Article

A Reference-Guided Double Pipeline Face Image Completion Network

Hongrui Liu 1,2 ©, Shuoshi Li 1 ©, Hongquan Wang 1 © and Xinshan Zhu 1,2* ©

1 School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; hrliu@tju.edu.cn (H.L.); lishuoshi1996@foxmail.com (S.L.); 1015203049@tju.edu.cn (H.W.)
2 State Key Laboratory of Digital Publishing Technology, Beijing 100871, China
* Correspondence: xszhu@tju.edu.cn; Tel.: +86-151-2258-0213

Received: 25 October 2020; Accepted: 14 November 2020; Published: 22 November 2020

Abstract: The existing face image completion approaches cannot be utilized to rationally complete damaged face images where their identity information is completely lost due to being obscured by center masks. Hence, in this paper, a reference-guided double-pipeline face image completion network (RG-DP-FICN) is designed within the framework of the generative adversarial network (GAN) completing the identity information of damaged images utilizing reference images with the same identity as damaged images. To reasonably integrate the identity information of reference images into completed images, the reference image is decoupled into identity features (e.g., the contour of eyes, eyebrows, nose) and pose features (e.g., the orientation of face and the positions of the facial features), and then the resulting identity features are fused with posture features of damaged images. Specifically, a lightweight identity predictor is used to extract the pose features; an identity extraction module is designed to compress and globally extract the identity features of the reference images, and an identity transfer module is proposed to effectively fuse identity and pose features by performing identity rendering on different receptive fields. Furthermore, quantitative and qualitative evaluations are conducted on a public dataset CelebA-HQ. Compared to the state-of-the-art methods, the evaluation metrics peak signal-to-noise ratio (PSNR), structure similarity index (SSIM) and L1 loss are improved by 2.22 dB, 0.033 and 0.79%, respectively. The results indicate that RG-DP-FICN can generate completed images with reasonable identity, with superior completion effect compared to existing completion approaches.

Keywords: face image completion; generative adversarial network; double pipeline; feature decoupling; feature fusion

1. Introduction

Face image completion has been extensively utilized as a research hotspot in the field of computer vision in the film industry and detection. Face image completion should effectively restore the reasonable identity of a damaged image, i.e., ensure that the completed image has the same identity as the original image. Mainly, traditional image completion approaches are based on patches or diffusion. The diffusion-based approach [1,2] is to iteratively propagate the low-level features from known regions to unknown areas along the mask boundaries to complete the small-hole area with weak structure. The patch-based approach [3–8] is to search for similar patches in the same dataset or image to complete the unknown areas. Both approaches assume unknown areas share similar content with known regions; thus they can directly match, copy and realign the known patches to complete unknown areas, but cannot predict unique content not present in the known regions. Thus, these methods are only appropriate for filling natural images with similar textures. Nevertheless, face images have stronger topology than natural images, hence, they cannot be completed by such traditional approaches.
Presently, the approach based on deep learning and a generative adversarial network (GAN) [9] has attracted research on face image completion. In such an approach, a trainable completion network is made mostly based on a convolutional neural network, and the completion network is utilized as a discriminator and generator for adversarial training to generate natural and real completed images. A context encoder was proposed by Pathak et al. [10] using the generative adversarial loss for training, based on which Iizuka et al. [11] proposed local and global discriminant loss. Yu et al. [12] put forward a completion plan of rough completion first and then fine completion to obtain a high-definition completion effect through long-range feature transfer using the semantic attention layer. Moreover, some scholars proposed partial convolution [13] and gated convolution [14] to enhance the network robustness to the mask shape. Given the problem of a single completion result, Zhang et al. [15] mapped the damaged images to probability distribution utilizing a modified variational encoder and GAN, so that multiple completion results can be constructed for one damaged image by sampling the probability distribution.

