electronics

MDPI

Article

Transformer-Based GAN for New Hairstyle Generative Networks

Qiaoyue Man [1], Young-Im Cho [1,*], Seong-Geun Jang [1] and Hae-Jeung Lee [2,*]

[1] Department of Computer Engineering, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si 461-701, Korea; manqiaoyue@gmail.com (Q.M.); zzangga97@gachon.ac.kr (S.-G.J.)

[2] Department of Food & Nutrition, College of Bionano Technology, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si 13120, Korea

[*] Correspondence: yicho@gachon.ac.kr (Y.-I.C.); skysea@gachon.ac.kr (H.-J.L.);
Tel.: +82-31-750-5800 (Y.-I.C.); +82-31-750-5968 (H.-J.L.)

Abstract: Traditional GAN-based image generation networks cannot accurately and naturally fuse surrounding features in local image generation tasks, especially in hairstyle generation tasks. To this end, we propose a novel transformer-based GAN for new hairstyle generation networks. The network framework comprises two modules: Face segmentation (F) and Transformer Generative Hairstyle (TGH) modules. The F module is used for the detection of facial and hairstyle features and the extraction of global feature masks and facial feature maps. In the TGH module, we design a transformer-based GAN to generate hairstyles and fix the details of the fusion part of faces and hairstyles in the new hairstyle generation process. To verify the effectiveness of our model, CelebA-HQ (Large-scale CelebFaces Attribute) and FFHQ (Flickr-Faces-HQ) are adopted to train and test our proposed model. In the image evaluation test used, FID, PSNR, and SSIM image evaluation methods are used to test our model and compare it with other excellent image generation networks. Our proposed model is more robust in terms of test scores and real image generation.

Keywords: face detection; convolutional neural network; generative adversarial networks; transformer; image fusion

check for updates

Citation: Man, Q.; Cho, Y.-I.; Jang, S.-G.; Lee, H.-J. Transformer-Based GAN for New Hairstyle Generative Networks. *Electronics* **2022**, *11*, 2106. https://doi.org/10.3390/electronics11132106

Academic Editor: Ahmed Abu-Siada

Received: 8 June 2022
Accepted: 2 July 2022
Published: 5 July 2022

1. Introduction

Good appearance, especially with regard to hairstyle, holds considerable importance for both women and men. A proper hairstyle can greatly improve a person's appearance. Once it becomes symbolic, an inappropriate hairstyle can also ruin one's confidence. In the event of a bad haircut, a person may require considerable time to grow sufficiently long hair to try a new hairstyle. Therefore, switching to a new hairstyle is a crucial decision. Moreover, before getting the actual haircut done, determining which new hairstyle would suit the person is very challenging. In this context, we studied the use of artificial intelligence to automatically and naturally fuse and generate new and suitable hairstyles. To naturally perform fusion to generate new hairstyle features, the face part is first detected [1], and the new hairstyle is subsequently matched. The traditional method [2] forcibly superimposes other people's hairstyles on one's avatar image by cutting images [3]. The hairstyles of different people do not match; therefore, they cannot be accurately integrated with the faces. Failure to consider the global consistency of an image results in significant synthesis artifacts; even though each part is composited in a highly realistic manner, different areas of the image can appear disjointed. For the generated image to appear reasonable and natural, we need to calculate the image details. While retaining the features of the original face image, carefully integrating the transition area between the new hair and the face reduces unnecessary artifacts and finally achieves our goal of generating realistic and natural face images with new hair.

Recently, with the rapid development of deep learning technology in several fields, outstanding achievements have been made in scientific research and practical applications. Among them, a Convolutional Neural Network (CNN), as a type of feedforward deep network, has demonstrated a strong ability in image feature learning and expression and has provided excellent performance in large-scale image processing [4,5]. Furthermore, Generative Adversarial Networks (GANs) proposed by Goodfellow et al. [6] have been widely used in the field of computer vision owing to their ingenious game adversarial learning mechanism and great potential for fitting data distribution. These research results greatly complement the image vision. Deep learning networks have overcome the shortcomings of traditional methods in understanding image semantics and have solved the semantic gap between low-level image features and high-level semantics to a certain extent. Simultaneously, researchers have introduced the attention mechanism [7] of the Natural Language Processing (NLP) field into computer vision tasks, which enhances the ability of image feature extraction. These excellent methods have gradually placed the deep learning technology on the forefront of the computer vision field.

Traditional GANs hairstyle generation tasks are trained on large amounts of data, and image features are mixed together to form a single synthetic image [8,9]. Although these methods can generate visually plausible image structures and textures, they often produce distorted structures or blurred textures that are inconsistent with the surrounding regions, making this task extremely challenging, especially the image generation task for a human face [10]. The main reason behind this can be explained as follows; the visual properties of the different parts of the image are not independent of each other. These problems can be solved by using (i) light, geometry, and other patches [11,12] for optimization; (ii) global and local discriminative optimization [13]; (iii) context attention [14–16] for optimizing the generation of image details and suppressing blur; (iv) transformers [17] for reconstructing the appearance priors by supplementing details such as textures. In the hairstyle generation task, the appearance of hair is largely influenced by ambient light and the colors transmitted from the underlying face, clothing, and background. Additionally, the geometry of a person's head and shoulders affects the shading and hair structure. Other challenges include blurring the background and detachment of facial areas exposing new parts of the face, such as the ears, forehead, or jawline. Although recent studies on GANs have been able to synthesize realistic hair or faces, combining them into a single, coherent, and plausible image remains difficult.

