electronics

MDPI

Article

Multi-Attention Infused Integrated Facial Attribute Editing Model: Enhancing the Robustness of Facial Attribute Manipulation

Zhijie Lin [1], Wangjun Xu [1,*], Xiaolong Ma [2,*], Caie Xu [1] and Han Xiao [3]

[1] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; linzhijie@zust.edu.cn (Z.L.); caiexu@163.com (C.X.)
[2] School of Management, Huzhou University, Huzhou 313000, China
[3] College of Science, University of Arizona, Tucson, AZ 85719, USA; hanxiao@arizona.edu
* Correspondence: 222108855052@zust.edu.cn (W.X.); xiaolongma@zjhu.edu.cn (X.M.)

Abstract: Facial attribute editing refers to the task of modifying facial images by altering specific target facial attributes. Existing approaches typically rely on the combination of generative adversarial networks and encoder–decoder architectures to tackle this problem. However, current methods may exhibit limited accuracy when dealing with certain attributes. The primary objective of this research is to enhance facial image modification based on user-specified target facial attributes, such as hair color, beard removal, or gender transformation. During the editing process, it is crucial to selectively modify only the regions relevant to the target attributes while preserving the details of other unrelated facial attributes. This ensures that the editing results appear more natural and realistic. This study introduces a novel approach called MAGAN (Combining GRU Structure and Additive Attention with AGU—Adaptive Gated Units). Moreover, a discriminative attention mechanism is introduced to automatically identify key regions in the input images that are relevant to facial attributes. This mechanism concentrates attention on these regions, enhancing the model's ability to accurately capture and analyze subtle facial attribute features. The method incorporates external attention within the convolutional layers of the encoder–decoder architecture, facilitating the modeling of linear complexity across image regions and implicitly considering correlations among all data samples. By employing discriminative attention in the discriminator, the model achieves more precise attribute editing. To evaluate the effectiveness of MAGAN, experiments were conducted on the CelebA dataset. The average precision of facial attribute generation in images edited by our model stands at 91.83%. PSNR and SSIM for reconstructed images are 32.52 and 0.957, respectively. In comparison with existing methodologies (AttGAN, STGAN, MUGAN), noteworthy enhancements have been achieved in the domain of facial attribute manipulation.

Keywords: facial attribute manipulation; adversarial generative networks; additive attention; external attention mechanism

Citation: Lin, Z.; Xu, W.; Ma, X.; Xu, C.; Xiao, H. Multi-Attention Infused Integrated Facial Attribute Editing Model: Enhancing the Robustness of Facial Attribute Manipulation. *Electronics* **2023**, *12*, 4111. https://doi.org/10.3390/electronics12194111

Academic Editor: Stefanos Kollias

Received: 31 August 2023
Revised: 20 September 2023
Accepted: 27 September 2023
Published: 30 September 2023

1. Introduction

Facial editing refers to the act of altering specific features of a particular countenance, such as gender, hair color, skin tone, and facial expression. This technology can vividly reproduce the facial images stored in people's minds and allow for effortless modifications. Its applications are extensive in domains like facial beautification [1] and human-computer interaction [2]. However, in the process of facial editing, some networks tend to focus more on certain attribute transformations while disregarding the preservation of the original identity features. This can result in inconsistent identity traits between the reconstructed image and the original one, and even distortion of facial features during the transformation. The challenge in current facial attribute editing lies in accurately transferring a given image from the source attribute domain to the target attribute domain while preserving