The completion results of the dataset-guided face image completion scheme [10–15] are natural and real on face images with completely lost identity information. However, the results are essentially average faces based on the entire dataset, due to their identities being uncertain. Therefore, Li et al. [16] proposed semantic parsing loss, effectively enhancing the similarity between the completion result and the topological structure of the original images. Some scholars [17–21] used landmarks and edges as prior knowledge to guide the completion process. Compared with the approach of [16], the visual quality of the completion results of such approaches is greatly improved with more stable training. However, these approaches can only provide a posture information guide to the completion procedure, and the completion results still have no certain identity.

To solve the identity uncertainty problem, a reference-guided double-pipeline face image completion network (RG-DP-FICN) is proposed in this paper to guide the entire completion process using the identity information of reference images. The aim of the network is to transfer the identity information from the reference images to completed images, thereby improving the identity plausibility of the completed images.

The key points of this paper are as follows:

1. The reference-guided completion network is used to determine the rationality of the identity of completion results.
2. A double-pipeline GAN is proposed to realize the decoupling and fusion of identity and posture features.
3. An attention feature fusion module is designed to restore the background information lost during fusion.

2. Materials and Methods

2.1. Dataset and Evaluation Metrics

All the experiments conducted in this paper exploit a high-quality human face dataset CelebA-HQ [22]. This dataset contains 30,000 high-quality face images coming from 6217 identities. It is trained and tested with the processing as follows. First, the images with the same identity in the dataset are summarized to obtain 6217 image lists. Each list contains a different number of face images. Then, the lists of only one image with the same identity are eliminated. Two images are selected in order from the list as the original image and the reference image, respectively, and 28,299 image pairs are obtained. Ultimately, 27,299 image pairs are chosen as the training set, while the remaining 1000 image pairs as the test set. All images are scaled to 256 × 256 × 3.

The evaluation metrics include peak signal-to-noise ratio (PSNR) [23], structure similarity index (SSIM) [23] and L1 loss. These metrics are used to measure the pixel-level difference and overall similarity between the completed image and the original image. For PSNR and SSIM, higher values
where, $I_a$ represents the completed image, $I_g$ represents the completed image, $i$ and $j$ represent the coordinate positions of pixels in $I_a$ and $I_g$; $H, W$ represent the width and height of the image respectively; $l, c$ and $s$ represent the image similarity of $I_a$ and $I_g$ in terms of luminance, contrast, and structure, respectively; $\mu_g, \mu_s, \sigma_s, \sigma_g$ represent the mean and variance of $I_a$ and $I_g$, $\sigma_{gs}$ represent the covariance of $I_a$ and $I_g$; $C_1, C_2, C_3$ represent the constant in order to avoid the denominator is 0, generally take $C_1 = (K1 \times L)^2$, $C_2 = (K2 \times L)^2$, $C_3 = (K3 \times L)^2$, $K1 = 0.01$, $K2 = 0.03$, $L = 255$.

2.2. Double-Pipeline GAN

Considering identity uncertainty problem in the completed image, reference image $I_a$ with the same identity as the damaged image $I_m$ is incorporated to guide the completion procedure. In this study, a reference guided double-pipeline face image completion network (RG-DP-FICN) is proposed (Figure 1). Two pipelines (reconstruction and completion) are constructed in this network, to reasonably incorporate the identity information of reference images. To realize the identity transfer from reference image to damaged image, the posture features (e.g., the orientation of face and the positions of the facial features) and identity features (e.g., the contour of eyes, eyebrows, nose) of reference images are decoupled through such double-pipeline network design, and the identity features of reference images are fused with the posture features of damaged images as follows.