In this study, we propose a novel transformer-based GAN for facial hairstyle image fusion by introducing a GAN-based semantic alignment step where occluded regions in an image are filled with semantically correct image content. Artifacts caused by transparency, reflections, or interactions of hair with faces are less noticeable when aligned with task-relevant semantic regions, such as hair. Our proposed method is superior at preserving details and encoded spatial information for the accurate and natural fusion of faces and hairstyles. In addition to our method's ability to transfer visual attributes, including specific details such as wrinkles, from multiple reference images, we perform image blending in the latent space. This enables us to synthesize coherent images. Our method avoids blending artifacts present in other methods and finds images that are globally consistent.

Following is a brief summary of the contributions of this work.

1. We propose a novel hairstyle generation composite network, Face Transformer Generative Hairstyle networks (FTGH), to overcome the single generation network and the abnormal situation of generating new hairstyle images of people by adding modules for extracting face masks and segmenting face regions to enhance the realistic effects of GAN image generation.
2. We designed a GAN image generation network incorporating transformers. To avoid the problem of insufficient accuracy of images generated by Vision Transformer (ViT), ResNet was used for image generation in the generator part and the transformer method was used to discriminate the generated images in the discriminator part.

3. We used the open-source datasets CelebA-HQ and FFHQ to train and validate our proposed model. Verification with various image evaluation standards revealed that our method is more robust than the other existing methods.

2. Related Work

Generative adversarial networks. GANs were proposed by Goodfellow et al. Owing to their emergence in the field of image generation and their realistic output, GANs have attracted increasing attention for use in image inpainting and image generation tasks, especially in the editing of face images. However, the original formulation of GAN suffers from training instability and mode collapse. Considerable research has been conducted to address these issues using various methods; for example, Zhu et al. proposed CycleGAN [18] to use cycle consistency loss to overcome the lack of paired training data. Choi et al. [19] proposed StarGAN, a unified model architecture that allows simultaneous training of multiple datasets with different domains in a single network, thereby improving image generation quality. Gu et al. [20] proposed mGANprior to generate multiple feature maps using multiple latent codes at an intermediate layer of the generator and combine them with adaptive channel importance to recover the input image. This overparameterization of the latent space significantly improves the image reconstruction quality. NVIDIA researchers proposed styleGAN [21,22], which is a series network model. This style-based generator method provides good performance in face static image restoration and hairstyle fusion and excellent performance in the field of video image restoration. Lin et al. [23] used a GAN approach combining ResNet and attention mechanism to generate high-quality face images for visual face information protection. Adela et al. [24] proposed, based on styleGAN face embedding method to generate new hairstyles, avoiding shadows at the intersection of facial contours and hair. As a result of these developments, GANs have achieved remarkable results in various tasks.

Visual Transformer. The transformer abandons the traditional CNN and RNN, and the entire network structure uses only the attention mechanism. NLP has led to rapid progress in a variety of natural language tasks; recently, NLP is being widely used in the field of computer vision with its appearance on some tasks previously dominated by CNNs. Among them, the seminal work, ViT [25], proposed a pure transformer architecture for image classification and presentation, with great potential in vision tasks. Carion et al. [26], an ensemble-based global loss, designed a DETR network for global segmentation, and the network performance significantly outperformed competing baselines. Srinivas et al. [27] proposed BoTNet, a transformer model incorporating the backbone architecture of self-attention and applied it to multiple computer vision tasks including image classification, object detection, and instance segmentation. Wu et al. [28] proposed Pale Transformer with pale self-attention (PS-Attention) that performs self-attention within pale regions. Compared with global self-attention, PS-Attention can significantly reduce computational and memory costs while capturing richer contextual information.

Transformer-based GANs. Recently, the research community has begun exploring the use of transformers for image-generation tasks with the hope that the increased expressiveness can benefit the generation of complex images. For example, Jiang et al. [29] proposed a TransGAN comprising a memory-and-transformer-based generator and a transformer-based patch-level discriminator that progressively increases the feature resolution while reducing the embedding size. Lee et al. [30] proposed ViTGAN, which integrated the ViT architecture into a GAN and improved spectral normalization to enhance Lipschitz continuity, achieving fairly-high performance. Wang et al. [31] proposed a convolution-free pyramid vision transformer (PVT) that overcomes the difficulty of porting the transformer to various dense prediction tasks, thereby reaching or even surpassing the performance of traditional CNNs. In this study, we propose a novel realistic hairstyle generation network for facial hairstyle synthesis image-editing tasks. In the proposed network, an efficient segmentation module is used to detect the face image in advance and extract the mask. For hairstyle generation, we use the transformer-based GAN. Because transformer efficiency is limited in image generation, ResNet is used in the generator to generate a new hairstyle and the transformer is used as a discriminator to focus on the generated area and assess the quality of the generated images. Finally, a realistic facial image with a new hairstyle is generated.

3. Methods

In this section, we provide a systematic description of the network structure of our model. The framework of the proposed FTGH network model is shown in Figure 1. This framework comprises two modules: Face segmentation (F) and Transformer Generative Hairstyle (TGH) modules.

The most important thing in the hairstyle generation task is that the generated hairstyle and the original face have overall consistency to avoid the appearance of artifacts, which requires accurate detection and segmentation of the face area and the hairstyle area. Therefore, we add the face segmentation module before generating the hairstyle to enhance the feature extraction ability and provide more accurate feature information for the generation network. The face segmentation module is responsible for detecting face features and extracting face area maps and face mask feature maps. Furthermore, in the TGH module, the transformer-based GAN calculates the hairstyles and people multiple times. After the facial features are fused, a new face image with a suitable and natural hairstyle is generated.

