electronics

Article

Color Face Image Generation with Improved Generative Adversarial Networks

Yeong-Hwa Chang [1,2,*], Pei-Hua Chung [1], Yu-Hsiang Chai [1] and Hung-Wei Lin [1]

[1] Department of Electrical Engineering, Chang Gung University, Taoyuan City 333, Taiwan
[2] Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243, Taiwan
* Correspondence: yhchang@mail.cgu.edu.tw

Abstract: This paper focuses on the development of an improved Generative Adversarial Network (GAN) specifically designed for generating color portraits from sketches. The construction of the system involves using a GPU (Graphics Processing Unit) computing host as the primary unit for model training. The tasks that require high-performance calculations are handed over to the GPU host, while the user host only needs to perform simple image processing and use the model trained by the GPU host to generate images. This arrangement reduces the computer specification requirements for the user. This paper will conduct a comparative analysis of various types of generative networks which will serve as a reference point for the development of the proposed Generative Adversarial Network. The application part of the paper focuses on the practical implementation and utilization of the developed Generative Adversarial Network for the generation of multi-skin tone portraits. By constructing a face dataset specifically designed to incorporate information about ethnicity and skin color, this approach can overcome a limitation associated with traditional generation networks, which typically generate only a single skin color.

Keywords: generative adversarial networks; image generation; image recognition

Citation: Chang, Y.-H.; Chung, P.-H.; Chai, Y.-H.; Lin, H.-W. Color Face Image Generation with Improved Generative Adversarial Networks. *Electronics* **2024**, *13*, 1205. https://doi.org/10.3390/electronics13071205

Academic Editors: Zongwei Wu, Fan Yang, Virginie Fresse, Chao Li and Slaviša Jovanović

Received: 26 January 2024
Revised: 20 March 2024
Accepted: 21 March 2024
Published: 25 March 2024

1. Introduction

The field of artificial intelligence (AI) has experienced rapid progress in recent years, driven by advancements in computing capabilities and the availability of large datasets. The progress in artificial intelligence has led to the development of versatile applications across various domains, such as image recognition [1–3], object detection [4–6], medical diagnosis [7–9], and Internet of Things (IoT) [10–12]. In many applications, the development of image recognition is relatively well-advanced. The progress in this field is attributed to the development of sophisticated algorithms, the availability of large datasets, and increased computing power. The development of the Convolutional Neural Network (CNN) can be viewed as a crucial point in the advancement of image recognition and computer vision. The convolutional layers in CNNs employ filters that scan the input image to detect local patterns. Motivated by the advent of CNN networks, there are some variants, including R-CNN, Fast R-CNN, Faster R-CNN [13–16], and Mask R-CNN [17–19]. In addition, recognition performance can be further improved using the Residual Network (ResNet) with deep learning.

While there have been significant advancements in image recognition, generative networks, which are used for tasks like image generation, still face several challenges. In the earlier stages of developing generative models, manual assessment of generated results followed by algorithm modifications was a common practice. Automatically judging the quality of results and optimizing generative models has been a crucial challenge in deep learning, especially for tasks involving generative models. In traditional supervised learning, the training process involves using labeled data to train a model to map input features to corresponding target labels. However, in generative networks, it is hard to define a correct answer for a generated result. This is the most difficult issue encountered by generative networks in applying deep learning. The emergence of Generative Adversarial Networks (GANs) has solved this problem of generative networks. GAN is centered around two models set against each other, namely, a generator and discriminator. These two models are trained simultaneously through adversarial training. In general, GANs provide a framework for training generative models that can be applied to various domains, including image generation and text-to-image synthesis. Image-to-image models focus on transforming one type of image into another, while text-to-image applications specifically involve the generation of images from textual descriptions. These concepts showcase the versatility of generative models in learning complex mappings and generating diverse types of data [20–23].

The integration of Generative Adversarial Networks (GANs) with transformer-based or diffusion-based models has indeed been a focus of recent research efforts. In a transformer-based GAN, both the generator and discriminator architectures are built upon transformer models, which are widely employed in natural language processing tasks for their ability to capture long-range dependencies and contextual information effectively. On the other hand, a diffusion-based GAN leverages the principles of diffusion models, which simulate the iterative process of diffusing noise into data to generate realistic samples. Several interesting works have explored these integrated learning approaches [24–29]. For instance, a recent study introduced a GAN network that employed transformer models to enhance text semantics, improving the relevance of generated images [24]. Another study presented a discriminator structure integrated with a specific transformer model to enhance the quality of image restoration [25]. Additionally, a transformer-based GAN was proposed for sonar images to remove speckle noises, thereby enhancing the accuracy of analysis tasks on sonar images [26]. Moreover, a method for imputing missing traffic state data was proposed based on a Diffusion Convolutional Neural Network–Generative Adversarial Network [27]. Furthermore, a method based on adversarial diffusion modeling was developed to enhance performance in medical image translation tasks [28]. Lastly, a denoising diffusion GAN was introduced to achieve large-step denoising, sample diversity, and training stability [29]. While transformer-based GANs and diffusion-based GANs offer advantages for image generation tasks, they also have certain limitations. Transformer models often require substantial computational resources, particularly for large-scale image generation tasks. On the other hand, diffusion models may have slow sampling speeds due to the multiple steps of noise diffusion required to generate each sample, compared to other GAN architectures.

Recently, the research related to GANs has attracted much attention; typical applications include image de-raining [30–32], image inpainting [25,33,34], image reconstruction [35–37], and image synthesis [38–40]. Image de-raining is a computer vision task that involves removing or reducing the effects of rain from images. Outdoor surveillance and autonomous driving are typical applications, situations where clear vision is crucial. In [25], a conditional Generative Adversarial Network was proposed for the single de-raining problem, considering both quantitative and visual performance. Image inpainting can be performed for various reasons, such as removing unwanted objects, restoring old photographs, or completing regions that are occluded or corrupted. In [35], an attention-generating adversarial image painting method was presented, aiming to capture global semantic data. Image reconstruction aims to produce a high-quality image from limited or noisy information; the applications for medical imaging and remote sensing are typical ones. In [38], a GAN framework was addressed to the recovery of high-quality PET images from filtered back projection PET images with streaking artifacts and high noise. Image synthesis refers to the generation of new images from scratch or the combination of existing images to create novel visual content. In [40], an auto-embedding Generative Adversarial Network was presented, aiming simultaneously to encode the global structure features and capture the fine-grained details.

The COCO dataset is among the most widely used and comprehensive datasets for object recognition and other computer vision tasks [13,16,41]. Researchers often use the

COCO dataset as a testbed for developing and testing models related to object recognition, instance segmentation, and image captioning. The progress of GANs has been significant, leading to impressive results in various domains, including image generation, style transfer, and data augmentation. However, there are challenges and considerations associated with the usage of GANs, including dataset limitations. This paper constructs a self-made image dataset for GAN learning that includes diverse skin colors. It can be verified that an image with colored skin can be generated from its sketch portrait.

This paper aims to contribute to the field of image generation by introducing an improved GAN designed for creating color images from sketches. The utilization of GPU computing is highlighted for efficient model training. Additionally, the focus on multi-skin tone portraits and the consideration of ethnicity and skin color in the dataset construction distinguish this research from traditional approaches. The comparative analysis adds a valuable reference point for understanding the strengths and weaknesses of various generative networks. Generating realistic and diverse images from sketches is a challenging task, and achieving success in this area has practical applications in various fields, including art, design, and entertainment. This paper contributes to the field of image generation by introducing an improved approach to creating color portraits from sketches using a Generative Adversarial Network (GAN) architecture. The key contributions of the paper can be summarized as follows:

1. This paper applies the Generative Adversarial Network architecture to improve the training process of the generative model. This enables a more effective transformation of a sketch image into a color portrait.
2. The use of the U-Net architecture for constructing the generative model allows for the retention of fine facial features, ensuring that the generated color portraits accurately capture the details from the input sketches.
3. Incorporating a Convolutional Neural Network as the discrimination model facilitates interactive training with the generative model. This adversarial training process enhances the quality and realism of the generated color portraits.
4. The creation of datasets classified based on ethnicity and skin color enables the training of generators capable of producing color portraits with diverse skin tones.

2. Materials and Methods

2.1. Image Generation

The rapid development of image recognition, particularly in the context of deep learning, can be attributed to several factors, including advancements in hardware, the emergence of specific architectures like Convolutional Neural Networks (CNNs), and the ability to easily identify and quantify the correctness of the recognition outcomes. During the training process, we provide the model with labeled examples, indicating the correct answers or ground truth for each input. It is expected that the model will produce the same answer. However, the situation with image generation is different compared to the cases of image recognition or classification. In image generation tasks, the goal is to create entirely new images rather than simply recognizing or classifying existing ones. There is no predefined ground truth for the generated images, and the produced outputs could be diverse. These situations make it difficult to develop image generation in the field of deep learning.