The reconstruction pipeline possesses all the information on reference images and the information is jointly obtained by the encoder $E$ and identity encoder $E_{id}$. The reconstruction process is as follows: First, the damaged image $I_m$ and the topological structure of the reference image $I_a$ are encoded via the encoder $E$ to attain posture feature $z_{ma}$. Then, $z_{ma}$ and identity feature $z_{id}$ extracted by $E_{id}$ are fused utilizing the identity transfer module to obtain $z_{maid}$. Ultimately, $z_{maid}$ is inserted into the decoder $F$ for decoding to attain the reconstructed image $I_{rec}$ (Equations (5) and (6)). Therefore, by supervising the similarity between the reconstructed image $I_{rec}$ generated via the pipeline, the reference image $I_a$ can effectively realize the decoupling of the posture feature $z_{ma}$ and the identity feature $z_{id}$. As for the completion pipeline, the damaged image $I_m$ and the topological structure $L_g$ are encoded by the encoder $E$ to obtain $z_m$. Then, the identity feature $z_{id}$ obtained by decoupling the reference image and the posture feature $z_m$ of the damaged image are fused utilizing the identity transfer module to obtain $z_{a}$. Ultimately, $z_{maid}$ is inserted into the decoder $F$ for decoding to generate the completed image $I_g$ with certain identity (Equations (7) and (8)).
where, \( I_g \) represents the original image; \( M \) is a mask; \( I_a \) and \( L_g \) denote the face topological structure obtained by connecting the landmarks of \( I_a \) and \( I_m \) predicted by the landmark predictor [20].

**Figure 1.** Overview of RG-DP-FICN with two parallel pipelines. The reconstruction pipeline (blue line) consisting of generator \( G \) and discriminator \( D_2 \) possesses all the information on reference images from \( I_a \), which is used only for training. The completion pipeline (yellow line) consisting of generator \( G \) and discriminator \( D_1 \) is used for training and testing. Two pipelines share identical generator \( G \) and identity encoder \( E_{id} \), the generator consists of encoder \( E \) and discriminator \( F \). The structure of the residual encoder and residual decoder is the same as [15]; Identity encoder module and identity transfer module is further described in Section 2.3; attention feature fusion module is further described in Section 2.4.

The generator \( G \) is trained to enhance the visual quality of \( I_g \) and the rationality of its topological structure through adversarial training with two discriminators \( D_1 \) and \( D_2 \). For \( D_1 \), the true sample and the false sample are the image pair \((L_g, I_g)\) and \((L_g, I_g)\), respectively. For \( D_2 \), the true sample
and the false sample are respectively the image pair \((L_a, I_a)\) and \((L_a, I_{rec})\). Such sample pair design can effectively enhance the image quality of \(I_g\) and \(I_{rec}\), and ensure the rationality of their topological structure.

2.3. Identity Encoder and Identity Transfer Module

The fusion of the identity information of reference images and the original information of damaged images is a challenging problem. This is because first, the reference images present different visual appearances due to gender, lighting conditions, and makeup, hence, it is difficult to obtain identity information. Second, there are differences in the face posture along with different proportions and appearances due to gender, lighting conditions, and makeup, hence, it is difficult to obtain identity information. Hence, to solve this problem, an identity transfer module and an identity encoder are designed in this paper (Figure 2) to extract the identity information and feature fusion, respectively. Next, the transfer process in the completion pipeline is introduced as an example.

![Figure 2. The identity encoder and identity transfer module.](image)

In the identity encoder, the residual block and the self-attention module \([24]\) are used first to extract and compress identity information of reference image \((256 \times 256 \times 3)\) to obtain a feature \((512 \times 4 \times 4)\). Then, the feature is mapped into the identity feature \(z_{id}\) \((512 \times 1 \times 1)\) through a fully connected layer. As a global operation, each unit of the identity feature \(z_{id}\) is the weighted sum of \(512 \times 4 \times 4\) feature values, which allows each unit in \(z_{id}\) to reason about reference image identity. Hence, \(z_{id}\) could well represent the identity of reference images.