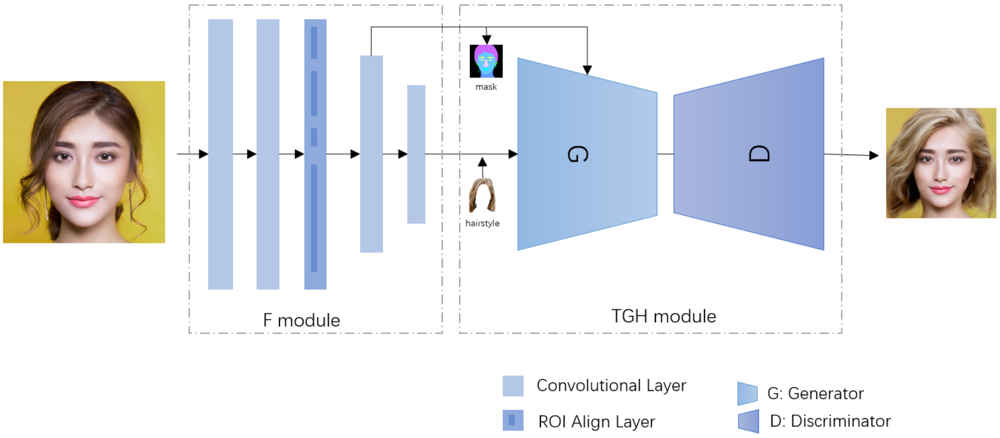

Figure 1. TGH Framework: Our proposed network comprises face segmentation (F) and transformer generative hairstyle (TGH) modules. The input image is extracted from the F module to extract the mask and detail features of the hairstyle and face. In the TGH module, the hairstyle image is referred to, and the face image with the new hairstyle is generated with the assistance of the mask image.

3.1. Face Segmentation (F) Module

The face segmentation (F) module is similar to the Mask R-CNN [32], as shown in Figure 2. Mask R-CNN is a composite network framework comprising RESNET-FPN, region proposal network, and fully convolutional network (FCN). However, in our task, multi-tasking and multi-target detection is not involved; only mask and facial features are required to be extracted. Therefore, we redesigned and optimized the network to make the face segmentation module more concise and efficient. ResNet-50 is more concise and has the same excellent performance. Additionally, it removes unnecessary fully connected networks and draws on the RoI (RoI Align) extraction method; therefore, it is used as the basic network for feature extraction.

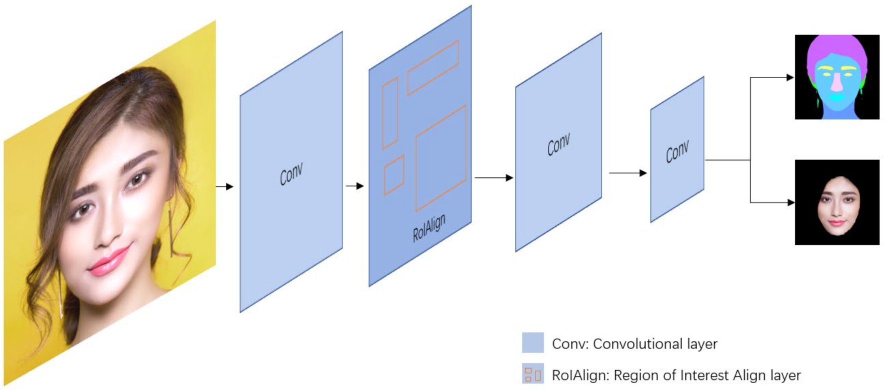
Conv
RoIAlign
Conv
Conv
Conv: Convolutional layer
RoIAlign: Region of Interest Align layer

Figure 2. Face segmentation (F) module.

The precise pooling layer of the RoI alignment is used to segment and extract facial features at the pixel level. In the traditional RoI pooling layer, the quantization operation leads to a deviation between the obtained features and the RoI. As shown in Figure 3, the multiple quantization operation deviations become larger, and this difference can have a negative impact on the predicted pixel mask. However, the RoI alignment retains the original feature data without quantification, and the obtained RoI features are more authentic. It is conducive to semantic recognition and detailed mask segmentation of face and hairstyle regions and provides help for the uniformity and naturalness of the generated images.

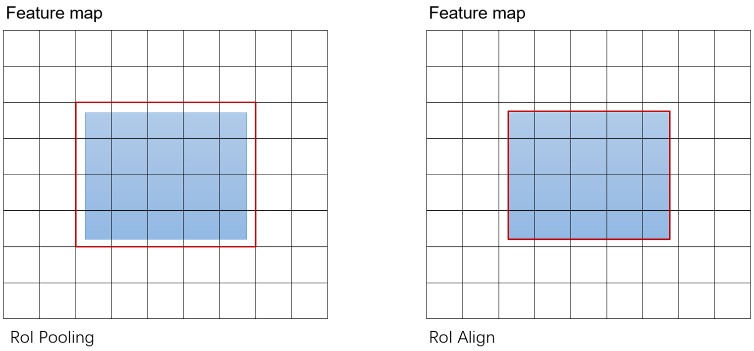

Figure 3. Comparison of RoI Pooling and RoI Align calculation methods.