Due to the difficulty in defining an effective loss function, the development pace of generative networks is relatively slower compared with other deep learning applications. It was not until 2014, when Ian Goodfellow released the paper "Generative Adversarial Network" (GAN) [38], that a researcher came up with the idea which could solve the biggest problem of generative networks. In the GAN framework, a generator and a discriminator are pitted against each other. The generator creates synthetic data, and the discriminator tries to distinguish between real and generated data. The objective is for the generator to produce increasingly realistic data, and for the discriminator to become better at distinguishing between real and fake samples. The process involves adversarial training, in which both networks iteratively improve their performance. The judgment in image recognition is influenced by the discriminative model, which is tasked with evaluating the quality of the outputs generated by the generative model. The discriminator plays a crucial role in distinguishing between real data from the training dataset and synthetic data generated by the generative network. Pre-trained discriminators can reduce the number of epochs required for training the entire GAN system so that the desired training results will be potentially faster and better [23].

The idea of Generative Adversarial Networks, first published in 2014, has been recognized as a turning point for generative models [42]. While GANs introduced a groundbreaking concept and demonstrated the ability to generate realistic data, their practical adoption initially faced issues such as long training times and training instability, leading to unsatisfactory results in some cases. This situation prevailed until the Deep Convolutional Generative Adversarial Network (DCGAN) was proposed; it was only then that the core idea of Generative Adversarial Networks was widely applied. The Deep Convolutional Generative Adversarial Network applies transposed convolution to improve the effect of generating the model, resulting in reduced training time. Transposed convolutional layers are placed in the generator to gradually upsample the low-dimensional feature maps. This process starts with a small spatial dimension and progressively increases it, creating a hierarchy of features. There have been various versions of Generative Adversarial Networks built upon the foundational concepts introduced by DCGAN.

Pix2Pix, introduced in 2016, is indeed a notable variant of GAN, specifically designed for image-to-image translation tasks [43–46]. It is particularly known for its use of an encoder–decoder architecture, which has been influential in various image conversion applications. Under the traditional encoder–decoder architecture, the encoder is used to capture image features, and some information will be lost in the process of transmitting to the decoder, resulting in the incoming image features not being fully represented. Therefore, U-Net was used in later variants; the main difference was that U-Net had a concatenate layer. The concatenate layer connects identical layers of the encoder and the decoder and transmits the output of each layer of the encoder to the corresponding decoder layer, which is merged before being sent to the next layer, in a manner such that the image position information is preserved.

2.2. System Architecture

The whole system described in this study is mainly divided into two parts: the user host and the GPU computing host, shown as Figure 1. The GPU computing host has a high computing efficiency, which aims to save time in model training. The system classifies, retouches, and pairs a large number of images on the GPU host and builds generation and discrimination models for adversarial training. The user host performs simple image preprocessing and generates images using the trained model from the GPU computing host. In this study, the GPU host is dealing with deep learning training. The first task is to set up an environment where the deep learning neural networks can be executed, for which the Ubuntu 20.04 is considered the operating system. After the generative model is trained on the GPU computing host, the user host only needs to perform simple image preprocessing, and the generative model trained by the GPU host can be executed without consuming too many computing resources.

The GPU host is responsible for both training a generative model and performing image retouching during the model-building process. The user host performs image pre-processing and uses the pre-trained generative model provided by the GPU host for image generation. In terms of environment construction, the user host and the GPU computing host use the same operating system, environment management platform, and GPU acceleration tools.

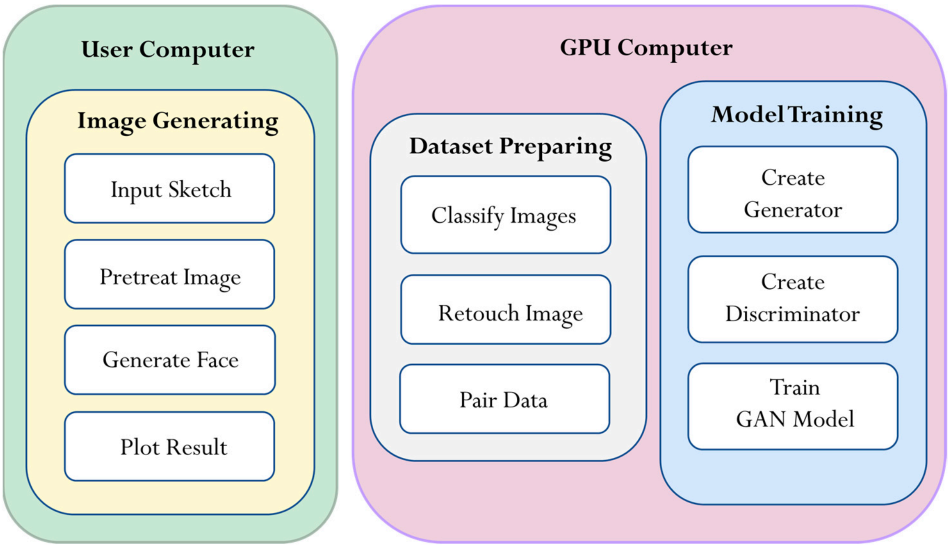

Figure 1. The scheme of the proposed system.

Deep learning frameworks help to accelerate the development and execution of algorithms, particularly in the context of utilizing NVIDIA GPUs (Graphics Processing Units). Frameworks often integrate with GPU-specific libraries like CUDA and CuDNN, which are optimized for deep learning tasks. Google's TensorFlow, Facebook's Pytorch, Berkeley's Caffe, and Microsoft's Cognitive Toolkit have indeed been popular open-source deep learning frameworks. Among the many deep learning frameworks, TensorFlow is an especially widely used open-source deep learning framework. It has been adopted by researchers for a variety of applications. TensorFlow, developed by Google, is the most widely used open-source mathematical computing framework. TensorFlow is designed to be highly flexible and can be built and run on various computing architectures, including systems with multiple CPUs and GPUs. In this study, TensorFlow 2.0 is adopted as the learning framework, and Anaconda is utilized for the environment management.

2.3. Experimental Methods

The experimental process, shown in Figure 2, is mainly divided into three steps: dataset preparation, model training, and image generation. In the dataset preparation part, the datasets used for model training are provided, including raw data, ethnicity-enhanced data, and skin color-enhanced data. In the model training step, the prepared datasets are used to train the GAN model, and then the trained generation model is provided for the subsequent image generation.

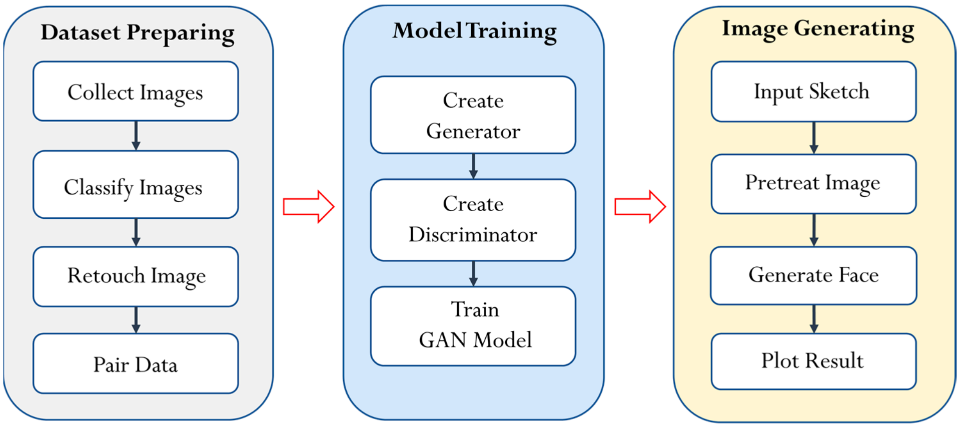

Figure 2. Experimental process.

2.3.1. Dataset Preparation

In this work, the goal is to transform images from sketches into realistic portrait images. The datasets used to train the generative model include the real portrait photos and the corresponding sketch images. The CelebA-HQ dataset, containing about 30,000 images of 1024 × 1024, is a collection of portraits of American celebrities [47–50]. In this study, the CelebA-HQ dataset was used as the image source of the self-built dataset to create sketch images. As trained by the traditional GAN scheme, the outcomes are shown in Figure 3. At this stage, only Caucasian images can be generated, no matter what the original skin color is. If the training set is composed entirely of images with a single skin color, it is expected that the generative model, when trained on such data, may have limitations in generating realistic images with diverse skin tones. The model learns from the patterns present in the training data, and if the data is not diverse enough, it might not generalize well to other skin tones. In this case, we added multiple-skin-tone images to the training dataset, but the output did not change as desired under the traditional GAN framework.

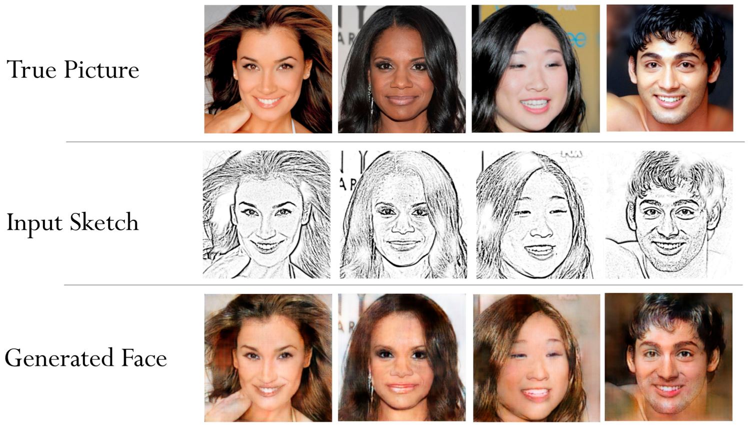

Figure 3. Preliminary image generation.

