electronics

MDPI

Article

GAN-Based High-Quality Face-Swapping Composite Network

Qiaoyue Man [1], Young-Im Cho [1,*], Seok-Jeong Gee [1], Woo-Je Kim [2] and Kyoung-Ae Jang [3]

[1] Department of Computer Engineering, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si 13120, Republic of Korea; manqiaoyue@gmail.com (Q.M.); daetung@gachon.ac.kr (S.-J.G.)
[2] Department of Industrial Engineering, Seoul National University of Science and Technology, 232 Gongneung-ro, Nowon-gu, Seoul 01811, Republic of Korea; wjkim@seoultech.ac.kr
[3] Datatree Company, 1406, Bando Ivy Valley, 204, Gasan Digital 1-ro, Geumcheon-gu, Seoul 08502, Republic of Korea; kajang@datatree.kr
* Correspondence: yicho@gachon.ac.kr; Tel.: +82-31-750-5800

Abstract: Face swapping or face replacement is a challenging task that involves transferring a source face to a target face while maintaining the target's facial motion and expression. Although many studies have made a lot of encouraging progress, we have noticed that most of the current solutions have the problem of blurred images, abnormal features, and unnatural pictures after face swapping. To solve these problems, in this paper, we proposed a composite face-swapping generation network, which includes a face extraction module and a feature fusion generation module. This model retains the original facial expression features, as well as the background and lighting of the image while performing face swapping, making the image more realistic and natural. Compared with other excellent models, our model is more robust in terms of face identity, posture verification, and image quality.

Keywords: image processing; face fusion; convolutional neural network; generative adversarial network

check for updates

Citation: Man, Q.; Cho, Y.-I.; Gee, S.-J.; Kim, W.-J.; Jang, K.-A. GAN-Based High-Quality Face-Swapping Composite Network. *Electronics* **2024**, *13*, 3092. https://doi.org/10.3390/electronics13153092

Academic Editors: Enzo Pasquale Scilingo and Gemma Piella

Received: 17 June 2024
Revised: 27 July 2024
Accepted: 30 July 2024
Published: 5 August 2024

1. Introduction

In recent years, deep learning has significantly empowered the field of computer vision, particularly in the processing of digital images [1,2]. Through deep neural networks, especially convolutional neural networks, significant progress has been made in various tasks of computer vision, including image classification, target detection, image segmentation, and image generation. Among them, the rise of generative network models has shown great potential in image generation, image restoration, and data enhancement, and more and more researchers are focusing on this field, and the face-swapping task is one of the research directions. Face swapping refers to the seamless replacement of features from one face to another, maintaining the characteristics of the target image such as facial expression, pose, and background. The technique has a wide range of applications in areas such as portrait appearance modification, video compositing, film production, privacy protection, facial animation, and augmented reality. With the continuous development and improvement of the technology, this technology will bring more innovative application scenarios and possibilities.

Face swapping is a challenging task in computer vision, involving the transfer of a source face's identity to a target face while preserving the target's facial attributes (such as facial expression, head pose, and background lighting). Early face-swapping techniques relied primarily on traditional image processing techniques and manual editing. Techniques such as image alignment and fusion, image sharpening, and distortion required extensive manual adjustments to ensure natural and realistic results. This may include adjusting the transparency of the fused region, correcting mismatched features, and smoothing transitions. For instance, Bitouk et al. [3] developed an automatic face replacement system that uses face detection software to extract facial features, in the established face graphic library, select similar candidate images from a facial graphic library, overlapping, matching, and mixing after their coordinates, and adjusting the facial color and lighting to facilitate the replacement of the facial features of the environment fusion. Timothy et al. [4] designed an active appearance model to model a bunch of images containing control points or masks into another more convenient method to transform faces through feature coverage. Volker et al. [5] used 3D modeling to estimate the three-dimensional shape and texture and all related scene parameters based on a single image. Manual interaction is reduced to clicking on a set of about seven feature points to exchange facial features. Aseem et al. [6] designed an interactive computer-aided framework called digital photomontage, which uses graph-cut optimization and gradient domain fusion to synthesize graphic features. Despite these advancements, early techniques were limited by algorithm accuracy and required extensive manual processing, and their synthesized images usually have poor realism and naturalness in the face of complex facial feature processing, as well as inefficiency.

The advent of deep learning technology has brought revolutionary innovations and efficient performance advantages, greatly promoting the development of face-swapping tasks. These methods learn facial feature representations from large-scale face datasets using deep convolutional neural networks (CNNs) [7,8], achieving better results than traditional techniques. Iryna et al. [9] described face swapping as a style migration problem, training a CNN to transfer the appearance of a target identifier from an unstructured collection of photographs, by describing this face-swapping problem as a style migration problem, thus achieving the purpose of face-swapping. Yuval et al. [10] used a standard fully convolutional network, trained on a 3D face dataset for face feature alignment, segmentation, and 3D shape estimation, followed by Poisson blending [11] to merge the source faces into the target image. Li et al. [12] proposed the Attribute-Conditioned Face Swapping Network, using an Image Enhancement Network (IEN) to restore high-resolution images from low-resolution images and a Face Exchange Module (FEM) to swap the faces, and designed a multi-domain feature fusion module (MDFFM) to integrate the identity feature, context feature, IEN feature, and attribute vector to obtain the final image. Although the CNN-based face swap model solves the problems of automation and efficiency, the swapped faces still have problems such as distorted facial features, abnormalities, blurred images, and poor quality.