the attribute-independent details. With the continuous development of deep learning, more and more generative models have been proposed, including GAN [3], VAE [4], and Diffusion [5] models. These models can transform low-dimensional data by generating high-dimensional image data, thus making significant progress in the field of face generation and solving many difficult problems. First of all, GAN has a generator and a discriminator. The generator is used to generate fake images from random noise, and the discriminator determines whether the input is a real image or a generated image. The two are constantly growing stronger in a minimax mutual game. Since the model itself is adversarial, we need to train two models at the same time, so it is difficult to train, which makes it difficult to achieve an optimal balance, and it is prone to mode collapse due to its poor training stability. VAE generates images by converting the original high-dimensional input into low-dimensional hidden layer encoding through the encoder and reconstructing the data from the encoding through the decoder. In order to generate images, we can add constraints to the encoder to force it to generate potential that obeys Gaussian distribution. Simply sampling it and passing it to the decoder produces a picture, but the resulting image is blurry. The difference between the diffusion model and generative networks such as VAE and GAN is that the diffusion model has two main process areas: forward diffusion and reverse diffusion. In the forward diffusion stage, the image is contaminated by gradually introducing noise until the image becomes completely random noise. Diffusion models have recently shown remarkable performance in image generation tasks, but these models are computationally demanding and training requires very large memory, because all Markov states need to be in memory at all times for prediction, meaning multiple instances of large deep networks are in memory at all times. Huang et al. [6] proposed a cooperative diffusion model that can control face generation and editing through multiple modalities simultaneously, without retraining single-modal models, and showing superiority in image quality and conditional consistency. DCFace [7] proposes a dual-conditional face generator based on the diffusion model, which controls intra- and inter-class variations by combining subject appearance (ID) and external factor (style) conditions. The author uses a novel patch-wise style extractor and time-step-dependent ID loss which enables DCFace to consistently produce facial images of the same subject under different styles with precise control. In recent years, cutting-edge research [8–10] has been predominantly based on the encoder–decoder architecture, which extracts the representation of the source image and reconstructs it guided by the target attribute vector. Moreover, attribute-independent constraints are incorporated into the facial attribute classifier to ensure accurate attribute transformations. Among them, refs. [11,12] directly convert the input image into an image with the target attributes instead of editing the input image within the appropriate attribute regions, inevitably leading to unnecessary modifications in attribute-unrelated parts. STGAN [10] introduced selective transfer units, replacing the target attribute vector with the differential attribute vector as the model's input. This approach allows STGAN to focus on the attributes that require modification, greatly improving the quality of image reconstruction and enhancing the flexibility of attribute transformations. However, it should be noted that STGAN still does not explicitly consider the attribute editing regions, and is thus unable to guarantee perfect preservation of content details in attribute-unrelated areas. MUGAN [13] introduced self-attention layers as a complement to the convolutional layers in the encoder–decoder, providing assistance for the model's generation. The self-attention mechanism helps simulate long-range dependencies between image regions and facilitates complex geometric constraints for image generation in GANs. However, it is important to note that its ability to manipulate attributes still falls somewhat short. Additionally, incorporating self-attention can result in an excessively large model size and prolonged training time.

In Figure 1, the effects of attribute manipulation, specifically the attribute "Bangs", are not pronounced in AttGAN, STGAN, and MUGAN. In the case of "Gender", the results generated by AttGAN alter the length of the hair. While we expect STGAN to preserve the content unrelated to the edited attribute, its generated output modifies the content of

the beard. Regarding the "Bald" attribute, our editing results outperform those of other models. To address these issues, we introduce the concept of "external attention" [14] as a complementary element to the convolutional layers of the encoder–decoder architecture. While self-attention, which calculates the correlations within the same sample to capture long-range dependencies, has limitations due to its quadratic computational complexity and disregard for connections between different samples, the introduction of external attention, constructed with only two linear layers and two normalization layers, offers a linear computational complexity and better models the linear complexity across image regions, while implicitly considering correlations among all data samples. This aids in achieving complex attribute constraints in generated images, making our model more robust in attribute decoupling. Moreover, external attention exhibits enhanced attribute decoupling capabilities, aiding in executing complex attribute constraints in image generation. This optimization not only significantly reduces training time but also yields better results than self-attention. To address the limitations of attribute manipulation in the model, we propose a discriminative attention mechanism that reduces information reduction and amplifies global dimension interaction features, enhancing the capabilities of the attribute classifier. This mechanism ensures a "superior" and "more stringent" classification performance, constraining the generator to produce desired facial attributes and thereby resolving the shortcomings in attribute manipulation capabilities. Our contributions are as follows:

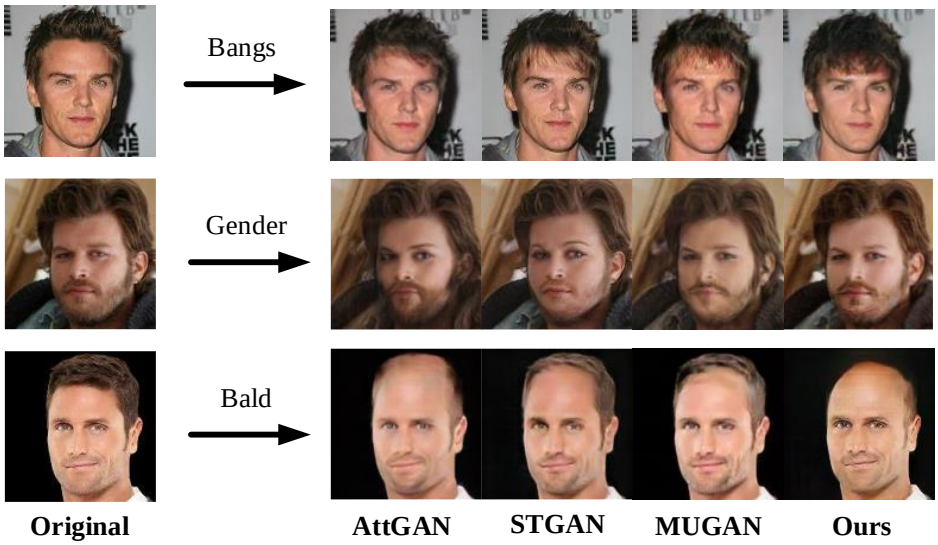

Figure 1. The facial editing results from AttGAN, STGAN, MU-GAN, and our approach, with the given target attributes being Bangs, Gender, and Bald.

(1) Enhanced attribute manipulation while preserving unrelated attributes: The integration of a novel skip-connection unit with the symmetric U-Net decoder-encoder architecture has elevated the capability to manipulate facial attributes. This contribution allows for precise attribute modification while preserving the integrity of unrelated attributes, ensuring that the editing process focuses only on the target attributes.

(2) Reduced model complexity with improved attribute decoupling: The incorporation of external attention mechanisms has reduced the complexity of the model. Simultaneously, this integration has enhanced the attribute decoupling capability, allowing for better separation and control over different facial attributes. Additionally, the image generation quality has improved, ensuring more realistic and visually appealing generated images.

(3) Strengthened attribute manipulation and discriminator performance: The proposal of a discriminative attention mechanism has brought about significant improvements in attribute manipulation capabilities. This mechanism enhances the discriminator's ability to distinguish and analyze facial attributes, thereby boosting the attribute classifier's effectiveness. This contributes to better attribute editing results and overall image quality.

2. Related Work

2.1. Encoder–Decoder Architecture

Hinton and Zemel [15] presented a seminal paper introducing an autoencoder network consisting of an encoder and a decoder. The encoder can process diverse inputs from CNNs, RNNs, and more, yielding feature vectors. Leveraging the feature vectors obtained by the encoder, the decoder generates results that closely resemble the desired output. In a related vein, Kingma and Welling [4] proposed the concept of variational autoencoders, which represents a specialized case within the encoder–decoder framework. The encoder learns the posterior distribution $p(z \mid x)$, encoding the input image x into latent space variables z. The decoder extracts new images from the latent vector z and generates novel samples akin to the input image. Notably, Nie et al. [16] employed a semi-supervised learning approach based on the VAE model to edit facial attributes in images.

2.2. Generative Adversarial Networks (GAN)

Currently, the majority of facial attribute modification methods are based on GANs. These models consist of a generator network G and a discriminator network D, engaged in a minimax game. The goal of the generator is to learn the distribution of the input data and generate new samples resembling the training samples, while the discriminator learns how to distinguish real samples from fake ones in the generator's output. GANs have been applied in various domains, with image synthesis being their origin and prime target. To integrate the merits of these two models, some approaches such as [17] have employed a hybrid model that combines VAEs and GANs. GANs, trained on a set of facial images, can generate similar facial photos based on input noise vectors. Since the introduction of GANs, many successful facial image generation models have been proposed. Among them, PgGAN [18] is an extension of GAN training that allows for more stable training of generator models capable of producing high-quality large-scale images. It starts from a very low-resolution image and gradually adds layers to increase the output size of the generator model and the input size of the discriminator model until the desired image size is achieved. StyleGAN [19] is an improvement upon the PgGAN architecture, enabling the control of separate style attributes of the created images by introducing inter-mediate latent spaces. StyleGAN generates high-resolution and realistic images, but sometimes unnatural speckles appear in the resulting images. In StyleGAN2 [20], adjustments have been made to the use of AdaIN, effectively avoiding these speckles.