Based on the CelebA-HQ Dataset, a new dataset was constructed based on classifications of ethnicity, namely Caucasian, African American, Asian, and Middle Eastern. There are 200 images of each category, and the samples from the dataset are shown in Figure 4. The architecture of the training model was re-planned, three independent generative models were added, and four generative models were used to generate images associated with different ethnicities for each of the four people.

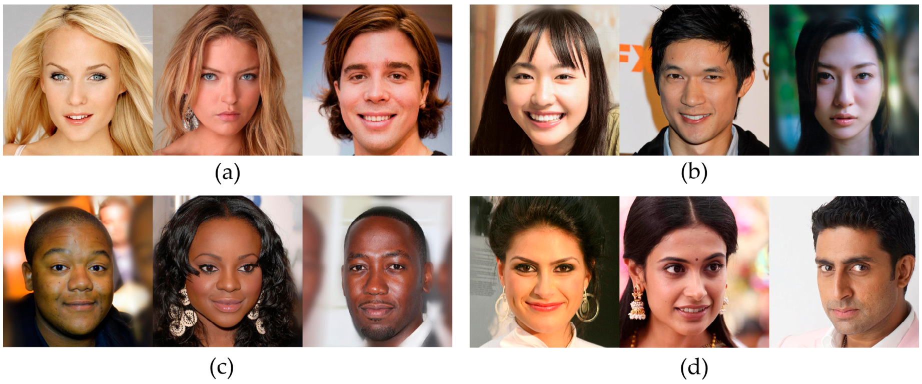

Figure 4. Samples based on ethnicity: (**a**) Caucasian, (**b**) Asian, (**c**) African American, and (**d**) Middle Eastern.

To further describe the possible diversity of skin tones, the following skin colors are considered: pale white, light brown, dark brown, and deep dark brown. This categorization helps acknowledge the diversity of skin tones, emphasizing the variations that exist within different populations. In this work, the U-Net network will be used to preserve the face structure, and the skin color tone can be preserved in the up-convolution's kernel. The samples of the skin tone classification are shown in Figure 5.

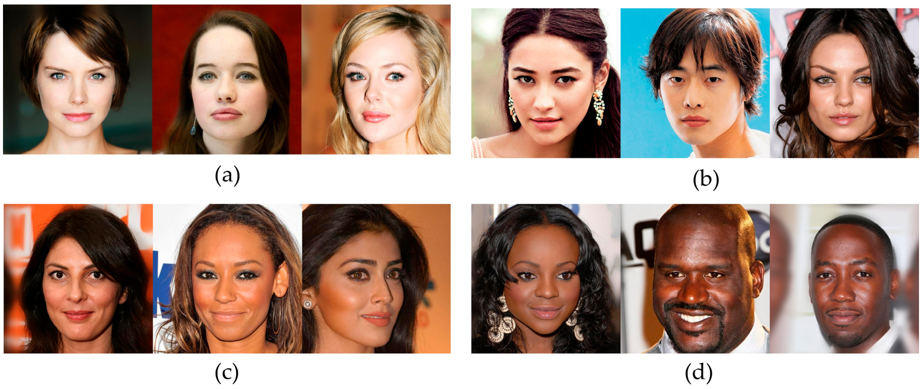

Figure 5. Samples based on the skin tone: (**a**) pale white, (**b**) light brown, (**c**) dark brown, (**d**) deep dark brown.

The CelebA-HQ dataset is considered a collection of original images, and the associated sketches can be obtained using Adobe Photoshop, as shown in Figure 6. The workflow to generate sketches from Photoshop includes serval steps. After opening the image, the process to be followed is to duplicate the background layer to preserve the original image and then convert the duplicated layer to grayscale. To preserve the characteristics of the sketches, Gaussian blur is applied to smooth the details of images. The negative image is created by utilizing the function of inverting color. Finally, a sketch-like effect can be generated by setting the blending mode of the duplicate layer. There is a set of paired images, in which each sketch image is associated with its corresponding original image. The pairs are used for image-to-image translation, in which the goal is to transform an image from a sketch into a second realistic image. In the image-to-image generative model, the paired images will be used as markers in training. The sketch image is the input of the generated model during training, and the original image is the expected output of the generated model.

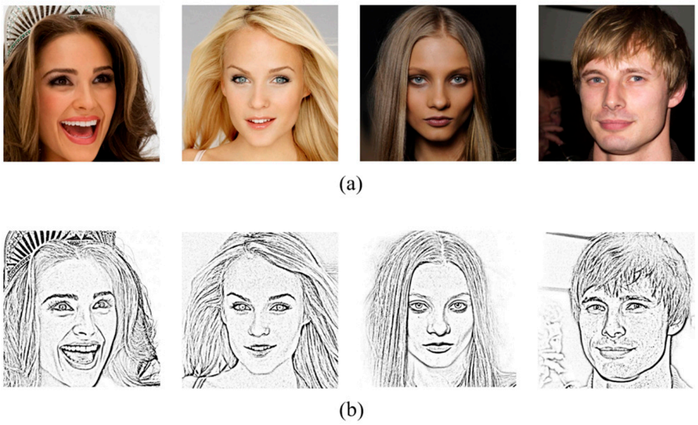

Figure 6. Samples of sketches: (**a**) real images and (**b**) sketches.

To assess the effectiveness of different data augmentation methods in improving image generation results, there was a particular focus on horizontal and vertical sketches. The two methods evaluated were left–right rotation and left–right flip. Experimental results indicated that left–right rotation yielded more accurate image generation for horizontal sketches, while left–right flip produced better results for vertical sketches. However, left–right rotation led to discoloration in certain areas, such as teeth, in vertical sketches, as shown as in Figures 7 and 8. Considering the typical drawing direction of users vertically, the left–right flip was chosen as the preferred data augmentation method for the dataset.

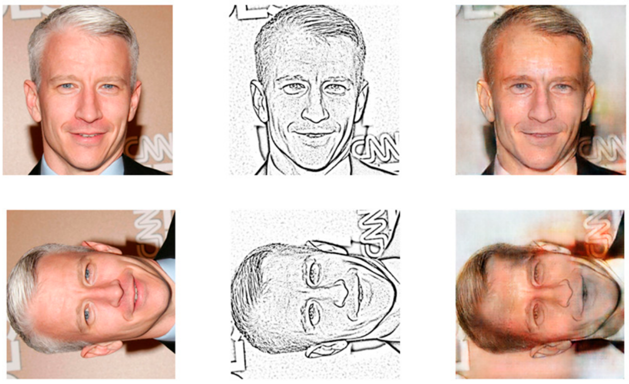

Figure 7. Image generation with left–right flip augmentation: real images (**left**), sketches (**middle**), and generated images (**right**).

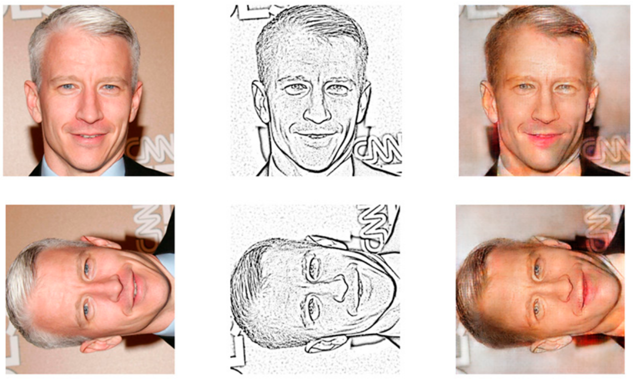

Figure 8. Image generation with left–right rotation augmentation: real images (**left**), sketches (**middle**), and generated images (**right**).

While preparing the images, it was also observed that some photos exhibited face occlusion or poor lighting conditions, as depicted in Figure 9. Using these photos as real images could potentially affect the accuracy of the discrimination judgments between real and fake images during the training process. Consequently, the training effectiveness of the generative model may be constrained, as the feedback of the discriminator is essential for guiding the learning process of the generator. These flawed photos were excluded from the self-built dataset in this study. In addition, it was observed that many photos appeared clear at first glance, but upon magnification, the noise became apparent, as illustrated in Figure 10. This situation could potentially lead the generative model to extract unnecessary features from the noisy images, ultimately degrading the quality of the generated output. Therefore, it is imperative to meticulously examine each photo by zooming in to check for noise and optimize the image quality of the dataset as much as possible. In CelebA-HQ, the majority of the images are of American celebrities. The dataset may likely be skewed towards individuals with lighter skin tones. Gathering a diverse set of images representing different skin tones can indeed be a challenging task. Combining multiple datasets with different color tones can help overcome the limitations in using a single source. Also, data augmentation techniques could be helpful for artificially increasing the diversity of the dataset. In this work, a total of 224 images were prepared for each of the four following datasets: Pale White, Light Brown, Dark Brown, and Deep Dark Brown. After image augmentation, 1792 images were prepared for the consequent model training.