With the emergence of generative adversarial networks (GANs) and the fact that they can successfully generate realistic fake face images after extensive research, researchers have begun to try to apply this technology to face-swapping tasks. Conditional GANs (cGANs) are used to transform images depicting real data from one domain to another and have inspired a variety of facial re-enactment schemes. Among them, the DeepFakes project utilizes cGAN to perform face swapping in videos, making it widely available to non-experts and receiving a lot of public attention. Among the projects, DeepFaceLab [13] and Faceswap [14] are the most outstanding. DeepFaceLab provides the necessary tools and an easy-to-use way to perform high-quality face swapping. It also provides a flexible and loosely coupled structure for those who need to enhance their pipeline with other functions without writing complex boilerplate code. Faceswap can replace faces in images or videos, and its generation effect is also excellent. At the same time, it can manually modify and train a variety of data to achieve better face-swapping results. Yuval et al. [15] proposed the face-swapping generative adversarial network (FSGAN), which combines recurrent neural network (RNN) with Poisson optimization and Poisson mixing loss, using a face blending network to achieve a seamless blending of two faces while preserving the target skin color and illumination conditions. Xu et al. [16] proposed a simple feedforward network FaceController to generate high-fidelity faces. They used 3D prior technology to separate the face identity, expression, and background and then embedded all the information into the adversarial network through the identity-style module for image generation. However, they still implicitly use 3D facial representations or rely on latent feature spatial domain separation for face swapping, which results in a significant loss of feature information and limits the quality of the generated image.

In summary, we proposed an efficient composite model for face-swapping tasks in this paper. The model solves the problem of feature information loss during face swapping, retains the original expression features, and its generated images are more realistic.

Our approach outperforms the best models available at this stage in terms of preserving the original face feature information, background, illumination, naturalness, and image quality. The key contributions of this paper are as follows:

1. We design a composite face-swapping generation network model to solve the feature loss and image blurring problems in the face-swapping process. The model includes two main modules: the facial feature extraction module, and the face feature fusion generation module.
2. To address the problem of unnatural facial feature fusion and poor image quality in face change tasks, we innovatively used a combination of variational autoencoders (VAEs) [17] and GANs to improve image quality post-face swap.
3. Our proposed model is experimentally validated to be more robust in face recognition, pose verification, and image quality assessment compared with other good models. In the validation of the assessment of image quality, our proposed model reduces the difference to 0.46 in the FID image quality score and obtains an excellent score of 0.91 in the SSIM score.

2. Related Works

The first task of face-swapping is the detection and recognition of faces [18]. The traditional methods, which typically recognize faces using one- or two-layer representations, employ filtered responses, feature code histograms, local descriptors, and feature transformations, which are less accurate for face recognition. Worse still, most of the traditional methods are unable to address unconstrained facial variations such as lighting, pose, expression, or camouflage. Therefore, these scurrying approaches often suffer from unstable performance or recognition errors in real-world applications. In the convolutional neural network-based approach, face recognition has made significant progress due to the widespread use of deep neural networks and the efficient processing power they have demonstrated. In DeepFace [19] and DeepID [20], face recognition is categorized as a multi-class classification problem and deep CNN models are introduced to learn features on large multi-identity datasets. In arcface [21], Additive Angular Margin Loss is proposed on the basis of SphereFace and CosFace for further increasing the intraclass compactness and interclass differentiation of extracted features to improve the discriminative ability of face recognition models. RetinaFace [22] utilizes joint extra-supervised and self-supervised multi-task learning to perform pixel-level face localization on faces of different scales to achieve accurate and efficient face localization in the wild. FaceNet [23] trains deep CNNs on nearly 200 million face images, and learns Euclidean spatial embeddings by using the ternary loss to achieve state-of-the-art performance.

The face-swapping task has long been a research interest in the computer graphics and computer vision communities. In the face swap task, a large number of distinctive modeling methods have emerged from long-term exploration and research, which are mainly divided into two research directions: 3D face feature-based methods and GAN-based methods. The 3D-based method, the earliest facial exchange method, requires manual participation in defining facial feature points. Later, a fully automatic facial feature coordinate alignment method was proposed Structural information, such as 3D models and landmarks, provides powerful prior knowledge. Justus et al. [24] proposed the face2face network, a method for real-time facial re-enactment of monocular target video sequences. They used a 3D deformable facial model (3DMM) [25] of two faces to achieve transfer of facial expressions from the source to the target face. However, 3D model-based approaches often fail to accurately reproduce expressions due to the limited expressive power of 3D face datasets. Generative adversarial networks (GANs) [26] have been shown to generate fake images with the same distribution as the target domain. Although it can successfully generate realistic appearances, training a GAN can be unstable, and this limits its application to low-resolution images.

However, subsequent methods have improved the stability of the training process [27]. Zhu et al. [28] proposed CycleGAN, which introduces cycle consistency loss to map images generated from the source domain back to the source domain, allowing the training of unsupervised universal transformations between different domains. Li et al. [29] Proposed the FaceShifter algorithm, a two-stage framework to achieve high-fidelity and occlusion-aware face-changing technology. It extracts target attributes at various spatial resolutions through a multi-level attribute encoder, and the generator adaptively integrates the face identity and attributes when synthesizing the face to generate a highly realistic replacement face. Chen et al. [30] introduced the ID Injection Module (IIM) to transfer the identity information of the source face to the target face at the feature level, solving the problem of identity restrictions, and at the same time using training loss constraints to prevent the target face attributes from being influenced by the source face. Wang et al. [31] designed a network, AP-Swap, consisting of a global residual attribute preserving encoder (GRAPE) and a landmark-guided feature entanglement module (LFEM). By performing landmark-based attribute preservation operations, the granularity of facial attributes is effectively preserved, which is used to improve the quality of the face-swapped image. Li et al. [32] proposed FaceSwapper, a network consisting of a decomposition representation module and a semantics-guided fusion module, which separates face identity and attribute information using an attribute encoder and an identity encoder. In addition, semantic information was introduced in the semantically guided fusion module to control the swapping region and to model pose and expression more accurately. These research methods have improved the quality of face-swapping to some extent, but there are still some problems such as feature anomalies after face-swapping, image blurring, and image quality degradation. For this reason, we propose a composite face-swapping generative network model in this paper to enhance the performance of face-swapping tasks. In the network, a feature extraction module with an attention layer is designed to improve the facial feature extraction capability, and a facial feature fusion generation module combining a Variational Autoencoder (VAE) network and a generative adversarial network (GAN) is innovatively constructed to solve the feature fusion anomalies, image quality degradation, and instability problems in the face-swapping task.