2.3. Image Translation

Image translation refers to the transformation between images, which, from a more abstract perspective, involves mapping between different visual domains. For instance, the task of colorizing black and white images entails a mapping from the "grayscale domain" to the "color domain". Models such as Pix2Pix [21], CycleGAN, and StarGAN employ research approaches that combine machine learning models to achieve facial attribute editing across single or multiple domains. Designing distinct transformation algorithms for images in different modes poses a significant challenge when dealing with image conversion tasks. However, Pix2Pix has accomplished image-to-image translation by utilizing pixel-to-pixel mapping based on GAN models, thereby demonstrating greater universality compared with previous image transformation techniques. Nevertheless, training the Pix2Pix model necessitates a substantial amount of paired data, which can be prohibitively expensive to acquire. To address this issue, CycleGAN utilizes a cycle-consistent loss function to achieve image translation between arbitrary pairs of visual domains, employing unsupervised learning and exhibiting extensive applicability. L2MGAN [22], a cross-domain image-processing model consisting of three modules, namely, a style encoder, style transformer, and generator, facilitates style transfer between real and generated images. The functionality of the style encoder lies in extracting style codes, while the style transformer factorizes the d-dimensional latent space into two parts, separating the extracted information into relevant attributes of interest and irrelevant attributes,

crucial for preserving the unrelated aspects since other attributes should not overwrite all unrelated information in the input image. PuppetGAN [23], another cross-domain image processing model, enables style transformation between real and synthetic images. This model employs a domain-uncertainty encoder (E) to map images from both domains into a disentangled latent space, where the interested attributes are isolated from other facial attributes. Furthermore, the model utilizes two distinct decoders, one for the real domain (GA) and the other for the synthetic domain (GB).

2.4. Facial Attribute Manipulation

Facial attribute editing is a wondrous and imaginative technique that bestows the power to alter specific attributes of faces in images or videos. It enables the transformation of facial traces of time, allowing the reversal of years gone by, or the conversion of masculine vigor into feminine grace, even instantly turning individuals into enviable celebrities. The methods employed for facial attribute editing can be categorized into optimization-based approaches (such as CNAI [24] and DFI [25]) and learning-based approaches (such as [26–28]). Optimization-based methods require individual optimization for each test image, incurring significant time costs for model training. Among the learning-based methods, the prevailing approach is the employment of generative adversarial networks (GANs), renowned for their fidelity in various image generation tasks. The early proposed approach, VAE/GAN [29], employed the combination of GAN and VAE, to learn a latent representation, and a decoder. By modifying the latent representation, it aimed to obtain the desired attribute information and subsequently decode it to accomplish the task of attribute editing. Subsequently, AttGAN [8] emerged as a facial editing framework, built upon the foundations of IcGAN [30] and Fader Networks [31]. Unlike the IcGAN and Fader Networks approaches, AttGAN maintains that excessive constraints on latent attributes limit the capacity of implicit representations, leading to overly smooth or distorted image generation. Within this framework, the authors employ attribute classification constraints to ensure accurate attribute variations in generated images. Similarly, StarGAN [9] enhances attribute editing performance through the optimization of necessary adversarial, attribute classification, and reconstruction losses, free from any potential limitations. The performance of STGAN [10] has been further elevated. In terms of attribute embedding, it employs differential attributes, providing more information for attribute transfer. To strike a balance between image generation quality and attribute manipulation, it is necessary to ensure high-quality generated images while precisely controlling attribute manipulation. This presents a conflicting challenge. To address this issue, STGAN draws inspiration from the principles of GRU [32,33] and introduces STU as a means of attribute embedding. It enables the transformation of attributes extracted in the encoder, thereby facilitating more accurate attribute editing. MUGAN [13] presents an alternative model based on a conditional decoder, wherein the target attribute vector is connected to the innermost representation of the encoder and fed into the decoder. In this model, the underlying network of the generator employs a symmetric U-Net structure, where skip connections selectively transfer decoder-side features to the encoder side using additive attention (AUC) to preserve attribute-independent details of the input image, thereby enhancing the quality of the generated images. On the other hand, self-attention complements the convolutional layers by addressing their limitations in capturing long-range dependencies between pixels and considering global geometric information within these layers. The discriminator D consists of two sub-networks: Dadv for discriminating between real and fake images, and Dc for attribute validation within generated images. In this study, we analyze the limitations of STGAN, AttGAN, and MUGAN and further develop a GAN that simultaneously improves attribute manipulation capabilities and image quality.

3. Proposed Method

To address the shortcomings of STGAN and AttGAN in preserving non-edited attributes, the excessive training time of the MUGAN model, and the limited capability for attribute manipulation, we have put forth a novel and enhanced method for facial attribute editing. We employ a symmetrical U-Net architecture to construct the generator, incorporating an additive attention mechanism and a skip connection unit, the AGU, which combines GRU. Furthermore, we incorporate EA (external attention [14]) to complement the information in the convolutional layers. In the discriminator's facial attribute classification model, we propose a discriminative attention mechanism (DAM) that accounts for both spatial and channel dimensions, thereby enhancing the model's attribute classification capabilities. Lastly, we present the objective function of the model.