3.2. Transformer Generative Hairstyle (TGH) Module

The internal structure of a GAN generator usually uses a deconvolution network (DCNN), which can be considered to be a stack of deconvolution layers. However, in the process of image generation, the input condition variables must pass through several layers to affect the final generated output. Simple convolutional networks are unable to generate realistic images. Designing generators based on the ViT architecture is not trivial, and the generated images are worse than those of CNN-based generators. Therefore, as shown in Figure 4, we adopted a ResNet-based CNN as the generator.

In the generator, we used ResNet-50 for image generation. We additionally tested ResNet-32, ResNet-101, and other networks with different layers; however, they did not provide significant performance gains. The ResNet in the generator has downsampling, residual blocks, and upsampling. The down-sampling layer adopts a strided convolution calculation without a pooling layer. The residual blocks do not change the width or height of the activation map. There are equal number of downsampling and upsampling layers. The downsampling computation uses a dilated convolution operation [33] to increase the receptive field size without applying subsampling or adding several convolutional layers. The generation process can be denoted as:

$$G(x) = f(g(h_x) + h_x) \tag{1}$$

where h_x is the input image, $G(x)$ is the final output, and $g(h_x)$ is the residual image to be learned by the generator.

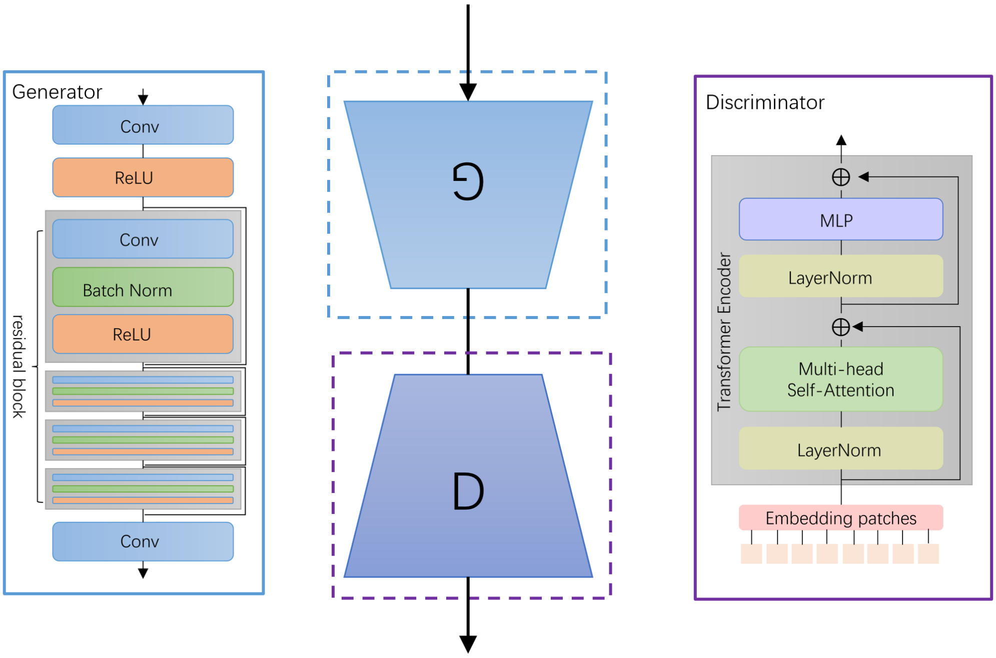

Figure 4. Transformer generative hairstyle (TGH) module. The module consists of a generator (G) and a discriminator (D). The generator uses the ResNet network, and the discriminator uses a transformer for image patch attention.

In the discriminator, unlike the generator used to generate pixel-exact images, the task of the discriminator is to distinguish between real and fake images. We divide the image generated by the generator into multiple patches, where each patch can be considered to be a "word". Unlike classifier image recognition tasks that focus on semantic differences, the discriminator is used to distinguish synthetic images from real ones by computationally simple tasks that focus on image details. Therefore, local visual cues have a greater influence on the discriminators. For the internal design of the transformer of the discriminator, we adopted the original transformer scheme, combining multi-head self-attention and Multilayer Perceptron (MLP) to determine whether the images generated by the generator are accurate. Transformers are more advantageous in image detail detection, which is of great help in the artifact suppression of hair regions and facial gaps in the hairstyle generation network. We employed a balanced learning rate [34] in the attention module and MLP to address the slow convergence of the transformer block in the discriminator.

Multi-Head Attention denoted as:

$$\begin{gathered}\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)\end{gathered} \tag{2}$$

where Q, K and V denoting matrices packing together sets of queries, keys and values, respectively. W^Q, W^K and W^V denoting projection matrices that are used in generating different subspace representations of the query, key and value matrices. W^O denoting a projection matrix for the multi-head output.

The TGH image generation process is a game between the discriminator D and the generator G, and the final equation can be denoted as:

$$\underset{G}{min}\underset{D}{max}V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{3}$$

where G is the generator, D is the discriminator, and z is the corrupted image.

4. Experiments

4.1. Datasets

We used the open-source datasets, CelebA-HQ [35] and FFHQ [36], to train and validate the proposed FTGH hair fusion generative network model and randomly selected 80% data for training and 20% for testing. CelebA-HQ is a high-definition image version dataset of CelebA, which contains 30,000 face images with a 1024 × 1024 resolution. FFHQ is a commonly used dataset for generating high-resolution images. It contains 70,000 high-quality PNG images at a resolution of 1024 × 1024 with considerable variation in age, ethnicity, and image background. Additionally, it has a good coverage of accessories, such as eyeglasses, sunglasses, and hats.