In the identity transfer module, \(z_m\) and \(z_{qm}\), which representing high-dimensional semantics, is globally adjusted step by step using the style rendering block to realize identity transfer. The Identity transfer module consists of four style rendering blocks, each of which contains two AdaIn layers. The specific structure of style rendering block is represented in Figure 3, which is employed twice through adaptation instance normalization (AdaIN) \([25]\), as follows:

\[
    f_{ij} = \frac{f_{ij} - u(f_{ij})}{\sigma(f_{ij})} \times \beta_{ij} + \alpha_{ij}
\]

where, \(f_{ij}\) represents the input feature, \(u(f_{ij})\) and \(\sigma(f_{ij})\) denote the mean and variance of \(f_{ij}\), respectively, \((\alpha_{ij}, \beta_{ij})\) represents the AdaIn affine parameter obtained by the fully connection layer mapping of identity feature \(z_{id}\), \(f_{ij}'\) is the rendered feature, \(i\) indicates the \(i\)-th style rendering block, and \(j\) shows the \(j\)-th AdaIn of each style rendering block.
Figure 3. The style rendering block.

To improve the rendering effect, the receptive field is expanded utilizing the dilated convolution [26] in the style rendering blocks, thus, identity rendering in various receptive fields is achieved by continuously using the style rendering blocks. The entire style rendering process is completed by 4 style rendering blocks. The AdaIn affine parameters required by style rendering blocks are obtained by mapping the identity feature $z_{id}$ extracted by the identity encoder module. Thus, the adjusted code $z_{u}$ has rationality of identity. Similarly, the code $z_{au}$ of the reconstruction pipeline also possesses the reasonable identity, due to it uses the same identity transfer operation as completion pipeline.

2.4. Attention Feature Fusion Module

After identity transfer utilizing the identity transfer module, the codes $z_{u}$ and $z_{au}$ are obtained with reasonable identity. However, AdaIn in the identity transfer module will globally adjust the feature map during transfer, thus, it leads to the loss of background information. Therefore, an attention feature fusion module is designed in this paper to combine the undamaged features of background information before transfer and the code with reasonable identity to restore the background information of face images. For instance, $f_{m}$ and $f_{u}$ obtained by $z_{u}$ sampling are fused in the case of completion pipeline.

Figure 4 represents the structure of the attention feature fusion module. First, the attention score $\delta$ is calculated utilizing $f_{u}$ (Equation (10)). Then, $f_{m}$ and $f_{u}$ are fused under the guidance of $\delta$ (Equations (11)–(14)). Ultimately, $\gamma_{m}$ and $\gamma_{u}$ are spliced and then the fusion feature $f_{g}$ is obtained through a residual block(Equation (15)). Similarly, the fusion feature $f_{rec}$ is obtained by fusing $f_{ma}$ and $f_{au}$ using this module.

$$\delta_{ij} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, (s_{ij}) = P(f_{ui})^{T}Q(f_{ui})$$

where, $N$ denotes the number of pixels in $f_{ui}$, and $P$ and $Q$ are $1 \times 1$ convolutions.
where, $O$ represents a $1 \times 1$ convolution, and $\gamma_m$ & $\gamma_u$ denote equilibrium parameters, $S$ represents a residual block.

**Figure 4.** The attention feature fusion module. The module contains three $1 \times 1$ convolutional layers, O, P, Q, to extract features pixel by pixel.

To verify the effectiveness of the attention feature fusion module, the additive operation is imposed for the multiple channels of the feature map $f_u$ and $f_g$ before and after fusion in this paper, for which the visualized results are represented in Figure 5. It is observed that the information before fusion is mainly concentrated in the face area, however, the background information is restored after fusion indicating that the module can effectively fuse the features.

**Figure 5.** The feature maps before and after attention feature fusion. x and y represent the coordinates of the feature map, and the color bar represents the range of the feature map values.