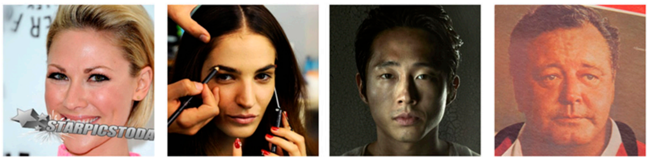

Figure 9. Flawed photos.

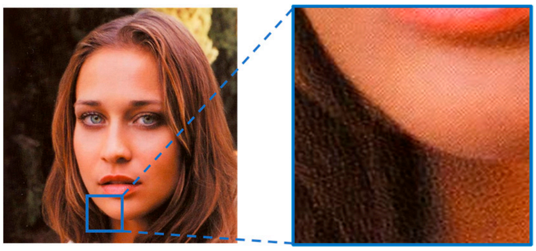

Figure 10. Noisy photo.

2.3.2. Model Training

In this paper, 70% of the total images are used for model training and the rest of the images, the remaining 30%, are used for validation. During each training epoch, the model learns from this set of images, adjusting its parameters to minimize the loss function. The validation set is used to evaluate the model performance after each epoch. The input of the traditional generative model is a multi-dimensional vector. It is often the case that the up-convolution layers are used in the decoder to transform the low-dimensional latent representation into a higher-dimensional output resembling the original input. Unlike the traditional GAN, the generative model of this study is applied to the image transformation, and the inputs to the generative model are the sketch images. This paper uses CNN as the encoder to construct the model in the form of an Encoder–Decoder [51,52]. The trained generative model can extract the facial features of the sketch image in the encoder stage and fill in the color features of the color portrait in the decoder stage. The goal of this work is to convert color-skinned portraits from sketch images. To preserve the image structure of the original sketch image as much as possible, the U-Net architecture is added to the model. The sub-layers of the Encoder–Decoder can be concatenated, and the spatial information in images can be preserved.

The training process of the traditional Generative Adversarial Network is depicted in Figure 11. A sketch from its paired image is fed into the generator for image generation, and the generated image is called a fake image. Both the fake and real images will be fed into the discriminator for distinction. The discrimination results will be fed back, not only to the generator, but also to the discriminator itself. The generator is updated based on the feedback from the discriminator when the generated fake image is judged as fake. By iteratively updating the parameters of the generator and discriminator based on the feedback from each other, the GAN learns to generate increasingly realistic images.

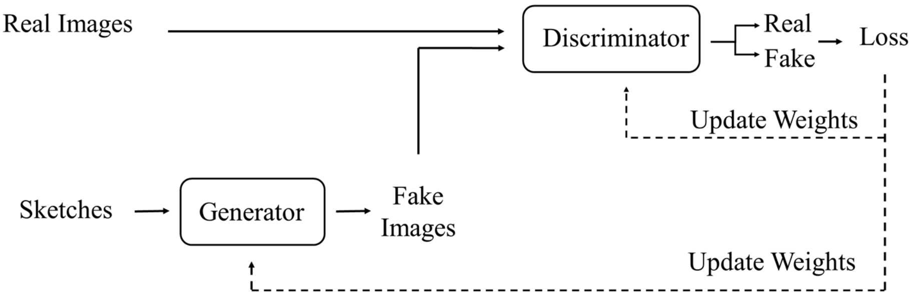

Figure 11. Traditional Generative Adversarial Network.

The proposed ethnicity-enhanced or color-enhanced Generative Adversarial Network (GAN) learning process involves training multiple generative models to specialize in generating images of people of different ethnicities or skin colors, as shown in Figure 12. In the case of the ethnicity-enhanced training process, the real images are divided into four racial groups: Caucasian, African American, Asian, and Middle Eastern. In the case of the skin color-enhanced training process, the real images are categorized into four skin color groups: pale white, light brown, dark brown, and deep dark brown. Paired sketch images corresponding to each racial or skin color group are fed into individual generative models. Each generative model is trained to generate images specific to its assigned racial or skin color group. The real images and their associated fake images produced by each generator are fed into a discriminator. The discriminator evaluates the authenticity of the real and fake images and provides feedback in the form of loss. The parameters of each generative model are updated based on the loss calculated from the discriminator's evaluation, aiming to improve the quality of the generated images relative to its assigned racial or skin color group. By training multiple generative models in this manner, the proposed GAN framework can effectively address the challenge of generating images with varied racial or skin color characteristics.

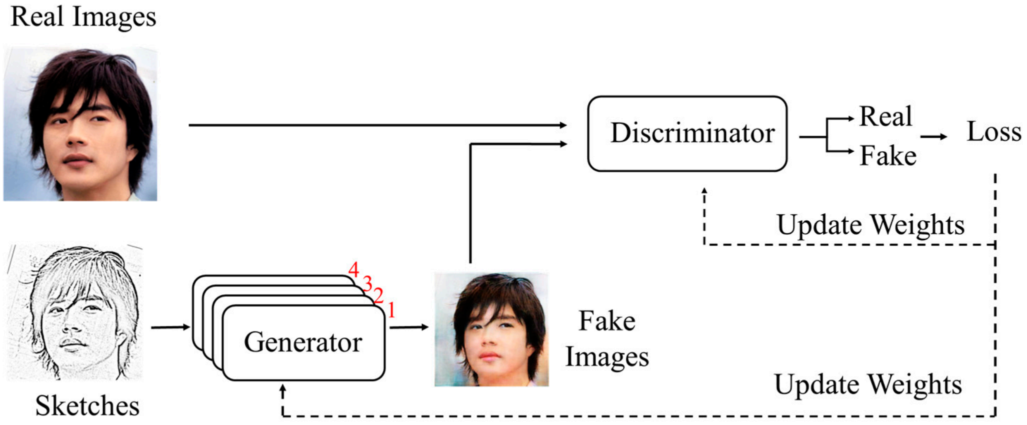

Figure 12. Ethnicity-enhanced/skin color-enhanced Generative Adversarial Network.

3. Experimental Results

3.1. Epoch Training Results

In the context of deep learning training, an epoch refers to one complete pass through the entire training dataset. The number of epochs represents how many times the learning algorithm will work through the entire training dataset. In general, too few epochs may result in underfitting, and too many epochs may lead to overfitting. In this work, the testing results regarding different numbers of epochs are shown in Figure 13. During the training process, the images in the image samples are randomly selected for testing. Significant improvements in facial features can be observed in the generated images from epoch 5 to epoch 50. Moreover, the improvement from epoch 100 to 200 is much more subtle. In this paper, the number of epochs selected for the model training is 200.

3.2. Ethnicity-Enhanced and Skin Color-Enhanced Training

Figure 14 renders the training result of the ethnographic classification dataset. From top to bottom, the top is the real image, the second row is the sketch image entering the generation model, and then at bottom is the training result of the original CelebA-HQ dataset. It can be seen that only single-skin-color images can be generated. The original CelebA-HQ dataset, being derived from celebrity faces, might have a lack of representation of diverse ethnicities and skin tones. As in Figure 12, the results generated by individual ethnicity-enhanced GAN are shown in Figure 14, where the blue box is the generated image corresponding to its real ethnography. For purposes of visual comparison, the generated images from pix2pix are also depicted in Figure 14. It is obvious that the blue-boxed images are much more similar to their real counterparts. For the purpose of skin color-enhanced GAN learning, the real images are categorized into four groups: Pale White, Light Brown, Dark Brown, and Deep Dark Brown. The testing results of the skin color-enhanced GAN model are shown in Figures 15–18. To describe the points of resemblance between the generated and real images, blue boxes are used to highlight the feasibility of the proposed skin color-enhanced image generation.

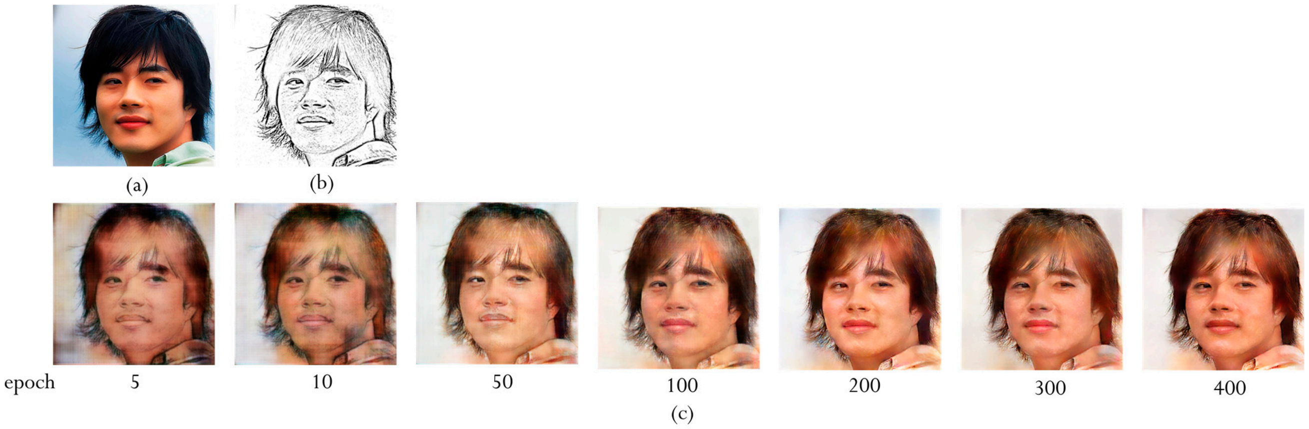

Figure 13. Image generation with different epochs: (**a**) real image, (**b**) sketch, and (**c**) generated images.

