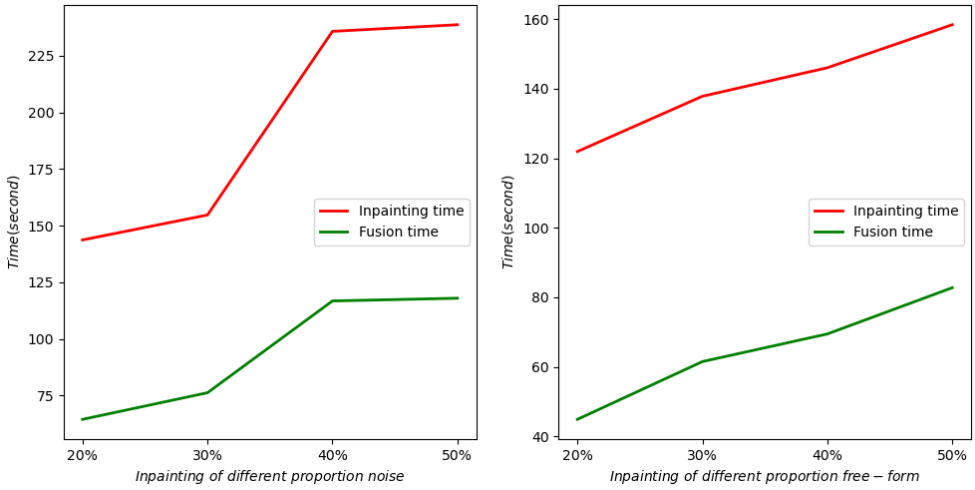


Figure 14. The graph of the average completion time of 1000 face images with random noise masks and random free-form brush masks.

4.4. Application and Failure Case

Because our method can complete high resolution images with random masks and large damaged area, it can be applied to face image editing, face image subtitle removal, face image watermark removal and other practical applications. The specific experimental results are given in Figure 15.

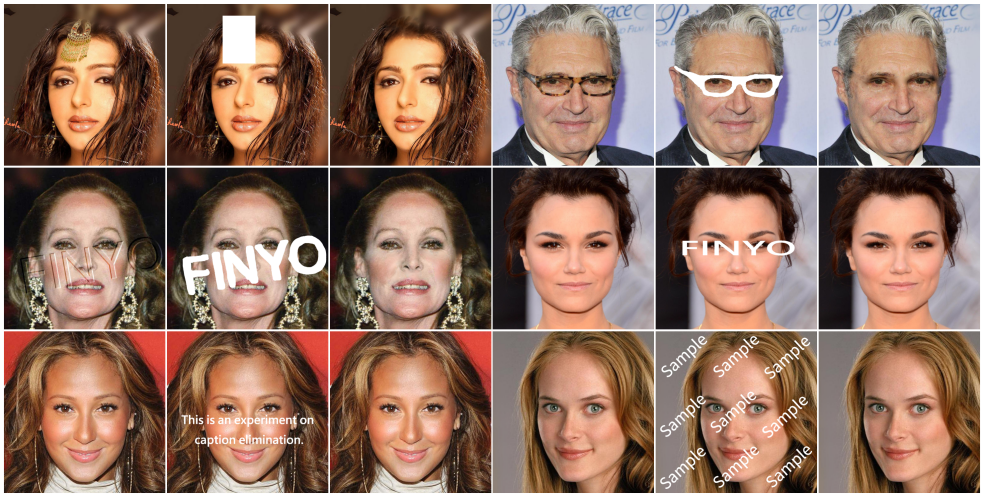


Figure 15. Different application experiments on face image inpainting. The experiments include watermark removal, text removal, and image editing.

Our method also has some failure cases shown in Figure 16. This is mainly because the main purpose of our face generation network is to generate face images. The generation of the face images with other objects (such as sunglasses) are not ideal.

Figure 16. An example of a failed face image inpainting.

4.5. Diversity of Completion Results

For an occluded face image, there should be multiple possible inpainting results. Although there are many methods that can complete reasonable inpainting, these methods often have only one output. The inpainting model proposed in this paper can obtain diverse inpainting results after simple adjustment. When our method completes a high-resolution face image, it will obtain a corresponding latent vector V_{Latent}. Further, by editing V_{Latent}, multiple completion results can be obtained. StyleGAN establishes a mapping relationship between an (18, 512) dimensional vector to a (1024, 1024, 3) dimensional vector(face image), which has some attribute such as age, gender, expression, etc. If we can explore the relationship between the change of the face attribute and the corresponding change of V_{Latent}, then the attributes of the generated face image can be edited and manipulated through the adjustment of V_{Latent}.