4.2. Implementation Details

During training, the algorithm flow is presented in Algorithm 1, we used the Adam solver [37] with reference to TTUR. Since our proposed hairstyle generation model uses G and D with different internal structures, the same learning rate cannot meet the training requirements. The important parameters of training are shown in Table 1. After experimental verification, a higher learning rate will lead to a decrease in the quality of generated images, finally selected unbalanced learning rates in the generator and discriminator: $l_g = 5 \times 10^{-5}$ and $l_d = 2 \times 10^{-4}$, with standard GAN losses and *R1* gradient penalty to train our networks. In addition, the same number of training steps will also affect the training results. In model training, we found that increasing the number of training steps s_g of the generator will produce a more accurate image, and too large a difference between s_g and discriminator s_d will also reduce the accuracy of the generated image. Figure 5 shows the loss curves of the generator and discriminator trained for 200 epochs.

Table 1. Training parameters.

l_g	Generator learning rate
l_d	Discriminator learning rate
s_g	Generator training steps
s_d	Discriminator training steps

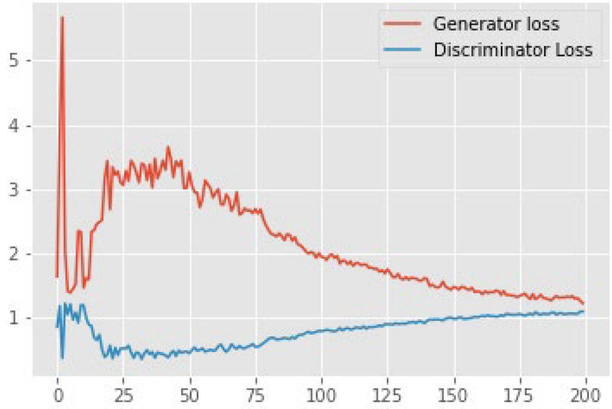

Figure 5. Training loss.

In network performance evaluation, a single test evaluation scheme cannot accurately evaluate the performance of the model; therefore, we use a variety of image generation quality evaluation methods: the Fréchet inception distance (FID), peak signal-to-noise ratio (PSNR), and SSIM (structural similarity index map) to measure the difference between the image generated by our proposed model and the real image.

FID: To calculate the real samples and generate the distance between them in the feature space, we first use the inception network to extract features and then use the Gaussian model to model the feature space. Finally, we solve the distance between the two features. A lower FID indicates a higher image quality and diversity. The following equation is used to calculate the FID.

$$\mathrm{FID}(x,g) = ||\mu_x - \mu_g||\frac{2}{2} + \mathrm{Tr}\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right) \tag{4}$$

PSNR: This is the ratio between the maximum value signal of the image and background noise. The larger the ratio, the higher the quality of the generated image. It is widely used as an image quality evaluation index in several image fields such as image super-resolution, compression, and denoising. It is calculated using the following equation.

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \tag{5}$$

SSIM: This is an index used to measure the similarity between two images, based on the following three aspects: brightness, contrast, and structure. The value range of the SSIM is [0,1]. The larger the value, the smaller the image distortion. It can be calculated using the following equation.

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}\cdot[c(x,y)]^{\beta}\cdot[s(x,y)]^{\gamma} \tag{6}$$

Algorithm 1 FTGH, using standard GAN losses and *R1* gradient penalty to train our networks. The number of steps to apply to the discriminator, *k*, is a hyperparameter, initialized *k* = 1.

```
for number of training iterations do
for k steps do
  • Sample batch of m noise samples {$z^{(1)}, \ldots, z^{(m)}$} from noise prior $p_g(z)$.
  • Sample batch of m noise examples {$x^{(1)}, \ldots, x^{(m)}$} from data generating distribution $p_{data}(x)$.
  • Update the discriminator by ascending its standard GAN losses and R1 gradient penalty:
      $\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right].$
end for
    • Sample batch of m noise samples {z (1), . . . , z(m)} from noise prior $p_g(z)$.
    • Update the generator by ascending its standard GAN losses and R1 gradient penalty:
      $\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$
```

4.3. Main Results

We used the FID image quality assessment metric to compare our method with other state-of-the-art convolution-based GAN methods. The proposed FTGH network achieved state-of-the-art FID scores, as presented in Table 2.

Table 2. Comparison of FID of different models.

Method.	**FID (↓)**
STGAN [38]	47.54
U-NetGAN [39]	43.99
MagGAN [40]	41.32
MASKGAN	37.55
LOHO [41]	35.50
FTGH (ours)	**21.72**

At the same time, we test different GAN internal generator and discriminator combinations, as presented in Table 3. In the GAN framework using the ResNet or Transformer network structure of the same generator and discriminator network combination, there is a certain gap between the performance and our proposed TGH model

Table 3. Comparison of FID of Different GAN.

Method	**FID (↓)**
G&D (ResNet)	29.2
G&D (Transformer)	44.7
TGH (G: ResNet/D: Transformer)	21.5

The image quality assessment method singly cannot accurately determine the performance of the image generation network. Therefore, we additionally adopt the PSNR and SSIM methods to test the performance of our proposed network. As presented in Table 4, our proposed network exhibits excellent performance.

Table 4. Comparison of PSNR and SSIM of different models.