2.5. Loss Function

According to the respective tasks of the generation pipeline and the reconstruction pipeline, in designing their loss function, it is considered that the reconstructed images should have a great similarity with the reference images, and the generated images should be real and natural with reasonable semantics.
First, reconstruction loss is introduced for the reconstruction pipeline to decrease the pixel-level difference between $I_{rec}$ and $I_a$. There is no complete face image information in the completion pipeline, hence, erroneous training results will be obtained when using the reconstruction loss forcibly. Therefore, spatially discounted reconstruction loss [12] is introduced, and the spatial discounted weight $W_{sd}$ is adjusted to apply a weaker pixel-level supervision to pixels farther from the nearest known pixel. The reconstruction loss and spatial reconstruction loss can be written as:

$$L_{app}^{g} = \sum_i \|I_{rec} - I_a\|_1$$  \hspace{1cm} (16)

$$L_{app}^{g} = \sum_i \|W_{sd} \times (I_g - I_a)\|_1$$  \hspace{1cm} (17)

$$W_{sd} = \gamma l$$  \hspace{1cm} (18)

where, $\|\cdot\|_1$ denotes the $L_1$ norm, $W_{sd}$ represents the weight of each pixel, $l$ is the distance of the pixel to the nearest known pixel. $\gamma$ is set to 0.9 in all experiments.

Then, the perceptual loss [27] is introduced for the reconstruction pipeline to decrease the perceptually difference between $I_{rec}$ and $I_a$. The output distance of the reconstructed image $I_{rec}$ and the reference image $I_a$ at different layers of the pre-training network is penalized to enhance their perceptual similarity. The perceptual loss is determined utilizing the feature map $\phi_i(\cdot)$ output by the pre-trained VGG19 network on the ImageNet dataset [28], as follows:

$$L_{pc} = \sum_i \frac{1}{N} \times \|\phi_i(I_g) - \phi_i(I_a)\|_1$$  \hspace{1cm} (19)

where, $\phi_i(I_g)$ and $\phi_i(I_a)$ denote the feature map corresponding to $I_g$ and $I_a$, respectively.

Ultimately, to decrease the difference in data distribution between the generated image and the real image, and enhance the visual quality of the generated image, the discriminant loss is introduced for the two pipelines. The discriminant loss for the completion pipeline is designed based on LSGAN [29] (Equation (20)), which could effectively improve the authenticity of the completed image $I_g$. Inspired by [30], the LSGAN-based discriminant loss is deformed (Equation (21)) for the reconstruction pipeline. Such a design can encourage the original features $D_2(I_a, L_a)$ and reconstructed features $D_2(I_{rec}, L_a)$ in the discriminator to be close together, so it can effectively enhance the semantic similarity between $I_{rec}$ and the reference image $I_a$ while ensuring the authenticity of the reconstructed image $I_{rec}$.

$$L_{adv}^{D_1} = \|D_1(I_g, L_g) - 1\|_1$$  \hspace{1cm} (20)

$$L_{adv}^{D_2} = \|D_2(I_{rec}, L_a) - D_2(I_{rec}, L_a)\|_1$$  \hspace{1cm} (21)

The total losses of the generation pipeline and the reconstruction pipeline are defined as follows:

$$L_g = \lambda_{app} \times L_{app}^{g} + \lambda_{adv} \times L_{adv}^{D_1}$$  \hspace{1cm} (22)

$$L_r = \lambda_{app} \times L_{app}^{r} + \lambda_{adv} \times L_{adv}^{D_2} + \lambda_{pc} \times L_{pc}$$  \hspace{1cm} (23)

where, $\lambda_{app}$, $\lambda_{adv}$, and $\lambda_{pc}$ denote the weighting factors.
3. Results and Discussion

3.1. Implementation Details

Our proposed model is implemented in PyTorch v1.2.0. The model is optimized using Adam optimizer [31] with decay index $\beta_1 = 0, \beta_2 = 0.9$, and learning rate $l_r = 0.0001$. The weights $\lambda_{rec}, \lambda_{adv}$ and $\lambda_{pc}$ of the loss function are set to 1, 0.1 and 0.1, respectively. On a single NVIDIA 2080Ti (11 GB), we trained our model on CelebA-HQ for three days with a batch size of 4. The convergence processes of the loss function and the test set evaluation metrics are shown in Figures 6 and 7. We found that the model achieved the best results at the 38th epoch.