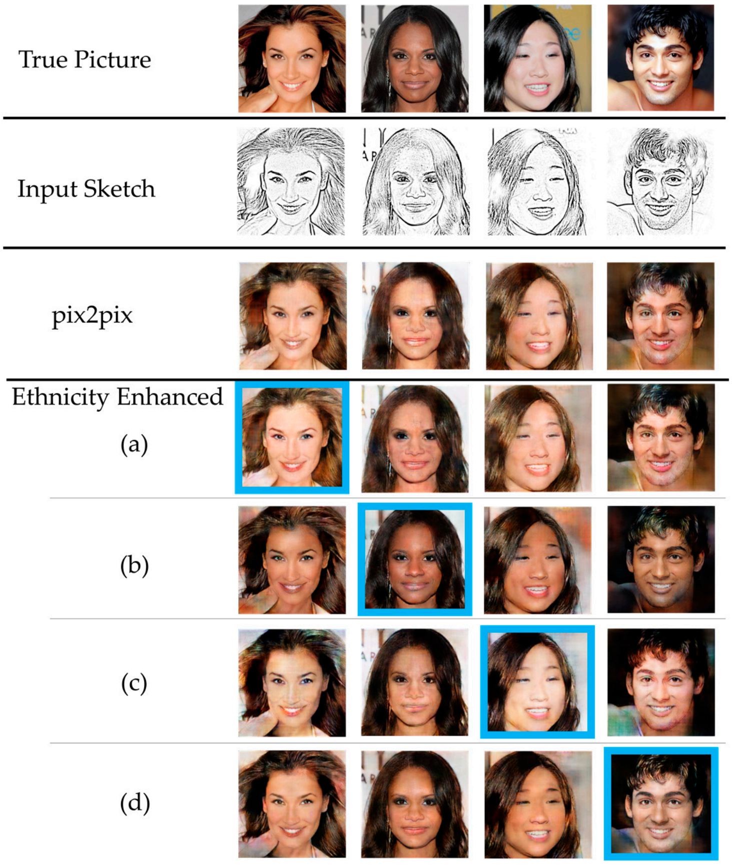

Figure 14. Image generation with ethnicity-enhanced GAN: (**a**) Caucasian, (**b**) African American, (**c**) Asian, and (**d**) Middle Eastern.

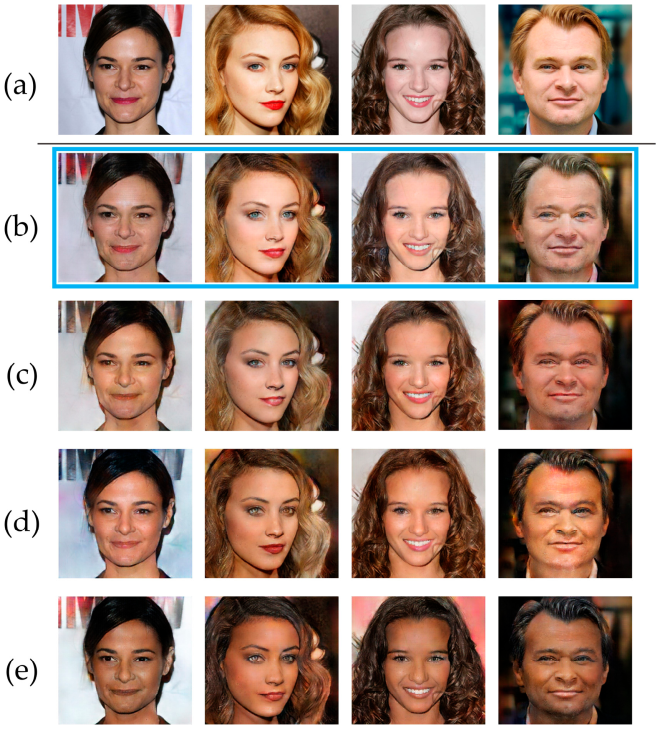

Figure 15. Image generation with skin color-enhanced GAN: (**a**) real image (pale white), (**b**) pale white, (**c**) light brown, (**d**) dark brown, and (**e**) deep dark brown.

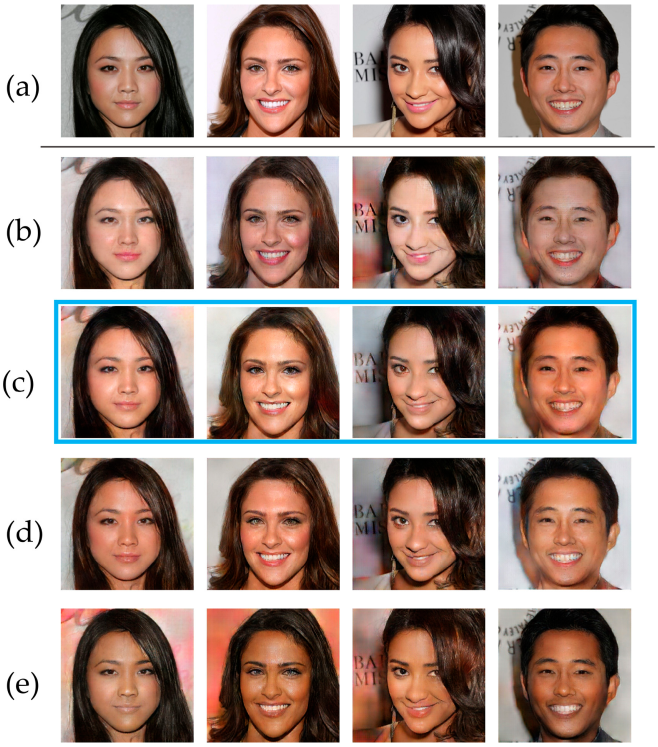

Figure 16. Image generation with skin color-enhanced GAN: (**a**) real image (light brown), (**b**) pale white, (**c**) light brown, (**d**) dark brown, and (**e**) deep dark brown.

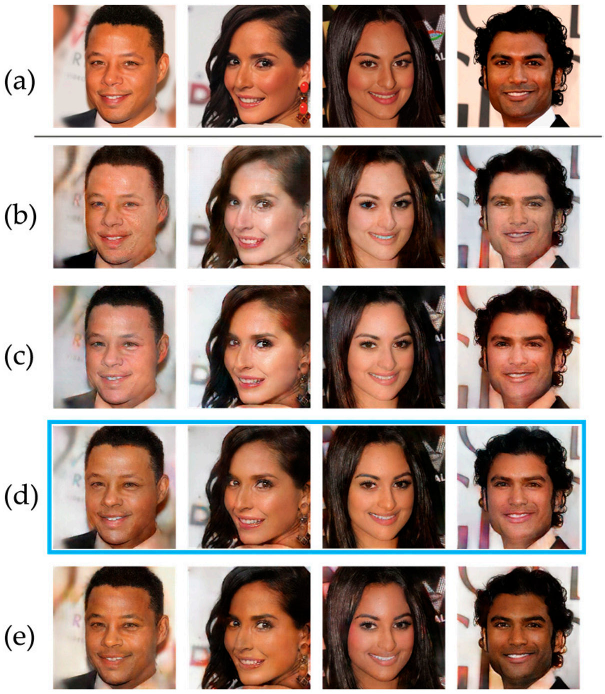

Figure 17. Image generation with skin color-enhanced GAN: (**a**) real image (dark brown), (**b**) pale white, (**c**) light brown, (**d**) dark brown, and (**e**) deep dark brown.

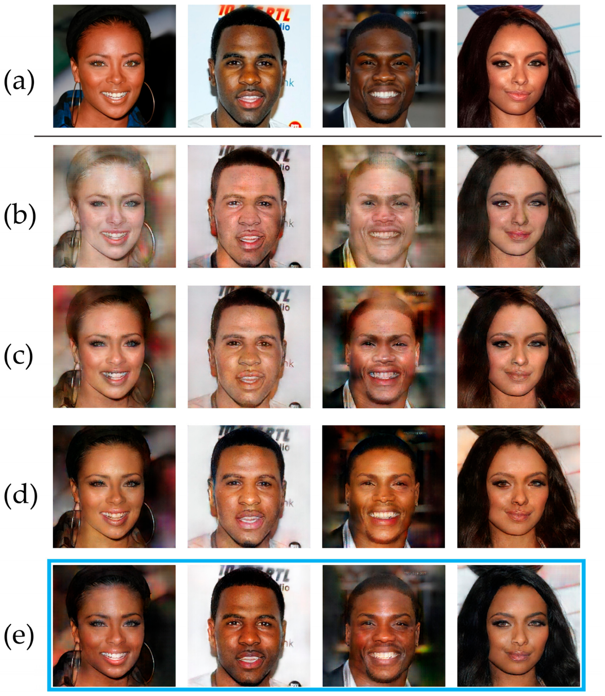

Figure 18. Image generation with skin color-enhanced GAN: (**a**) real image (deep dark brown), (**b**) pale white, (**c**) light brown, (**d**) dark brown, and (**e**) deep dark brown.