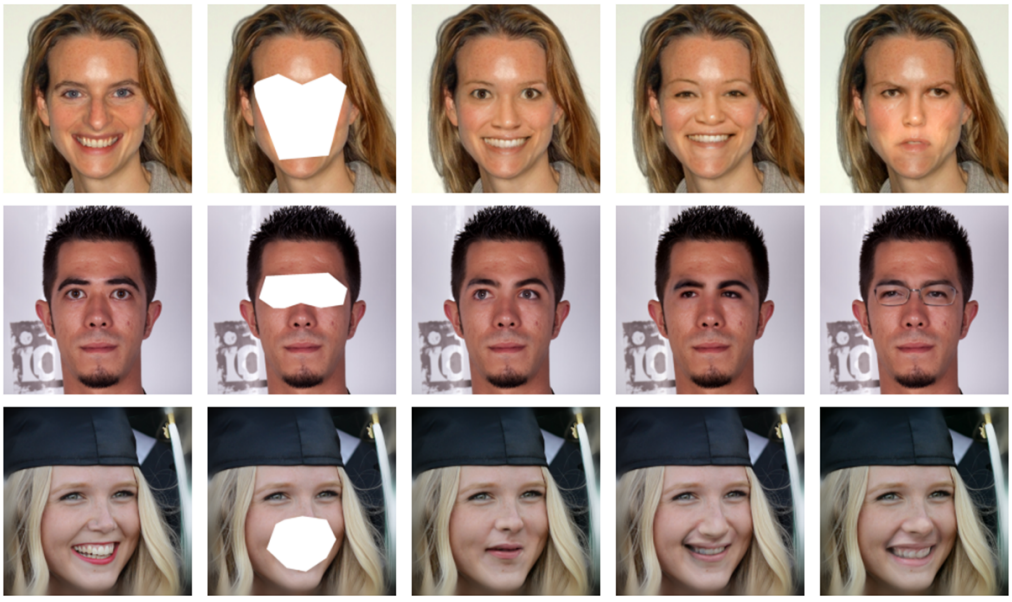
We can first use the face feature extraction method to extract the attribute features of each face image, and use the numerical value to represent each attribute to constitute the

feature vector V of each face image attribute. For the convenience of operation, the range of the value for each dimension in the vector V is normalized to the $[0, 1]$. Then, for the data of each dimension in V (such as the data representing the age feature), we explore the relationship between the change of its value and the change of the corresponding latent vector value. This changing relationship can be represented by the direction vector $\vec{e}_i$, which has the same dimension as the latent vector. $\vec{e}_i$ is calculated by the ratio of the difference value of the latent vector to the difference value of the corresponding attribute. In order to obtain an accurate $\vec{e}_i$, it is necessary to calculate the mean of the ratios for each pair of sample in a large number of face attribute data-sets, which is very difficult. In fact, the solution of $\vec{e}_i$ can be carried out through an optimization strategy. First, the median $v^m_{i,j}$ of the i-th dimension (corresponding to an attribute) of the feature vector V_j is used as the dividing line to establish a new two-class sample. For the samples whose corresponding attribute values are lower than $v^m_{i,j}$ in the original samples, the labels $y_{i,j}$ are set to 0. The samples whose corresponding attribute values are higher than $v^m_{i,j}$, the labels $y_{i,j}$ are set to be 1. V_j represents the attribute feature of the j-th face sample image. Then we construct the objective function and use logistic regression to solve the binary classification problem. The objective function is defined as: $\vec{w}_i v_{i,j} + b = y_{i,j}$, where $v_{i,j}$ is the value of the i-th dimension of V_j and $y_{i,j}$ is the label value of the binary classification. By using this optimization method, $\vec{w}_i$ can be solved in a relatively fast time and $\vec{e}_i$ can be approximately represented by using $\vec{w}_i$. According to the latent vector of the face image to be completed and the direction vector $\vec{e}_i$, the following formula can be used to realize the editing of the corresponding features of the face to achieve Diversification of completed results.

$$N_{Latent} = V_{Latent} + Coef * \vec{e}_i. \tag{12}$$

where $Coef$ is the editing parameter. Different latent vectors can be obtained by giving different values to $Coef$.

In this way, by editing the latent vector V_{Latent}, we can achieve the generation of face images with different attributes, such as facial expressions. Then, the image to be inpainted and the generated image are fused by the Poisson fusion method to achieve the diversity of repair results. Figure 17 shows the diversity inpainting results we obtained by editing the latent vector.



(a) Original images **(b) Masked images** **(c) Results** **(d) Results** **(e) Results**

Figure 17. Examples of diversity inpainting results.

5. Discussion

As a highly popular research topic in the field of image inpainting, the completion of face images has been widely studied in recent years. However, high-resolution face image inpainting has always represented a challenge due to the difficulty of training high-resolution generative networks and the fact that faces have clear semantic features. The StyleGAN network was proved to be able to generate a variety of rich high-resolution face images by adjusting the latent vector, which renders the repair of high-resolution face images feasible. The method proposed in this paper modifies the latent vector iteratively through various loss functions to control the StyleGAN network to generate the face image most similar to the undamaged section of the face image.Then, our method synthesizes the inpainted result image through Poisson fusion. The method proposed in this paper can also be employed for other types of high-resolution image inpainting by training specific high-resolution generative networks of the corresponding type. In the future, for face image inpainting, we intend to complete the inpainting by adding facial structure guidance components. For example, according to the face image to be inpainted, we can first obtain the contour content of the face to be repaired according to the prior information of the face structure, and use it as the completion guidance condition to complete the face image inpainting. In addition, different components (contours, textures) of the face image can be decomposed and face image inpainting can be conducted on the different components. All of these research contents can be combined with our method. We also expect our method to be applied to various applications of face completion.