Method	**PSNR (dB) (↑)**	**SSIM (↑)**
STGAN	17.92	0.72
U-NetGAN	18.55	0.75
MagGAN	20.43	0.78
MASKGAN	20.75	0.80
LOHO	22.28	0.83
FTGH (ours)	**30.10**	**0.92**

In previous studies on facial hairstyle generation, hard transitions such as copy and paste areas were usually used. Owing to the misalignment of images, several artifacts would be created between the fusion of hair and face, resulting in unnatural images. Our proposed FTGH network provides pixel-level face-mask information and facial features. In the image generation task shown in Figure 6, this spatial information can be used to find the hair and facial boundaries, perform new hairstyle fusion, and generate a clear and artifact-free natural image.

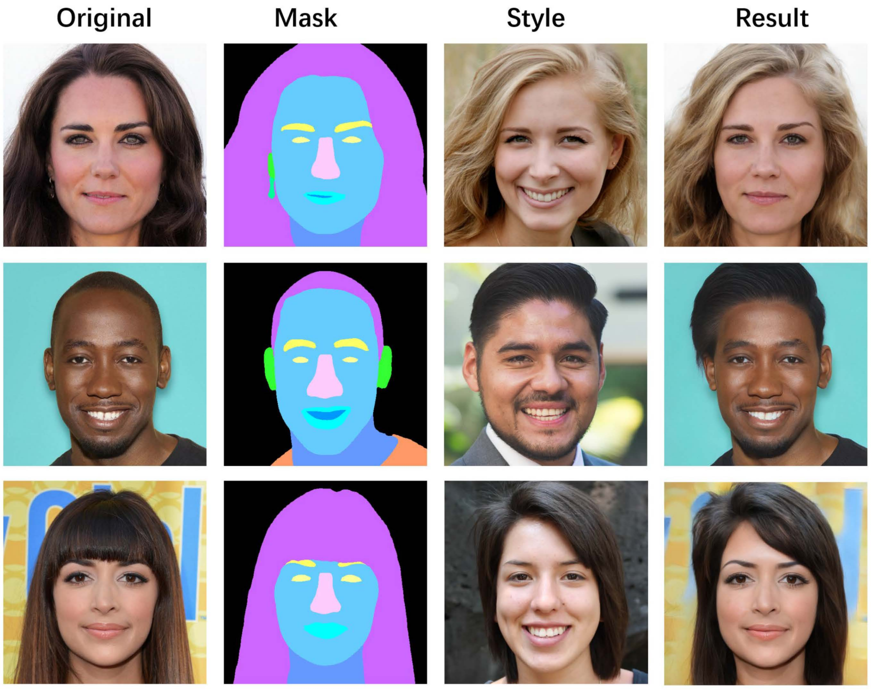

Figure 6. Examples of new hairstyles generated by the FTGH network.

A comparison between the results generated by our method and those by other image generation networks is presented in Figure 7. It can be observed that, although some methods can generate relatively clear face images with new hairstyles, the performance of these networks is limited regarding the generated hair and face details, with unnatural hairstyles, blurred edges, and facial feature changes. In summary, our proposed method generated new hairstyles that are clearer and more natural on the premise of preserving the original facial features.

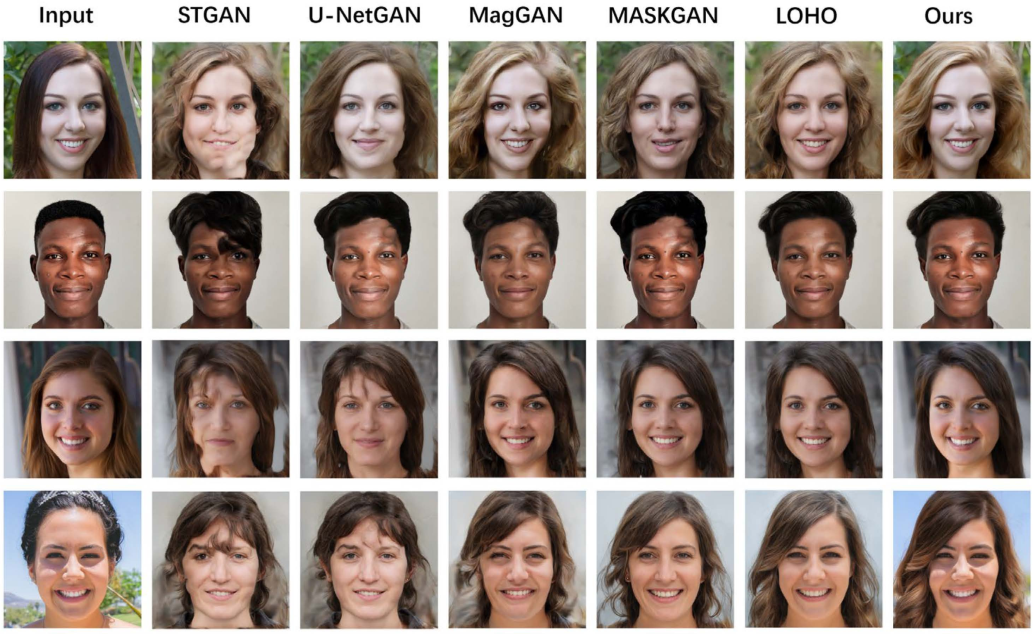

Figure 7. Comparison of the results of our proposed hair generation network model and those of other excellent network models.