3.3. Test Comparisons

In the context of Generative Adversarial Networks, there are two main categories; one is information-to-image, and the other is image-to-image. In information-to-image applications, the goal is to generate images based on textual descriptions. The quality assessment of a generative model, especially in information-to-image applications, often involves evaluating whether the generated result can be identified as the desired object or concept described in the input text. In image-to-image applications, the primary goal is to generate an output image that closely resembles a given real input image. The quality assessment involves examining the generated images in comparison to real images or ground truth images. In the area of image-to-image generation, commonly addressed criteria include the Structural Similarity Index Measure (SSIM) and the Feature Similarity Index Measure (FSIM) [53–56]. SSIM is a widely used metric for assessing the structural similarity between two images. It quantifies the perceived quality of an image by considering luminance, contrast, and structure. FSIM is another metric designed to evaluate the quality of images, which it performs by measuring the similarity in feature space. The ethnicity-enhanced and skin tone-enhanced generation models were compared with the benchmark test of SSIM and FSIM with pix2pix, as shown in Tables 1 and 2. All values were averaged over 70 images. It can be concluded that the ethnicity-enhanced model is better than the pix2pix, and the skin color-enhanced is the best among the three image translation models. Compared to the traditional pix2pix, there is an average improvement of 14.3% in SSIM and 5.3% in FSIM with the skin color-enhanced mode. Samples of sketch-to-image are shown in Figures 19 and 20.

Table 1. Comparison with SSIM benchmark.

	PW	LB	DB	DDB	Average
pix2pix	0.702	0.657	0.733	0.756	0.712
Ethnicity-enhanced	0.703	0.710	0.746	0.792	0.738
Skin color-enhanced	0.798	0.836	0.807	0.813	0.814

Table 2. Comparison with FSIM benchmark.

	PW	LB	DB	DDB	Average
pix2pix	0.633	0.634	0.648	0.615	0.633
Ethnicity-enhanced	0.632	0.637	0.662	0.633	0.641
Skin color-enhanced	0.638	0.681	0.686	0.657	0.666

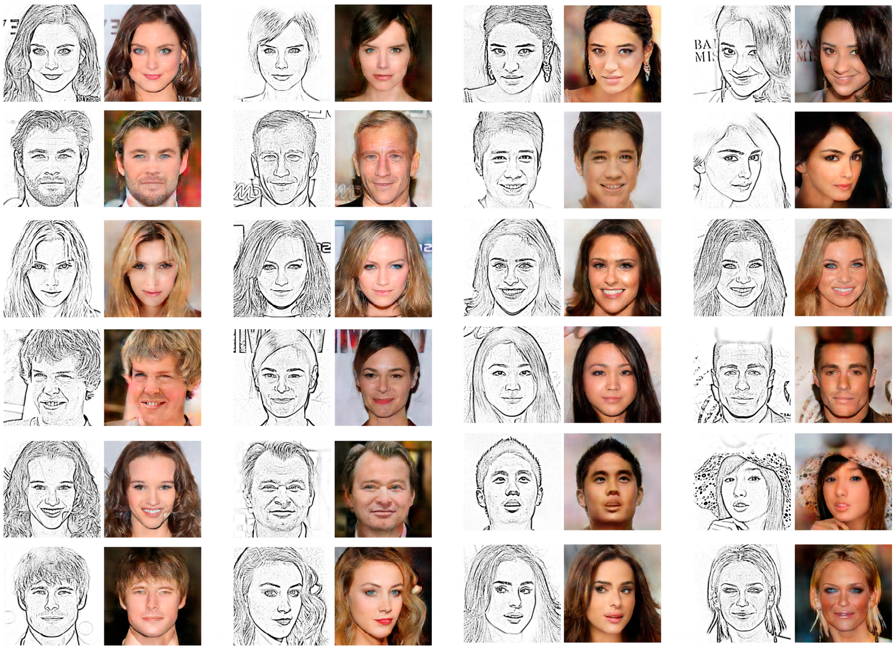

Figure 19. Sketch-to-image generation: pale white and light brown.

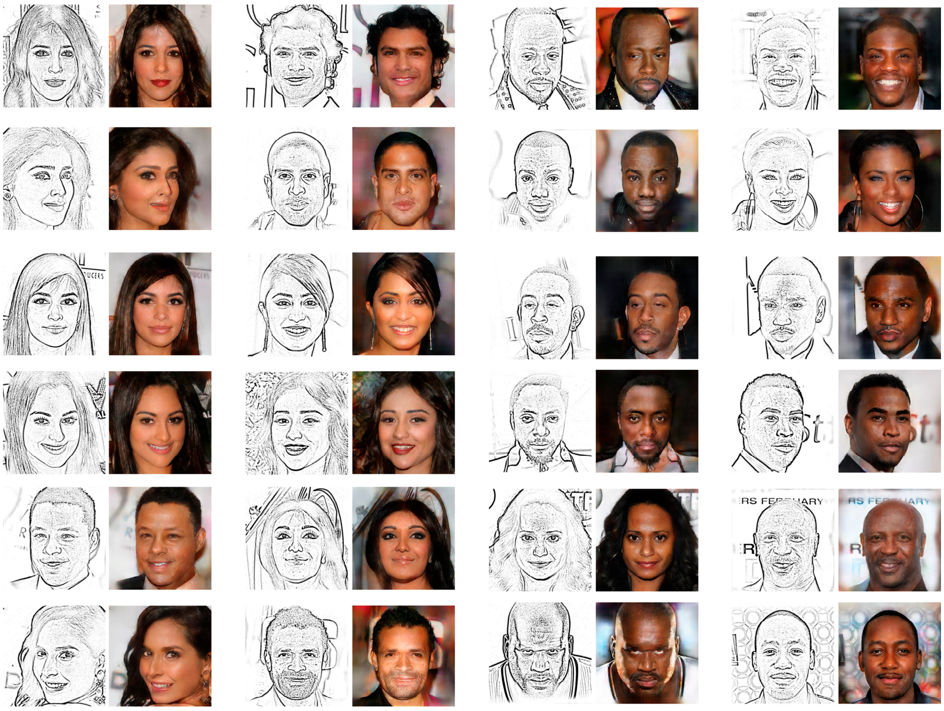

Figure 20. Sketch-to-image generation: dark brown and deep dark brown.

4. Discussion

The quality of the generation of color skin images has been improved with the ethnicity-enhanced or skin color-enhanced GAN model. However, there exists a problem in the background generation. In Figure 21, there are two images in a group; the left image is the real image, and the right image is the generated image. Even though the proposed enhanced models successfully generate face images, there are challenges with generating corresponding backgrounds. The differences in style and color among the backgrounds is too large, making it difficult to fully learn the relationship during training. Addressing this challenge requires handling diverse background styles and colors. We have considered this problem when building the training dataset. We have also tried to choose candidate images with less difference in their backgrounds. In practice, nearly 27,000 images out of the 30,000 images available in the CelebA-HQ dataset were selected for the current training set. If the criterion of background selection is further raised, the number of image samples in the dataset may be insufficient.

There are two ways to solve this problem. The first is to increase the source of images, choosing the same background style as much as possible. Due to the possibility of diverse backgrounds as to both color and style, the scale of the source images should be increased. The second method is to remove the background and retain the face portrait, which can make the generation network focus on the desired task without the complexity introduced by diverse backgrounds. Based on the current training datasets, the generation of face images is attained, but the generation of backgrounds still needs to improve. When training a generative model, especially in scenarios where real images have diverse backgrounds, issues with the discriminator misjudging real images as fake or vice versa can occur. The produced loss during model training is biased, and the face generation could be adversely affected with respect to its color.

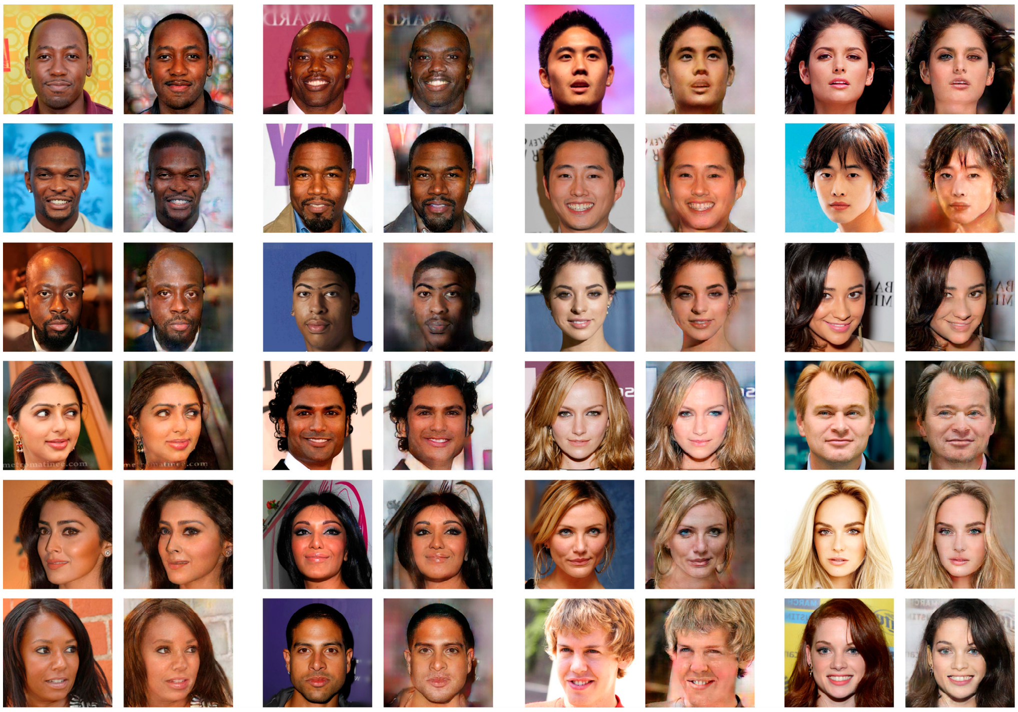

Figure 21. Sketch-to-image generation with diverse backgrounds, group images: real images (**left**) and generated images (**right**).