3. Methods

We designed a composite face-swapping generation network model, as shown in Figure 1. The network consists of two parts, the facial feature extraction module and the face feature fusion generation module. In the face-swapping task, we first need to detect and identify the face in the image and extract it. The face image is scanned using 3D feature points to determine the face region for cropping and alignment. Then, the corresponding face features are extracted by the facial feature extraction network. Afterward, the two-channel parallel generation network in the face feature fusion module realizes the high-quality face-swapping image generation.

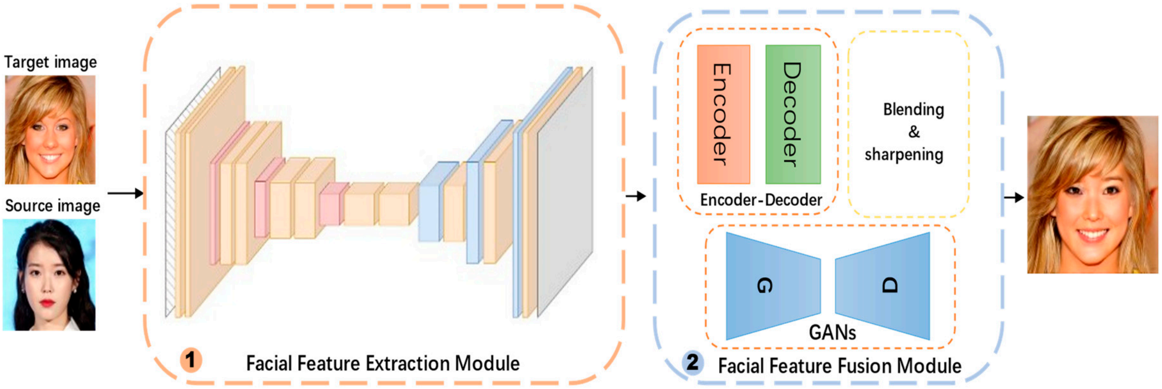

Figure 1. Our proposed composite face swap generation network framework. The framework contains two modules: (1) facial feature extraction module; (2) facial feature fusion module.

3.1. *Facial Feature Extraction Module*

The facial feature extraction module is shown in Figure 2. Firstly, the source and target images are preprocessed. Here, a 3D feature point detector is used for face alignment and pre-cropping of the face part of the image to obtain the preprocessed images I_t and I_s. The design of the feature extraction network module considers high efficiency and flexibility, so it borrows from the U-Net [2] framework. Since the original U-Net has the defect of losing contextual information, in our design of the face feature extraction module, we add the attention layer [5] after the convolutional layer, and output the accurate face feature maps, F_t and F_s, after Skip Connections, and computation of the deconvolutional layer.

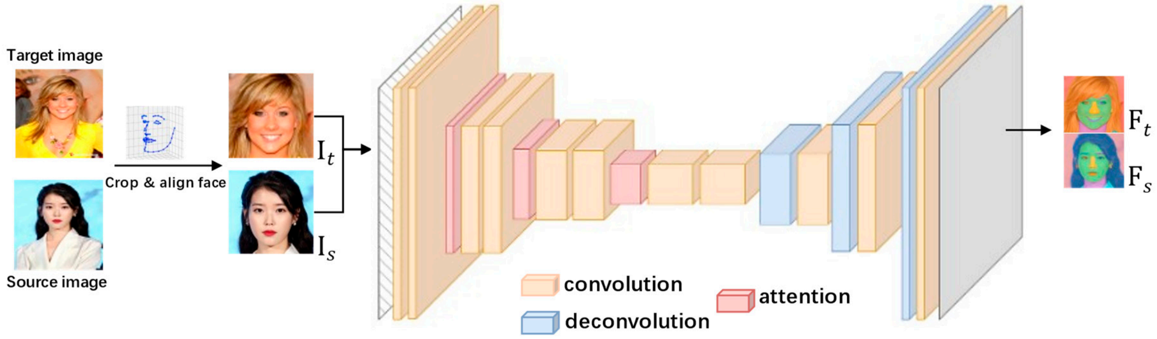

Figure 2. Facial feature extraction module. This module consists of a three-dimensional feature point detector responsible for the preliminary extraction of the face part in the image, and a face feature extraction network containing an attention layer extracts facial features.

In the face feature extraction network, the feature extraction loss contains reconstruction loss and feature matching loss. whose function is:

$$\begin{aligned}\mathcal{L}_{FE} &= \lambda_1\mathcal{L}_{\text{reconstruction}} + \lambda_2\mathcal{L}_{\text{feature matching}} \\ &= \frac{1}{N}\sum_{i=1}^{N} \parallel I_i - \hat{I}_i \parallel^2 + \frac{1}{N}\sum_{i=1}^{N} \parallel F_i - \hat{F}_i \parallel\end{aligned} \tag{1}$$

where I_i is the i-th input image, and $\hat{I}_i$ is the i-th image reconstructed after extracting feature through the U-Net and attention layer. F_i is the feature of the i-th input image, $\hat{F}_i$ is features reconstructed after extracting features through the U-Net and attention layers, and N is the number of images in the batch.