We tested different hairstyles of female and male avatars. In female hairstyle generation tasks, hairstyles are more diverse, such as long, short, curly, and braided hair; hair covering the face and/or cheeks; and the wearing of earrings. This is a challenge for hairstyle-generation networks.

Figure 8 shows the proposed FTGH network which generates different hairstyles on different faces. Concurrently, in the male hairstyle generation task, there are fewer cases of facial occlusion because men prefer short hair. This is advantageous for the generation network and more conducive for generating clear and natural face images. As shown in Figure 9, our proposed network exhibits excellent image generation performance.

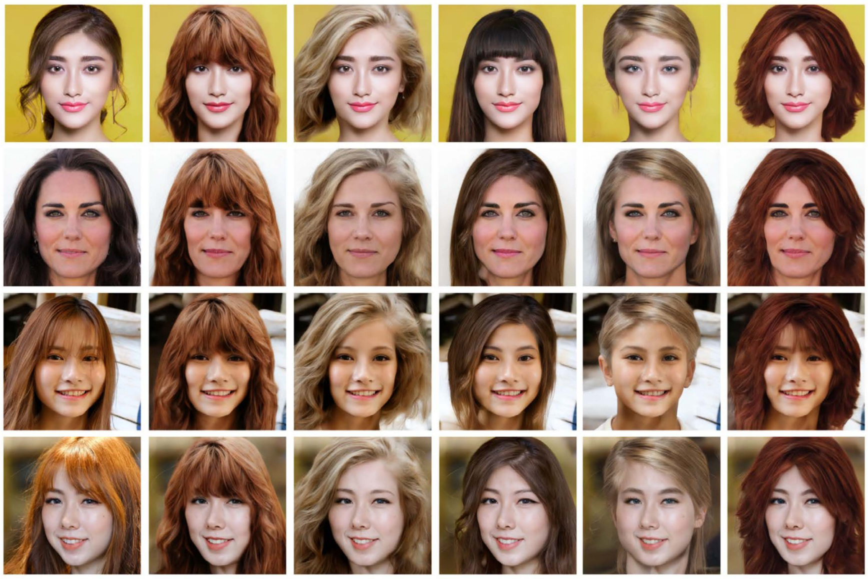

Figure 8. Comparison of different new hairstyles for women.

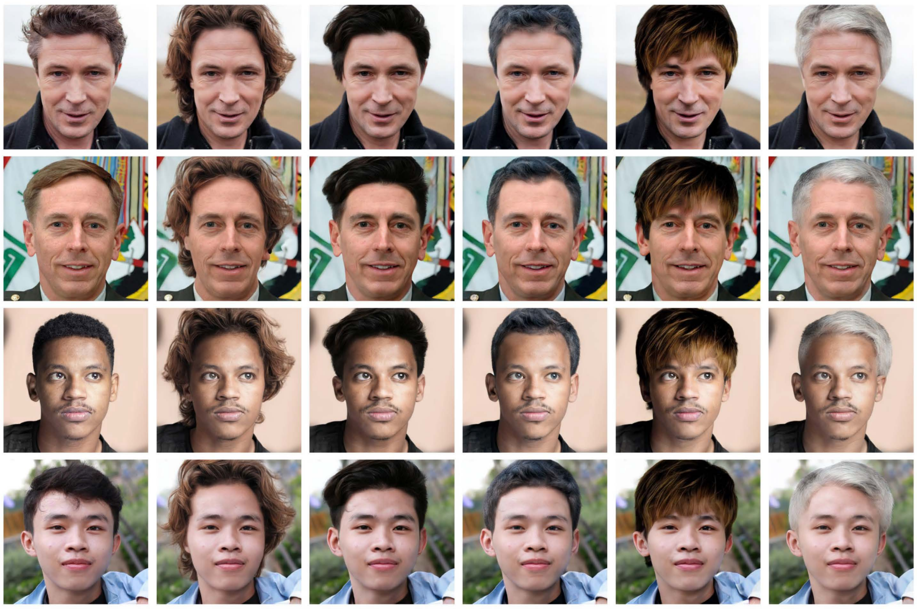

Figure 9. Comparison of different new hairstyles for men.

5. Discussion

Our proposed FTGH network for new hairstyle generation can generate new, clear, artifact-free, and natural hairstyles for both females and males and has excellent performance. The FTGH network uses the open-source datasets, CelebA-HQ and FFHQ, for training. Most of the images in the datasets use frontal unobstructed face images; therefore, during image testing, we observed that, in some specific face images, such as those wearing glasses, when face occlusion (hand, hand joints, etc.) is used for new hairstyle generation, our network provides limited performance, as shown in Figure 10. Specifically, for images showing faces wearing glasses, the generated face image of a new hairstyle and face parts is accurate; however, the restoration of the leg part of the glasses is limited when the face is occluded by the hand, and the restoration of hand details in the generated image is not sufficiently natural. In future research, we will continue to modify and improve our network to achieve more accurate and natural generation of face images with new hairstyles in various scenarios.