5. Conclusions

This paper utilizes the Generative Adversarial Network architecture to improve the training of the generative network, leading to the development of a system capable of transforming sketch images into color portraits. The U-Net architecture was employed to construct the generative model, ensuring the retention of facial features. Meanwhile, a convolutional neural network was utilized as the discrimination model to iteratively train with the generative model. By collecting and classifying numerous images, two image datasets were established, categorized by ethnicity and skin color. This classification enabled the improvement of the quality of image generation. In addition, an additional dataset was created, comprising sketches paired with their corresponding real images.

This study aims to address a limitation of traditional generation models, specifically their ability to generate only a single skin tone. This work has successfully trained generators that can generate images with four different skin tones. Experimental tests were conducted to validate the effectiveness of the trained models. The experimental results indicate that the skin color-enhanced model has demonstrated improvements compared to a traditional image generation model, as measured by SSIM and FSIM. In the future, the multi-skin color image generation model trained in this paper can be used to expand the sample size of the dataset and further optimize the image quality. When preparing the dataset, Mask R-CNN could be employed to remove the background. This would help avoid interference from the background for both the generation model and the discrimination model during the training process.

Author Contributions: Conceptualization, Y.-H.C. (Yeong-Hwa Chang) and P.-H.C.; methodology, Y.-H.C. (Yeong-Hwa Chang) and P.-H.C.; software, P.-H.C.; validation, P.-H.C. and Y.-H.C. (Yu-Hsiang Chai); formal analysis, Y.-H.C. (Yeong-Hwa Chang) and P.-H.C.; investigation, Y.-H.C. (Yeong-Hwa Chang) and H.-W.L.; resources, Y.-H.C. (Yeong-Hwa Chang) and P.-H.C.; data curation, P.-H.C. and Y.-H.C. (Yeong-Hwa Chang); writing—original draft preparation, Y.-H.C. (Yeong-Hwa Chang) and P.-H.C.; writing—review and editing, Y.-H.C. (Yeong-Hwa Chang) and H.-W.L.; visualization, Y.-H.C. (Yeong-Hwa Chang) and Y.-H.C. (Yu-Hsiang Chai); supervision, Y.-H.C. (Yeong-Hwa Chang) and H.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in https://github.com/nperraud/download-celebA-HQ and https://paperswithcode.com/dataset/celeba-hq (accessed on 26 December 2023).

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

References

1. Elfaki, A.O.; Messoudi, W.; Bushnag, A.; Abuzneid, S.; Alhmiedat, T. A Smart Real-Time Parking Control and Monitoring System. *Sensors* **2023**, *23*, 9741. [CrossRef] [PubMed]
2. Zhang, W. An improved DBSCAN Algorithm for Hazard Recognition of Obstacles in Unmanned Scenes. *Soft. Comput.* **2023**, *27*, 18585–18604. [CrossRef]
3. Chang, Y.-H.; Zhang, Y.-Y. Deep Learning for Clothing Style Recognition Using YOLOv5. *Micromachines* **2022**, *13*, 1678. [CrossRef] [PubMed]
4. Lee, J.C.; Kim, Y.; Moon, S.; Ko, J.H. A Reconfigurable Neural Architecture for Edge–Cloud Collaborative Real-Time Object Detection. *IEEE Internet Things J.* **2022**, *9*, 23390–23404. [CrossRef]
5. Eversberg, L.; Lambrecht, J. Generating Images with Physics-Based Rendering for an Industrial Object Detection Task: Realism versus Domain Randomization. *Sensors* **2021**, *21*, 7901. [CrossRef] [PubMed]
6. Wang, G.; Zhou, M.; Wei, X.; Yang, G. Vehicular Abandoned Object Detection Based on VANET and Edge AI in Road Scenes. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14254–14266. [CrossRef]
7. Joshi, R.C.; Khan, J.S.; Pathak, V.K.; Dutta, M.K. AI-CardioCare: Artificial Intelligence Based Device for Cardiac Health Monitoring. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 1292–1302. [CrossRef]
8. Lee, S.E.; Lee, E.; Kim, E.-K.; Yoo, J.H.; Park, V.Y.; Youk, J.H.; Kwak, J.Y. Application of Artificial Intelligence Computer-Assisted Diagnosis Originally Developed for Thyroid Nodules to Breast Lesions on Ultrasound. *J. Digit. Imaging* **2022**, *35*, 1699–1707. [CrossRef]
9. Samaddar, P.; Mishra, A.K.; Gaddam, S.; Singh, M.; Modi, V.K.; Gopalakrishnan, K.; Bayer, R.L.; Igreja Sa, I.C.; Shalil Khanal, S.; Hirsova, P.; et al. Machine Learning-Based Classification of Abnormal Liver Tissues Using Relative Permittivity. *Sensors* **2022**, *22*, 9919. [CrossRef]
10. Behzadipour, F.; Raeini, M.; Mehdizadeh, S.; Taki, M.; Moghadam, B.; Bavani, M.; Lloret, J. A Smart IoT-based Irrigation System Design using AI and Prediction Model. *Neural Comput. Appl.* **2023**, *35*, 24843–24857. [CrossRef]
11. Taneja, A.; Rani, S.; Breñosa, J.; Tolba, A.; Kadry, S. An improved WiFi Sensing based Indoor Navigation with Reconfigurable Intelligent Surfaces for 6G Enabled IoT Network and AI Explainable Use Case. *Future Gener. Comput. Syst.* **2023**, *19*, 294–303. [CrossRef]
12. Chuang, S.-Y.; Sahoo, N.; Lin, H.-W.; Chang, Y.-H. Predictive Maintenance with Sensor Data Analytics on a Raspberry Pi-Based Experimental Platform. *Sensors* **2019**, *19*, 3884. [CrossRef] [PubMed]
13. Zhang, S.; Yu, S.; Ding, H.; Hu, J.; Cao, L. CAM R-CNN: End-to-End Object Detection with Class Activation Maps. *Neural Process. Lett.* **2023**, *55*, 10483–10499. [CrossRef]
14. Lyu, H.; Qiu, F.; An, L.; Stow, D.; Lewison, R.; Bohnett, E. Deer survey from drone thermal imagery using enhanced faster R-CNN based on ResNets and FPN. *Ecol. Inform.* **2024**, *79*, 102383. [CrossRef]
15. Li, L.; Wang, X.; Yang, M.; Zhang, H. An Accurate Shared Bicycle Detection Network based on Faster R-CNN. *IET Image Process* **2023**, *17*, 1919–1930. [CrossRef]
16. Chen, S.; Li, Z.; Tang, Z. Relation R-CNN: A Graph Based Relation-Aware Network for Object Detection. *IEEE Signal Process. Lett.* **2020**, *27*, 1680–1684. [CrossRef]
17. Butler, J.; Leung, H. A Novel Keypoint Supplemented R-CNN for UAV Object Detection. *IEEE Sens. J.* **2023**, *23*, 30883–30892. [CrossRef]
18. Wan, C.; Chang, X.; Zhang, Q. Improvement of Road Instance Segmentation Algorithm Based on the Modified Mask R-CNN. *Electronics* **2023**, *12*, 4699. [CrossRef]
19. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
20. Zhang, Y.; Hu, B.; Huang, Y.; Gao, C.; Yin, J.; Wang, Q. HQ-I2IT: Redesign the Optimization Scheme to Improve Image Quality in CycleGAN-based Image Translation Systems. *IET Image Process* **2024**, *18*, 507–522. [CrossRef]
21. Liang, Z.; Huang, J.X.; Antani, S. Image Translation by Ad CycleGAN for COVID-19 X-Ray Images: A New Approach for Controllable GAN. *Sensors* **2022**, *22*, 9628. [CrossRef] [PubMed]
22. Liao, W.; Huang, Y.; Zheng, Z.; Lu, X. Intelligent Generative Structural Design Method for Shear Wall Building based on Fused-text-Image-to-Image Generative Adversarial Networks. *Expert Syst. Appl.* **2022**, *210*, 118530. [CrossRef]
23. Xu, Z.; Wu, S.; Jiao, Q.; Wong, H.-S. TSEV-GAN: Generative Adversarial Networks with Target-aware Style Encoding and Verification for Facial Makeup Transfer. *Knowl.-Bases Syst.* **2022**, *257*, 109958. [CrossRef]
24. Naveen, S.; Kiran, M.S.R.; Indupriya, M.; Manikanta, T.V.; Sudeep, P.V. Transformer Models for Enhancing AttnGAN based Text to Image Generation. *Image Vis. Comput.* **2021**, *115*, 104284. [CrossRef]