3.2. *Facial Feature Fusion Generation Module*

The module, shown in Figure 3, consists of a dual generative network based on VAE and GAN. It mainly fuses the cropped face features to achieve the perfect face swap. In the VAE-based face feature fusion generation network, it first performs feature blending on

F_s and F_t computed from face feature extraction. After that, the feature map is encoded in the encoder and the decoder extracts the data from the latent space layer for feature reconstruction, since the VAE-generated image has some blurring quality problems. For this reason, we add a sharpening layer to further sharpen the features after completing the generation of the face fusion image, to enhance the image quality. A loss function combining reconstruction loss and KL divergence loss is used to stabilize the network during feature training. Its loss equation is:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{KL}} \tag{2}$$

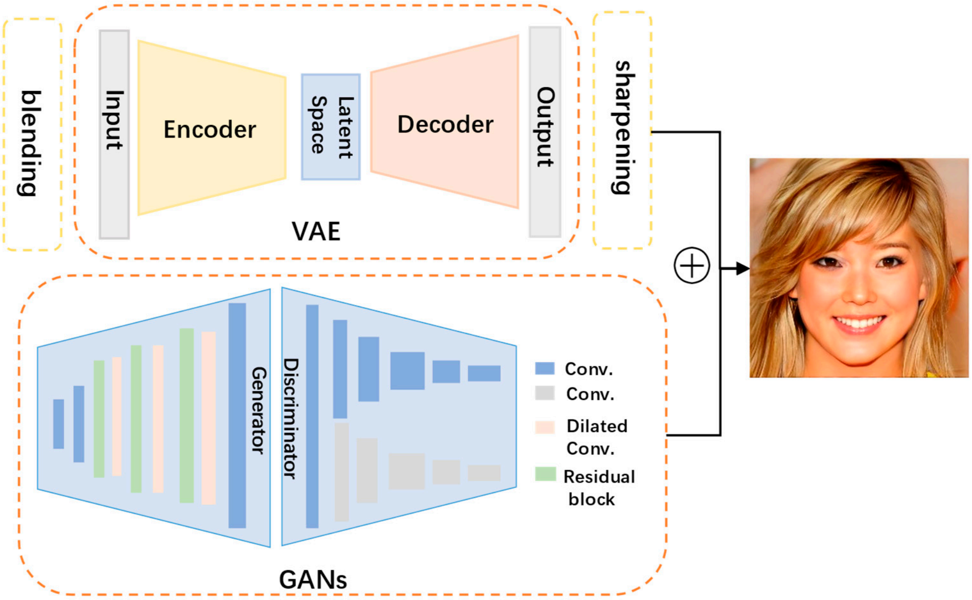

Figure 3. Facial feature fusion module. The upper part of the module uses blending and sharpening algorithms and VAE (Variational Autoencoder) network to stably generate face images. The GAN framework in the lower part includes a generator composed of ResNets and a discriminator composed of local and global convolutional networks to improve the quality of generated face images.

In the GAN-based facial feature fusion generation network, we reconstructed the generators and discriminators in the GAN framework. In the generator part, the ResNet is used as the main network for the reconstruction and generation of the fusion of face features F_t and F_s, and the size of the feature convolutional sensory field of view is crucial for the generation of face feature texture in the fusion generation training. Here, we add a dilated convolution [33] layer instead of the upper and lower sampling layers to increase the feature sensory field of view, and at the same time, reduce the excessive and useless convolutional computation. In the construction of the discriminator, consider that the discriminator is only to discriminate the quality of the image generated by the generator, so there is no need to design anything too complex as compared to the generator. Here, we use a discriminator composed of global and local convolutional networks to discriminate the feature authenticity of the image generated after face swapping. In the discriminator, the global convolutional network consists of multiple convolutional layers and one fully connected layer. All convolutional layers use 2 × 2-pixel spans to reduce the image resolution. The local convolutional network uses a similar architecture, except that the input image block size is half that of the global discriminator. The authenticity of the swapped face is finally confirmed by integrating the global and local convolutional network outputs into the sigmoid activation function computation. Its network loss can be expressed as:

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

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

Finally, the composite face-swapping generation network loss is:

$$\mathcal{L}_{\mathrm{FG}} = \mathcal{L}_{\mathrm{vae}} + \mathcal{L}_{\mathrm{gan}} \tag{4}$$

4. Experiments

4.1. Trainning and Implementation Details

Datasets. In model training and testing, we selected a variety of datasets for experiments. CelebA-HQ [34]: The CelebFaces Attributes—High Quality dataset comprises a total of 30,000 high-quality images generated from the celeba dataset after denoising and super-resolution by a generative adversarial network. FFHQ [35]: The Flickr-Faces—High Quality dataset is a high-quality image set containing 70,000 PNG images with a resolution of 1024 × 1024, which are crawled from Flickr and automatically aligned and cropped using dlib, and it contains considerable diversity in terms of age, ethnicity, and image background. FaceForensics++ [36]: These data are derived from 977 YouTube videos, all of which contain trackable frontal surfaces and are unobstructed, and were processed using four automated face processing methods, Deepfakes, Face2Face, FaceSwap, and NeuralTextures, to create a face forgery dataset consisting of 1000 raw video sequences comprising a face forgery dataset.

Implementation Details: In the composite face-swapping generative network, several network modules are involved, each of which needs to be trained individually. For the training of the face feature extraction module, the learning rate is set to 0.0001, and 1000 rounds of training are performed to achieve excellent feature extraction performance. For the face feature fusion module, both VAE and GAN are trained for 100,000 rounds, where the learning rate of VAE is set to 0.0001, and in GAN, the learning rate of the generator is 0.00001 and the learning rate of the discriminator is 0.0001. For data selection, the face image datasets of CelebA-HQ and FFHQ with a uniform image size of 512 × 512 are used for the training of the facial feature extraction module and the facial change fusion generation module. In addition, for the video dataset, we refer to and borrow the relevant experimental setup of FaceShifter, and align and crop the face; the cropped image is 256 × 256 in size and covers the entire face and some background areas. Each video in the video dataset FaceForensics++ is uniformly sampled and processed at 10 frames per second to obtain 10,000 aligned faces. Manual checking of aligned faces was also performed to prevent detection errors. After data cleaning, all of the corresponding frame images in the video are processed and extracted for testing. The experimental configuration environment is shown in Table 1.