![Figure 6. The curve of loss function convergence.](image)

![Figure 7. The curve of evaluation metrics convergence.](image)

3.2. Completion Consistency

A comparison is made between RG-DP-FICN proposed in this paper and generative image inpainting with contextual attention (CA) [12], generative landmark guided face inpainting (LaFIN) [20], and pluralistic image completion (PIC) [15] on the same database CelebA-HQ. Notice that for CA, LaFIN and PIC, the pre-trained models on the CelebA-HQ are given, these methods are tested utilizing the code and its network weights provided by the original authors.

To visually demonstrate the superiority of the face image completion approach proposed in this paper, a qualitative evaluation is performed for the completion results of this approach, CA, LaFIN, and PIC (Figure 8). The completion result of PIC and CA are essentially average faces based on the entire dataset, due to their identity are uncertain. Therefore, their completion results can possess certain realism, however, the identity is clearly different from the ground-truth. For LaFIN, the landmarks is introduced as prior knowledge of topological structure, nevertheless, its completion result still has no reasonable identity owing to the lack of identity information guidance, particularly manifests as the different contour of eyes, eyebrows, nose from that in the original image. For RG-DP-FICN, the reference image of the same identity is introduced to guide the completion process. Utilizing the identity transfer module for transferring the identity of the reference image to the damaged image, the identity features obtained by decoupling and the posture features of the damaged image are fused.
Moreover, the background information lost during identity transfer is restored by the attention feature fusion module. In conclusion, RG-DP-FICN can generate the completed images with reasonable topological structure and identity compared to the above three schemes.

![Image 1](a) SSIM: 0.935  
![Image 2](b) SSIM: 0.864  
![Image 3](c) SSIM: 0.893  
![Image 4](d) SSIM: 0.859  
![Image 5](e) SSIM: 0.910

Figure 8. The qualitative evaluation on CelebA-HQ datasets: (a) masked input; (b) ground-truth; (c) results of CA; (d) results of PIC; (e) results of LaFIn; (f) results of RG-DP-FICN.

To objectively demonstrate the superiority of the face image completion approach proposed in this paper, the completion results of this approach, CA, LaFIn, and PIC are quantitatively assessed. The quantitative evaluation results are represented in Table 1. As can be seen from the numbers in Table 1, LaFIn is superior over PIC and CA in most cases, as it employs the landmark information to guide completion process. PSNR and SSIM of RG-DP-FICN are far higher than those of other schemes, while L1 loss is lower compared to the other schemes. It is inferred that in comparison with the existing advanced schemes, the RG-DP-FICN proposed in this paper has a stronger capability of face image completion.

| Method      | PSNR | SSIM   | L1 Loss(%) |
|-------------|------|--------|------------|
| CA [12]     | 24.39| 0.863  | 2.88       |
| PIC [15]    | 24.86| 0.871  | 2.65       |
| LaFIn [20]  | 26.22| 0.905  | 2.37       |
| Our method  | 28.44| 0.938  | 1.58       |

Table 1. The quantitative results on the CelebA-HQ test set consisting of 1000 face images.
3.3. Completion Diversity

To further validate the model’s identity transfer capabilities, the reference face images of different identities are chosen for the damaged images utilizing the center mask (Figures 9 and 10). Covering the face attributes completely, the completion results will represent obvious differences in the position and contour of the eye and the shape of the nose based on the different reference face images. It can be seen that the completion network in this paper effectively guided the identity of the completion result.

![Figure 9](image9.png)

**Figure 9.** The completion results of a masked IMAGE6 using different reference face images: (a) masked input; (b) ground-truth; (c) reference face images; (d) results of RG-DP-FICN.