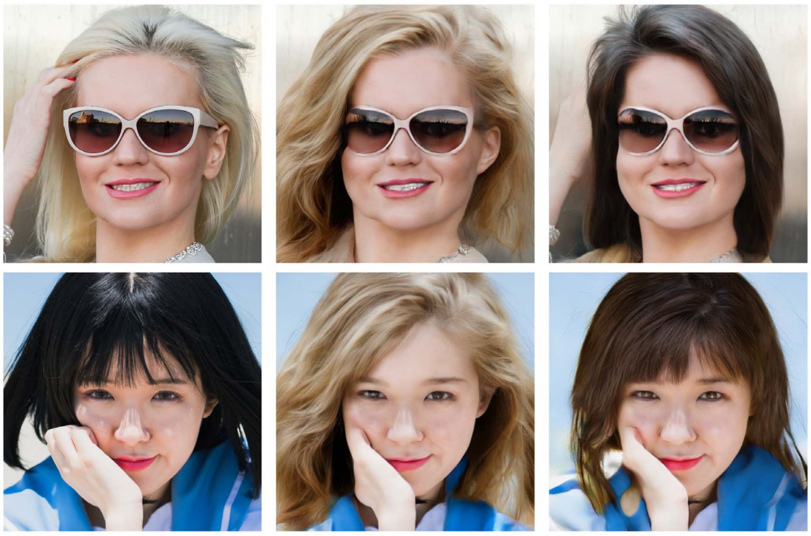

Figure 10. Face image with face occlusion.

6. Conclusions

We proposed a novel hairstyle generation network called the FTGH. The network model comprises two modules: the F and TGH modules. The F module is responsible for the accurate extraction of face and hairstyle area masks from the original image. The TGH module uses a combination of transformers and GANs to generate high-quality, high-resolution new hairstyles. The traditional GAN model offers limited performance in image generation, particularly in the image fusion generation of specific areas of an image. The proposed method uses a combination of multiple modules. First, the specific area of the image to be generated is confirmed, and subsequently, image generation is performed. For image generation, ResNet was used as the generator for the fusion generation of specific images, and transformers were used as the discriminator for focusing on specific regions generated and for discriminating the quality of the entire image generation. Our model is more robust than other excellent models in high-quality 1024 × 1024 resolution image tests provided by the CelebA-HQ and FFHQ open-source datasets.

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

Funding: This work supported by the Gachon University research fund of 2019 (GCU-2019-0794) and partly supported by the Gachon University research fund of 2019 (GCU-2019-0389).

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

References

1. Liu, Z.; Ping, L.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
2. Pasupa, K.; Sunhem, W.; Loo, C.K. A hybrid approach to building face shape classifier for hairstyle recommender system. *Expert Syst. Appl.* **2019**, *120*, 14–32. [CrossRef]
3. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
4. 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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
5. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
6. Ian, G.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. Advances in neural information processing systems. *arXiv* **2014**, arXiv:1406.2661.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. *arXiv* **2017**, arXiv:1706.03762.
8. Natsume, R.; Yatagawa, T.; Morishima, S. Rsgan: Face swapping and editing using face and hair representation in latent spaces. *arXiv* **2018**, arXiv:1804.03447.
9. Yin, W.; Fu, Y.; Ma, Y.; Jiang, Y.; Xiang, T.; Xue, X. Learning to Generate and Edit Hairstyles. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1627–1635.
10. Li, Y.; Liu, S.; Yang, J.; Yang, M. 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.
11. 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]
12. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Processing* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
13. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph. TOG* **2017**, *36*, 1–14. [CrossRef]
14. 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.
15. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Alexei, A.E. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2536–2544.
16. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.
17. Wan, Z.; Zhang, J.; Chen, D.; Liao, J. High-Fidelity Pluralistic Image Completion with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4692–4701.
18. Zhu, J.-Y.; Park, T.; Isola, P.; Alexei, A.E. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
19. Choi, Y.; Choi, M.; Kim, M.; Ha, J.; Kim, S.; Choo, J. Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
20. Gu, J.; Shen, Y.; Zhou, B. Image Processing Using Multi-Code Gan Prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3012–3021.
21. Tero, K.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of Stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
22. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 852–863.
23. Lin, J.; Li, Y.; Yang, G. FPGAN: Face de-identification method with generative adversarial networks for social robots. *Neural Netw.* **2021**, *133*, 132–147. [CrossRef] [PubMed]
24. Šubrtová, A.; Čech, J.; Franc, V. Hairstyle Transfer between Face Images. In Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 1–8.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
27. Srinivas, A.; Lin, T.-S.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.

28. Wu, S.; Wu, T.; Tan, H.; Guo, G. Pale Transformer: A General Vision Transformer Backbone with Pale-Shaped Attention. *arXiv* **2021**, arXiv:2112.14000. [CrossRef]
29. Jiang, Y.; Chang, S.; Wang, Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 14745–14758.
30. Lee, K.; Chang, H.; Jiang, L.; Zhang, H.; Tu, Z.; Liu, C. Vitgan: Training gans with vision transformers. *arXiv* **2021**, arXiv:2107.04589.
31. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
32. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
33. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
34. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
35. Lee, C.-H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards Diverse and Interactive Facial Image Manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.
36. 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.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; Wen, S. Stgan: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3673–3682.
39. Schonfeld, E.; Schiele, B.; Khoreva, A. A U-Net Based Discriminator for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8207–8216.
40. Wei, Y.; Gan, Z.; Li, W.; Lyu, S.; Chang, M.; Zhang, L.; Gao, J.; Zhang, P. Maggan: High-Resolution Face Attribute Editing with Mask-Guided Generative Adversarial Network. In Proceedings of the Asian Conference on Computer Vision, Online, 30 November–4 December 2020.
41. Saha, R.; Duke, B.; Shkurti, F.; Taylor, G.W.; Aarabi, P. Loho: Latent Optimization of Hairstyles via Orthogonalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1984–1993.