Table 1. Experimental environment configuration.

	Configuration Information
Operating Systems	Windows 10
Development Languages	Python 3.11
Frameworks	Pytorch 2.1.0 + CUDA 11.8
CPU	AMD 5800X
GPU	NVIDIA RTX 3090 24 G (x2)
Memory	48 GB

4.2. Competing Methods

At this stage, most face swap networks encounter issues after facial feature changes. Most of the generated face images have some problems, such as image blurring, abnormal face features, insufficient feature fusion with a sense of tearing, and inability to maintain the overall appearance of the original face, as shown in Figure 4. Compared with other excellent models, in the composite model we proposed, the image generated after changing the face preserves the overall appearance of the original face while integrating new facial features, with higher clarity and naturalness.

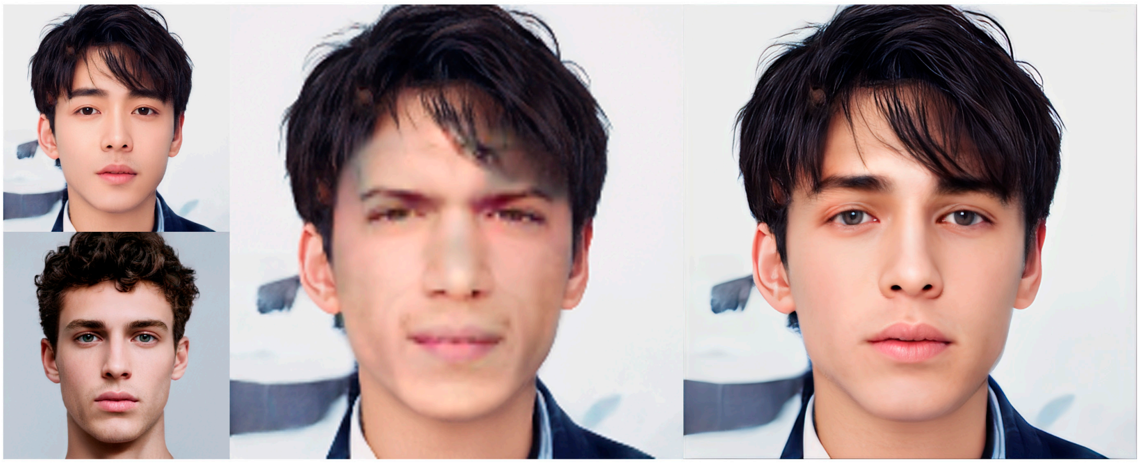

Figure 4. Comparison of image features generated after face swap.

After the face swap, how can we determine whether the face swap was successful? We cannot accurately verify it with the naked eye, so we introduced identity retrieval and posture. We verify whether the face identity after the face swap is successfully integrated and whether the face posture has changed. In the quantitative comparison of model metrics, to validate the effectiveness of our proposed composite network model, we quantitatively compared the ID retrieval and pose of the face with those of excellent models from other studies. We adopted the method proposed by Faceshifter to extract the identity vector and used cosine similarity to measure the identity distance. Faces were exchanged using the same source and target in FaceForensics++. For each exchanged face in the test set, the nearest face in the original video frame was identified, and its correspondence to the correct source video was checked. The accuracy of this identity determination is referred to as ID retrieval. Additionally, we used a pose estimator to estimate the poses in the generated and original frames and computed their average L2 distances.

In the ID detection and pose evaluation comparison experiments, we studied and compared the emergence of superior models at this stage. Each was tested 10 times, and the average scores were selected for model performance comparison. As presented in Table 2, our proposed composite network model performs excellently compared with other network models in both ID detection and pose evaluation comparison experiments, in which the ID retrieval accuracy is improved to 97.82, the L2 distance is reduced to 1.55 in pose assessment, and the model performance is better than other network models. As shown in Figure 5, when handling the cross-gender face-switching task with different skin colors and facial expression features, our designed network model maintains the original overall expression features while switching faces, making the face-switching task natural and unobtrusive.

Table 2. Performance comparison of different models.

Methods	ID Retrieval	Pose
Faceswap [14]	54.19	5.73
FSGAN [15]	60.34	5.28
Deepface [13]	81.96	4.29
Simswap [30]	92.65	2.74
Ours	97.82	1.55