![Figure 10](image10.png)

**Figure 10.** The completion results of a masked IMAGE8 using different reference face images: (a) masked input; (b) ground-truth; (c) reference face images; (d) results of RG-DP-FICN.
4. Conclusions

Considering the problem of identity uncertainty in the results of face image completion, RG-DP-FICN is proposed in this paper. First, an identity transfer module is designed to realize the identity guidance of face image completion results. Then, an attention feature fusion module is designed to effectively restore the background information of the face image. In comparison with various advanced approaches, the completion results of RG-DP-FICN are more real and natural in the subjective visual effect and possessed significantly enhanced objective pixel-level similarity and overall similarity. Hence, it effectively solves the above problem.

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

Funding: This study was supported by the National Natural Science Foundation of China (Grant No. 61972282), and by the Opening Project of State Key Laboratory of Digital Publishing Technology (Grant No. Cndplab-2019-Z001).

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

References

1. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics And Interactive Techniques, Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
2. Elad, M.; Starck, J.L.; Querre, P.; Donoho, D.L. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). Appl. Comput. Harmon. Anal. 2005, 19, 340–358. [CrossRef]
3. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 2009, 28, 24. [CrossRef]
4. Bertalmio, M.; Vese, L.; Sapiro, G.; Osher, S. Simultaneous structure and texture image inpainting. IEEE Trans. Image Process. 2003, 12, 882–889. [CrossRef] [PubMed]
5. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 2004, 13, 1200–1212. [CrossRef] [PubMed]
6. Huang, J.B.; Kang, S.B.; Abuja, N.; Kopf, J. Image completion using planar structure guidance. ACM Trans. Graph. (TOG) 2014, 33, 1–10. [CrossRef]
7. Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D.B.; Sen, P. Image melding: Combining inconsistent images using patch-based synthesis. ACM Trans. Graph. (TOG) 2012, 31, 1–10. [CrossRef]
8. Hays, J.; Efros, A.A. Scene completion using millions of photographs. ACM Trans. Graph. (TOG) 2007, 26, 4. [CrossRef]
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. 2014. Available online: https://papers.nips.cc/paper/2014/file/5ca3e9b122618f06494c97b1afccf3-Paper.pdf (accessed on 16 November 2020).
10. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
11. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. ACM Trans. Graph. (ToG) 2017, 36, 1–14. [CrossRef]
12. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
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 International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4471–4480.

15. Zheng, C.; Cham, T.J.; Cai, J. Pluralistic image completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1438–1447.

16. Li, Y.; Liu, S.; Yang, J.; Yang, M.H. Generative face completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3911–3919.

17. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* 2019, arXiv:1901.00212.

18. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-aware image inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5840–5848.

19. Song, L.; Cao, J.; Song, L.; Hu, Y.; He, R. Geometry-aware face completion and editing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27–28 January 2019; Volume 33, pp. 2506–2513.

20. Yang, Y.; Guo, X. Generative Landmark Guided Face Inpainting. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Nanjing, China, 16–18 October 2020; pp. 14–26.

21. Jo, Y.; Park, J. SC-FEGAN: Face Editing Generative Adversarial Network with User’s Sketch and Color. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1745–1753.

22. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* 2017, arXiv:1710.10196.

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. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Okinawa, Japan, 16–18 April 2019; pp. 7354–7363.

25. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.

26. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* 2015, arXiv:1511.07122.

27. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, The Netherlands, 11–14 October 2016; pp. 694–711.

28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

29. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.

30. Bao, J.; Chen, D.; Wen, F.; Li, H.; Hua, G. CVAE-GAN: fine-grained image generation through asymmetric training. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2745–2754.

31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* 2014, arXiv:1412.6980.

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

© 2020 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 (http://creativecommons.org/licenses/by/4.0/).