25. Zhou, M.; Liu, X.; Yi, T.; Bai, Z.; Zhang, P. A Superior Image Inpainting Scheme using Transformer-based self-supervised Attention GAN Model. *Expert Syst. Appl.* **2023**, *233*, 120906. [CrossRef]
26. Zhou, X.; Tian, K.; Zhou, Z.; Ning, B.; Wang, Y. SID-TGAN: A Transformer-Based Generative Adversarial Network for Sonar Image Despeckling. *Remote Sens.* **2023**, *15*, 5072. [CrossRef]
27. Zhang, C.; Zhou, L.; Xiao, X.; Xu, D. A Missing Traffic Data Imputation Method Based on a Diffusion Convolutional Neural Network–Generative Adversarial Network. *Sensors* **2023**, *23*, 9601. [CrossRef] [PubMed]
28. Özbey, M.; Dalmaz, O.; Dar, S.; Bedel, H.A.; Özturk, S.; Güngör, A.; Çukur, T. Unsupervised Medical Image Translation with Adversarial Diffusion Models. *IEEE Trans. Med. Imaging* **2023**, *42*, 3524–3539. [CrossRef]
29. Xiao, H.; Wang, X.; Wang, J.; Cai, J.-Y.; Deng, J.-H.; Yan, J.-K.; Tang, Y.-D. Single Image Super-Resolution with Denoising Diffusion GANs. *Sci. Rep.* **2024**, *14*, 4272. [CrossRef]
30. Zhang, H.; Sindagi, V.; Patel, V.M. Image De-Raining Using a Conditional Generative Adversarial Network. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3943–3956. [CrossRef]
31. Yang, H.; Guo, J.; Xin, Y.; Cai, J.; Zhang, M.; Zhao, X.; Zhao, Y.; He, Y. Multi-scale Fusion and Adaptively Attentive Generative Adversarial Network for Image De-raining. *Appl. Intell.* **2023**, *53*, 30954–30970. [CrossRef]
32. Lu, B.; Gai, S.; Xiong, B.; Wu, J. Single Image Deraining with Dual U-Net Generative Adversarial Network. *Multidimens. Syst. Signal Process.* **2022**, *33*, 485–499. [CrossRef]
33. Bansal, N.; Sridhar, S. HEXA-GAN: Skin Lesion Image Inpainting via Hexagonal Sampling based Generative Adversarial Network. *Biomed. Signal Process. Control* **2024**, *89*, 105603. [CrossRef]
34. He, L.; Zhenping Qiang, Z.; Shao, X.; Lin, H.; Wang, M.; Da, F. Research on High-Resolution Face Image Inpainting Method based on StyleGAN. *Electronics* **2022**, *11*, 1620. [CrossRef]
35. Du, Q.; Ren, X.; Wang, J.; Qiang, Y.; Yang, X.; Kazihise, N. Iterative PET Image Reconstruction using Cascaded Data Consistency Generative Adversarial Network. *IET Image Process.* **2020**, *14*, 3989–3999. [CrossRef]
36. Wei, W.; Zhang, D.; Wang, H.; Duan, X.; Guo, C. Utilizing the Neural Renderer for Accurate 3D Face Reconstruction from a Single Image. *Neural Process. Lett.* **2023**, *55*, 10535–10553. [CrossRef]
37. Wang, Y.; Luo, Y.; Zu, Y.; Zhan, C.; Jiao, B.; Wu, Z.; Zhou, X.; Shen, D.; Zhou, L. 3D Multi-modality Transformer-GAN for High-quality PET Reconstruction. *Med. Image Anal.* **2024**, *91*, 102983. [CrossRef]
38. Guo, Y.; Chen, Q.; Chen, J.; Wu, Q.; Shi, Q.; Tan, M. Auto-Embedding Generative Adversarial Networks for High Resolution Image Synthesis. *IEEE Trans. Multimed.* **2019**, *21*, 2726–2737. [CrossRef]
39. Sushko, V.; Zhang, D.; Gall, J.; Khoreva, A. Generating Novel Scene Compositions from Single Images and Videos. *Comput. Vis. Image Underst.* **2024**, *239*, 103888. [CrossRef]
40. Sharma, O.; Sharma, A.; Kalia, A. MIGAN: GAN for Facilitating Malware Image Synthesis with Improved Malware Classification on Novel Dataset. *Expert Syst. Appl.* **2024**, *241*, 122678. [CrossRef]
41. Zhang, W.; Fu, C.; Chang, X.; Zhao, T.; Li, X.; Sham, C.-W. A More Compact Object Detector Head Network with Feature Enhancement and Relational Reasoning. *Neurocomputing* **2022**, *499*, 23–34. [CrossRef]
42. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
43. Li, S.; Lin, J.; Yang, X.; Ma, J.; Chen, Y. BPFD-Net: Enhanced Dehazing Model based on Pix2pix Framework for Single Image. *Mach. Vis. Appl.* **2021**, *32*, 124. [CrossRef]
44. Sajichandrachood, O.M.; Sethunadh, R. Detection and Segmentation of Radio Frequency Interference from Satellite Images using Attention-GANs. *Astron. Comput.* **2023**, *45*, 100769.
45. Muyuan Liu, M.; Su, X.; Yao, X.; Hao, W.; Zhu, W. Lensless Image Restoration Based on Multi-Stage Deep Neural Networks and Pix2pix Architecture. *Photonics* **2023**, *10*, 1274.
46. Fujioka, T.; Satoh, Y.; Imokawa, T.; Mori, M.; Yamaga, E.; Takahashi, K.; Kubota, K.; Onishi, H.; Tateishi, U. Proposal to Improve the Image Quality of Short-Acquisition Time-Dedicated Breast Positron Emission Tomography Using the Pix2pix Generative Adversarial Network. *Diagnostics* **2022**, *12*, 3114. [CrossRef] [PubMed]
47. Chen, T.; Zhang, X.; Hamann, B.; Wang, D.; Zhang, H. A Multi-level Feature Integration Network for Image Inpainting. *Multimed. Tools Appl.* **2022**, *81*, 38781–38802. [CrossRef]
48. YUN Pang, Y.; Mao, J.; He, L.; Lin, H.; Qiang, Z. An Improved Face Image Restoration Method Based on Denoising Diffusion Probabilistic Models. *IEEE Access* **2024**, *12*, 3581–3596. [CrossRef]
49. Man, Q.; Cho, Y.-I.; Jang, S.-G.; Lee, H.-J. Transformer-Based GAN for New Hairstyle Generative Networks. *Electronics* **2022**, *11*, 2106. [CrossRef]
50. Shen, L.; Yan, J.; Sun, X.; Li, B.; Pan, Z. Wavelet-Based Self-Attention GAN With Collaborative Feature Fusion for Image Inpainting. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 1651–1664. [CrossRef]
51. Salem, M.; Valverde, S.V.; Cabezas, M.; Pareto, D.; Oliver, A.; Salvi, J.; Rovira, A.; Vier Llado, X. Multiple Sclerosis Lesion Synthesis in MRI Using an Encoder-Decoder U-NET. *IEEE Access* **2019**, *7*, 25171–25184. [CrossRef]
52. Saha, A.; Zhang, Y.-D.; Satapathy, S.C. Brain Tumour Segmentation with a Muti-Pathway ResNet Based UNet. *J. Grid Comput.* **2021**, *19*, 43. [CrossRef]
53. Oyelade, O.N.; Ezugwu, A.E.; Almutairi, M.S.; Saha, A.K.; Abualigah, L.; Chiroma, H. A Generative Adversarial Network for Synthetization of Regions of Interest based on Digital Mammograms. *Sci. Rep.* **2022**, *12*, 6166. [CrossRef] [PubMed]

54. Wang, L.; Zhang, S.; Gu, L.; Zhang, J.; Zhai, X.; Sha, X.; Chang, S. Automatic Consecutive Context Perceived Transformer GAN for Serial Sectioning Image Blind Inpainting. *Comput. Biol. Med.* **2021**, *136*, 104751. [CrossRef] [PubMed]
55. Oyelade, O.N.; Ezugwu, A.E. EOSA-GAN: Feature Enriched Latent Space Optimized Adversarial Networks for Synthesization of Histopathology Images using Ebola Optimization Search Algorithm. *Biomed. Signal Process. Control Biomed. Signal Process. Control* **2023**, *84*, 104734. [CrossRef]
56. Manu, C.M.; Sreeni, K.G. GANID: A Novel Generative Adversarial Network for Image Dehazing. *Vis. Comput.* **2023**, *39*, 3923–3936. [CrossRef]

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.