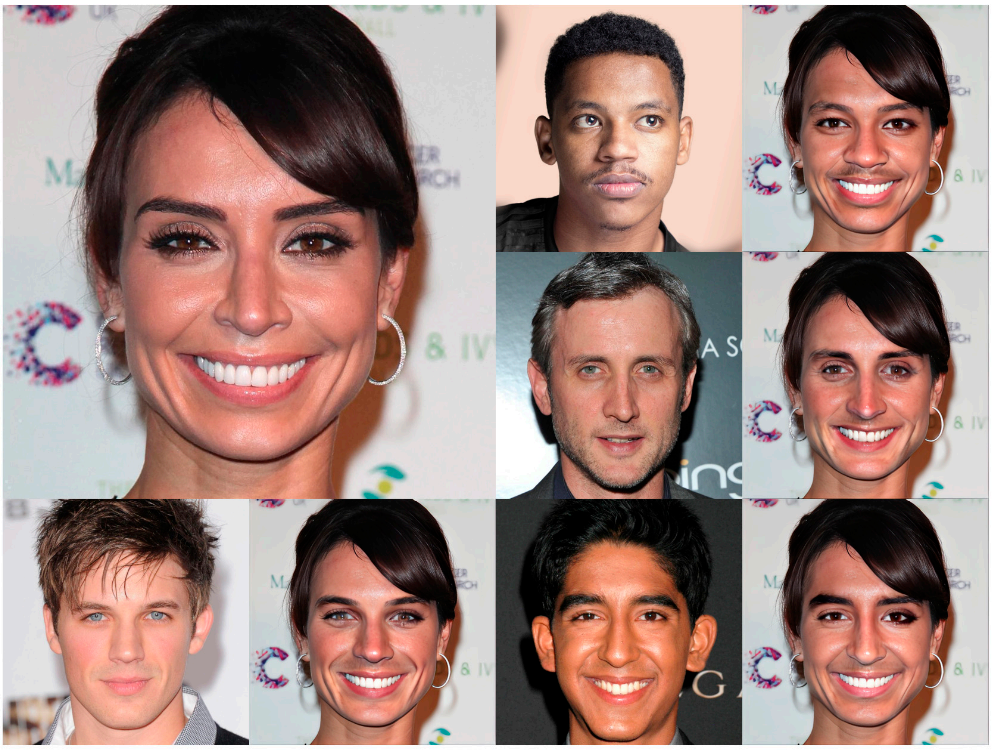

Figure 5. Face swapping of gender-specific facial features.

4.3. Performance Evaluation Metrics

In the face-swapping task, it is important to evaluate the overall quality of the image generated after face swapping. Most of the face-swapping models have serious degradation in the quality of the generated images after face swapping. In the image quality verification experiment, we used the Fréchet inception distance (FID) and structural similarity index measure (SSIM) to evaluate the quality of the images generated after face replacement.

The FID is a metric used to measure the difference between images generated by a generative model and real images. It is calculated by comparing their distribution distance in the feature space of the Inception v3 [37] model. The feature vector used by the FID is the high-dimensional vector output of the penultimate fully connected layer of the Inception v3 model. This vector captures the visual feature information of an image. The formula used is as follows:

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

where g and r represent the generated image and the real image, respectively, and μ_g and μ_x denote the mean values of the respective eigenvectors. Σ_g and Σ_x denote the covariance matrix of the respective eigenvectors, and Tr denotes the trace of the matrix.

The SSIM method evaluates image quality by examining multiple regions of the image and comparing statistics such as structure, brightness, and contrast within these regions. This algorithm considers the characteristics of human image perception and is more in line with the visual characteristics of the human eye. The equation is as follows:

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

where $l(x,y)$ is the luminance comparison, $c(x,y)$ is the contrast comparison, and $s(x,y)$ is the structure comparison.

In the composite face transformation generation network, we propose that quality of the generated face image after swapping is compared with other excellent models, in which the proposed model and other compared models are tested on 1000 images generated by them, respectively, and the average value is selected for the model performance comparison. As shown in Table 3, FID and SSIM image quality reviews perform ahead of other networks. In the FID test, the difference between images was reduced to 0.46, and in the SSIM image quality assessment, its performance was improved to 0.91. In the face-swapping task, the generated images generally suffer from blurred images, distorted facial features, changes in overall appearance, and low naturalness. In Figure 6, our model effectively suppresses these issues compared to other models. The facial images generated by our model after face swap are more realistic and naturally retain their original overall appearance.

Table 3. Comparison of image quality performance between different models.

Methods	FID	SSIM
Faceswap	0.57	0.75
FSGAN	0.63	0.54
Deepface	0.59	0.80
Simswap	0.53	0.85
Ours	0.46	0.91

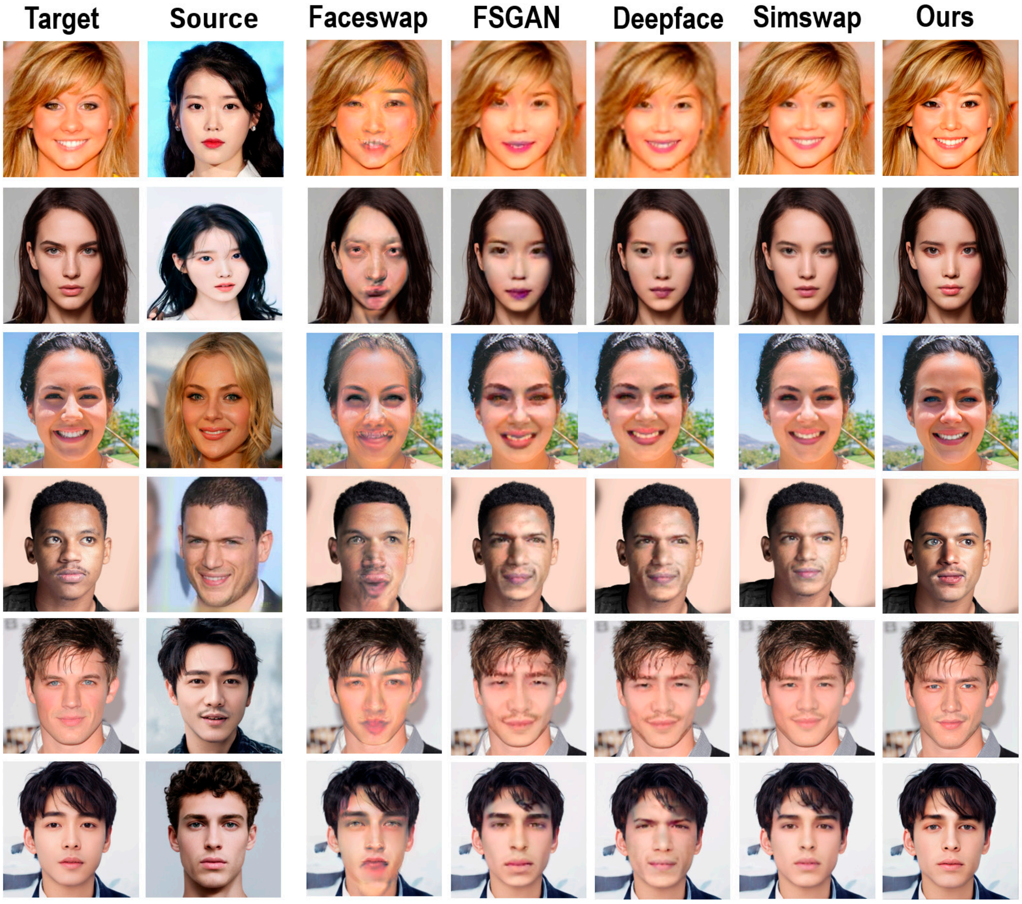

Figure 6. Comparison with other excellent models for face swapping.