6. Conclusions

A new high-resolution face images completion method has been proposed in this paper. This method first uses ResNet to predict the latent vector of a face image, and then generates high-resolution face images through the StyleGAN model to repair damaged face imagery. This method can successfully synthesize the missing content in the face image, so that the inpainted image is semantically effective and visually reasonable. StyleGAN model can first train the generator and discriminator with low spatial resolution, and gradually add layers to the generator and discriminator network to increase the resolution of the generated image. In this way, the problem that it is difficult to converge in training high-resolution networks has been solved. Although the proposed model is only used for facial image completion, after training the generation network with other types of imagery, this method can be used for the restoration of other types of image restoration.

Our proposed method performs iterative repair process by calculating the loss of the known area of the image to be repaired and the corresponding area of the generated image, so the shape of the mask areas are not a factor that directly affects the completed results. There is no need to train and modify our model for different masks, which renders our method easy to apply to face image editing, face-image watermark removal, face image subtitles removal and so forth.

In addition, when the incomplete image is repaired, we can obtain the latent vector corresponding to the inpainted image. Next, we can then modify the vector value representing the specific area in the latent vector, finally generating a new face image through the modified latent vector. In this way, our method can realize the editing of specific facial regions and the diversity of face image completion results with the same mask. This method can also be used to edit facial features, such as the repaired image results of a person with or without glasses, which can be successfully obtained by using this method.

Further, the method proposed in this paper can be extended to repair and edit various high-resolution images.

Author Contributions: Conceptualization, Z.Q. and L.H.; methodology, X.S.; software, X.S. and H.L.; validation, L.H. and M.W.; formal analysis, Z.Q. and F.D.; investigation, L.H.; resources, Z.Q. and L.H.; data curation, X.S.; writing—original draft preparation, L.H.; writing—review and editing, Z.Q. and F.D.; visualization, X.S.; supervision, Z.Q.; project administration, Z.Q.; funding acquisition, L.H. and Z.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the projects of Natural Science Foundation of China (Grant No. 12163004), the Yunnan Fundamental Research Projects (Grant No. 202001AT070135, No. 202002AD080002), the key scientific research project of Yunnan Police College (19A010).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Networks
GAN	Generative Adversarial Network
DCGAN	Deep Convolutional GAN
ResNet	Deep Residual Network
PGGAN	Progressive Growing GAN
AdaIN	Adaptive Instance Normalization
VGGNet	Visual Geometry Group Networks
SSIM	Structural SIMilarity
MS-SSIM	MultiScale Structure SIMilarity
LPIPS	Learned Perceptual Image Patch Similarity
CE	Context Encoders
GLCIC	Globally and Locally Consistent Image Completion
PSNR	Peak Signal-to-noise Ratio

References

1. Bertalmío, M.; Sapiro, G.; Caselles, V.; Coloma, B. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, Washington, DC, USA, 19–23 July 2020.
2. Qiang, Z.; He, L.; Xu, D. Exemplar-Based Pixel by Pixel Inpainting Based on Patch Shift. In Proceedings of the CCF Chinese Conference on Computer Vision (CCCV), Tianjin, China, 11–14 October 2017; pp. 370–382. [CrossRef]
3. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680. [CrossRef]
4. Zhao, D.; Cai, J.; Qi, Y. Convincing 3D Face Reconstruction from a Single Color Image under Occluded Scenes. *Electronics* **2022**, *11*, 543. [CrossRef]
5. Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544. [CrossRef]
6. Yeh, R.A.; Chen, C.; Yian Lim, T.; Schwing, A.G. Semantic Image Inpainting with Deep Generative Models. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6882–6890. [CrossRef]
7. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
8. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–14. [CrossRef]
9. Li, Y.; Liu, S.; Yang, J.; Yang, M.H. Generative face completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3911–3919.
10. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
11. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
13. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.

14. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4471–4480.
15. Yang, Y.; Guo, X.; Ma, J.; Ma, L.; Ling, H. Lafin: Generative landmark guided face inpainting. *arXiv* **2019**, arXiv:1911.11394.
16. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp.770–778.
18. Pérez, P.; Gangnet, M.; Blake, A. Poisson image editing. *ACM Trans. Graph.* **2003**, *22*, 313–318. [CrossRef]
19. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
20. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1501–1510.
21. Lipton, Z.C.; Tripathi, S. Precise recovery of latent vectors from generative adversarial networks. *arXiv* **2017**, arXiv:1702.04782.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
24. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
25. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.