In the statistical charts, as shown in Figure 7, it is more clearly demonstrated that our proposed model is more robust than other good models regarding multiple image generation quality assessment and face feature assessment.

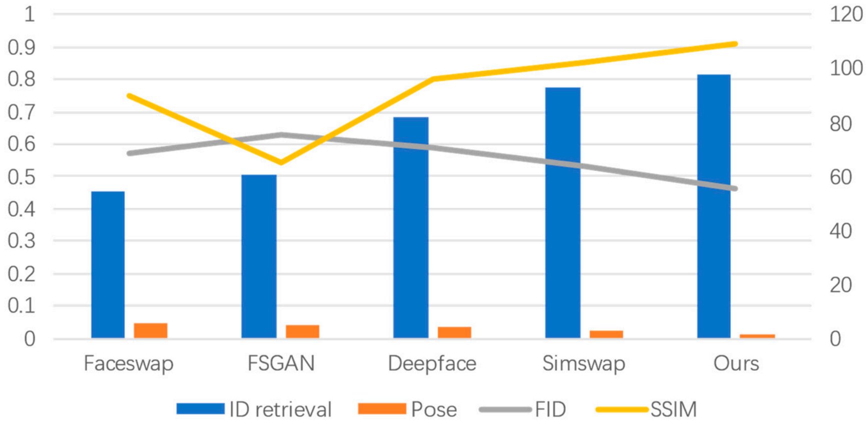

Figure 7. Model performance comparison chart.

When faced with different skin colors, expressions, complex backgrounds, and cross-gender face-swapping tasks, as shown in Figure 8, our proposed model generates face images that retain the original background and lighting characteristics, resulting in a more natural appearance. Even in non-frontal (side face) situations, the model still exhibits excellent performance.

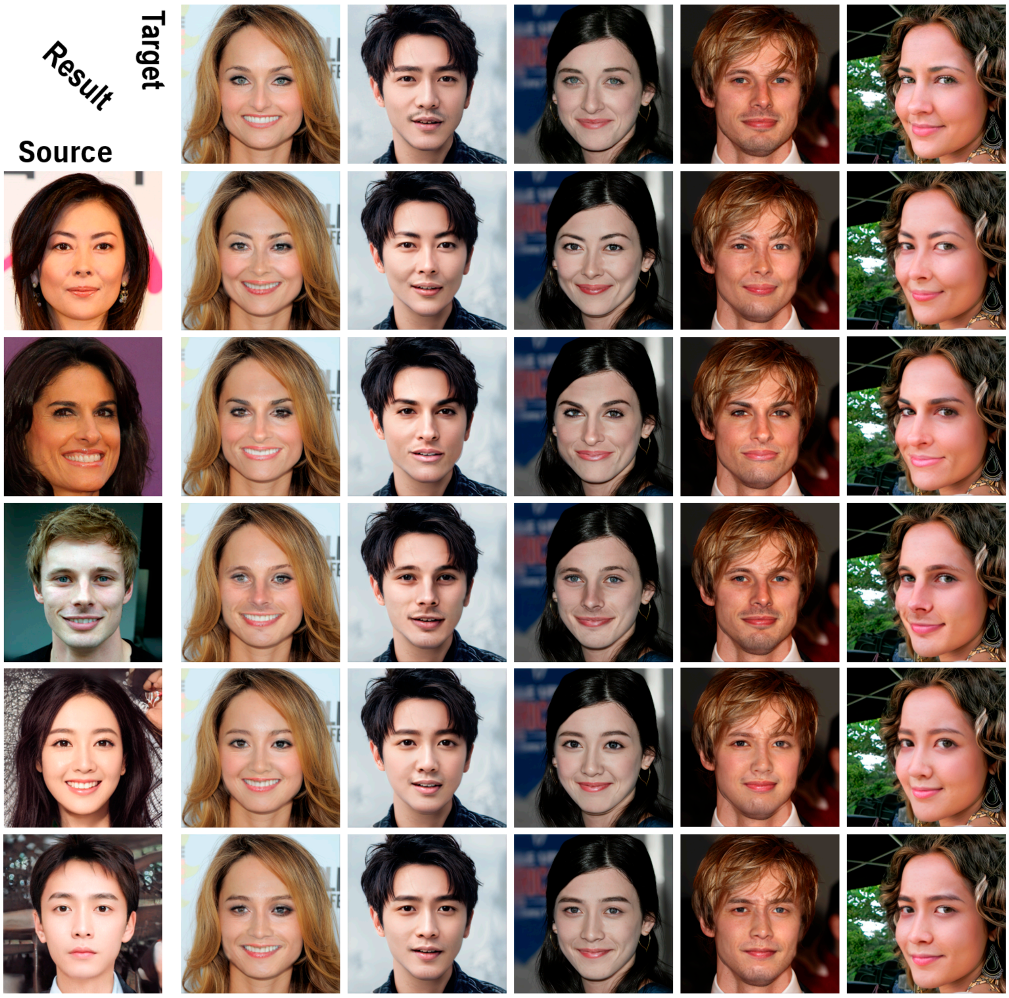

Figure 8. Different types of face-swapped images.

5. Conclusions and Discussion

Although the face-swapping task has been studied for a long time, at this stage, most of the models use a single network for the face-swapping task. There are still some problems that are hard to ignore, with the appearance of unstable face-swapping features, fusion blurring, and the appearance of accidents, such as a strong sense of boundary. For this reason, in this paper, we design a composite face-swapping generative network modeling method to solve these problems. We introduce a facial feature extraction network module with an attention layer to enhance the face feature extraction capability. In face swapping, the facial feature fusion generation module consisting of VAE and GAN solves the unstable image generation of a single GAN network and designs a GAN using it as a generator and local and global convolutional networks as discriminators for enhancing image generation performance. After experimental verification, our network is feasible and efficient. In the validation of the assessment of image quality, our proposed model reduces the difference to 0.46 in the FID image quality score and obtains an excellent score of 0.91 in the SSIM score.

At this stage, face-swapping technology is still too demanding for computer computation to be deployed in small, low-energy devices, and how to address the light weight of the network is a direction we should consider and research in the future. In addition, we should investigate how to identify the authenticity of face-swapped pictures and solve the security problems caused by the current emergence of many face-swapping technologies and related software, which cannot distinguish between the authenticity and falsity of face-swapped pictures, which is another future research direction of ours.

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

Funding: This work was supported by project for Collabo R&D between Industry, Academy, and Research Institute funded by Korea Ministry of SMEs and Startups in 2023 (RS-2023-00283448) and Korea Institute of Marine Science and Technology Promotion (KIMST) Funded by the Ministry of Oceans and Fisheries (RS-2022-KS221571).

Institutional Review Board Statement: All subjects gave their informed consent for inclusion before they participated in the study. Ethics approval is not required for this type of study. The study has been granted exemption by the Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: All datasets utilized in this article are open-source and publicly available for researchers. Interested individuals can obtain the datasets using the following link: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 5 December 2023), https://github.com/NVlabs/ffhq-dataset (accessed on 12 December 2023), https://justusthies.github.io/posts/faceforensics++/ (accessed on 21 February 2024).

Conflicts of Interest: Author Kyoungae Jang was employed by the company Datatree Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. 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. [CrossRef]
2. Man, Q.; Cho, Y.I.; Jang, S.G.; Lee, H.J. Transformer-based gan for new hairstyle generative networks. *Electronics* **2022**, *11*, 2106. [CrossRef]
3. Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P.; Nayar, S.K. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.* **2008**, *27*, 1–8. [CrossRef]

4. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
5. V Blanz, V.; Scherbaum, K.; Vetter, T.; Seidel, H.P. Exchanging faces in images. In *Computer Graphics Forum*; Blackwell Publishing, Inc.: Oxford, UK; Boston, MA, USA, 2004; Volume 23, pp. 669–676.
6. Agarwala, A.; Dontcheva, M.; Agrawala, M.; Drucker, S.; Colburn, A.; Curless, B.; Salesin, D.; Cohen, M. Interactive digital photomontage. *ACM Trans. Graph.* **2004**, *23*, 294–302. [CrossRef]
7. Phan, H.; Nguyen, A. DeepFace-EMD: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20259–20269.
8. Chang, F.J.; Tuan Tran, A.; Hassner, T.; Masi, I.; Nevatia, R.; Medioni, G. Faceposenet: Making a case for landmark-free face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 1599–1608.
9. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast face-swap using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3677–3685.
10. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 98–105.
11. Pérez, P.; Gangnet, M.; Blake, A. Poisson image editing. In *ACM SIGGRAPH 2003 Papers (SIGGRAPH '03)*; Association for Computing Machinery: New York, NY, USA, 2003; pp. 313–318. [CrossRef]
12. Li, A.; Hu, J.; Fu, C.; Zhang, X.; Zhou, J. Attribute-conditioned face swapping network for low-resolution images. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2305–2309.
13. Liu, K.; Perov, I.; Gao, D.; Chervoniy, N.; Zhou, W.; Zhang, W. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognit.* **2023**, *141*, 109628. [CrossRef]
14. FaceSwap, "FaceSwap". Available online: https://github.com/MarekKowalski/FaceSwap/ (accessed on 15 November 2019).
15. Nirkin, Y.; Keller, Y.; Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 7184–7193.
16. Xu, Z.; Yu, X.; Hong, Z.; Zhu, Z.; Han, J.; Liu, J.; Ding, E.; Bai, X. Facecontroller: Controllable attribute editing for face in the wild. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 3083–3091.
17. Razavi, A.; Van den Oord, A.; Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14866–14876.
18. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015; British Machine Vision Association: Glasgow, UK, 2015.
19. Taigman, Y.; Yang, M.; Ranzato, M.A.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
20. Sun, Y.; Wang, X.; Tang, X. Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1891–1898.
21. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
22. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv* **2019**, arXiv:1905.00641.
23. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
24. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
25. Blanz, V.; Romdhani, S.; Vetter, T. Face identification across different poses and illuminations with a 3d morphable model. In Proceedings of the Fifth IEEE International Conference on Automatic Face Gesture Recognition, Washinton DC, USA, 20–21 May 2002; IEEE: Piscataway, NJ, USA, 2002; pp. 202–207.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
27. 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.
28. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
29. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv* **2019**, arXiv:1912.13457.

30. Chen, R.; Chen, X.; Ni, B.; Ge, Y. Simswap: An efficient framework for high fidelity face swapping. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2003–2011.
31. Wang, T.; Li, Z.; Liu, R.; Wang, Y.; Nie, L. An efficient attribute-preserving framework for face swapping. *IEEE Trans. Multimed.* **2024**, *26*, 6554–6565. [CrossRef]
32. Li, Q.; Wang, W.; Xu, C.; Sun, Z.; Yang, M.-H. Learning disentangled representation for one-shot progressive face swapping. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–17. [CrossRef] [PubMed]
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. 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.
36. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 1–11.
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

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