As shown in Figure 2, the upper section depicts the generator, which bridges the encoder Genc and the decoder Gdec through the AGU. The generator has the ability to selectively transform the encoder representation to complement the decoder representation, thereby enabling editing of the source image based on the given target attribute vector. The lower section illustrates the discriminator D, which takes both the source image and the edited image as inputs. The discriminator D consists of two sub-networks, namely the adversarial discriminator Dadv and the attribute classifier Dc. We apply DAM to the intermediate features of the discriminator.

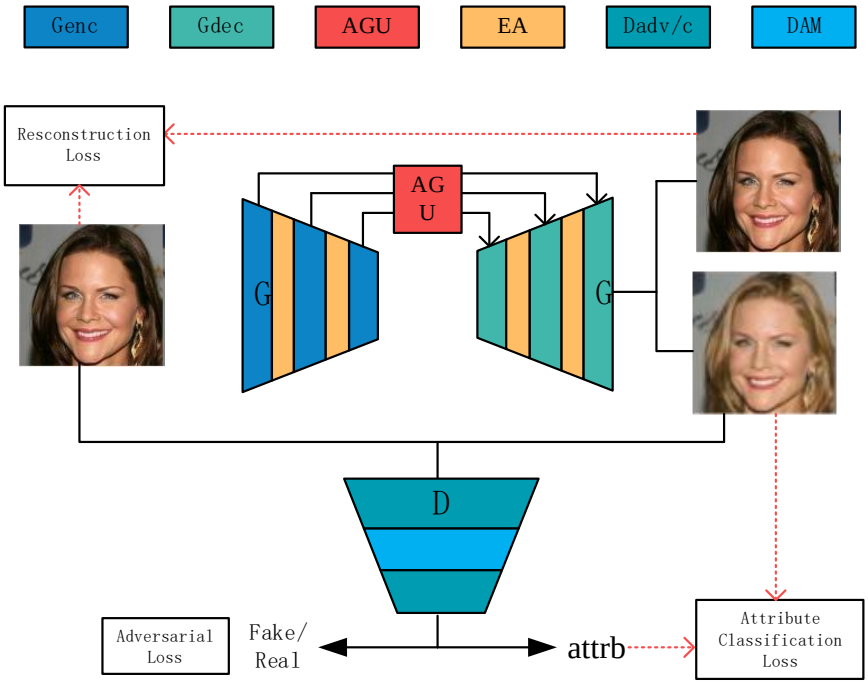

Figure 2. Overview of the MAGAN model.

3.1. Generator

Attention U-Net connection: Figure 3 illustrates the proposed architecture of the generator, wherein we employ the Attention Guided Unit (AGU) to selectively transfer attribute-independent representations between the encoder and decoder. By connecting the encoder and decoder representations through AGU, we aim to enhance image quality and preserve fine details. AGU, as depicted in Figure 4, is a structural unit that combines additive attention and GRU. Similarly to other encoder and decoder attention mechanisms, it facilitates the integration of attribute-related information. The encoder consists of key(k) and value(v) components, while the decoder consists of the query(q) representation.

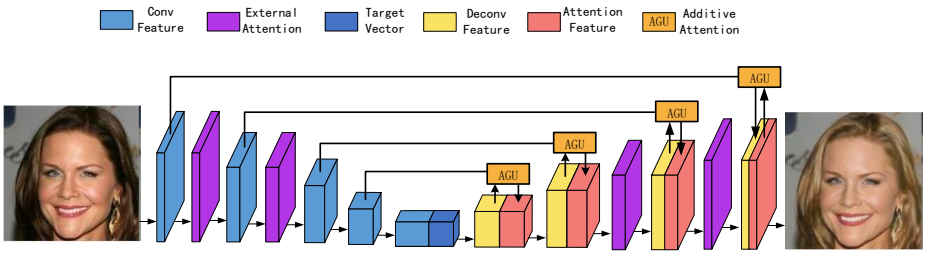

Figure 3. The architectural design of MAGAN's generator. The purple blocks represent the utilization of external attention to augment the feature information, followed by the selective transmission of encoded representations as supplemental information to the decoder through AGU.

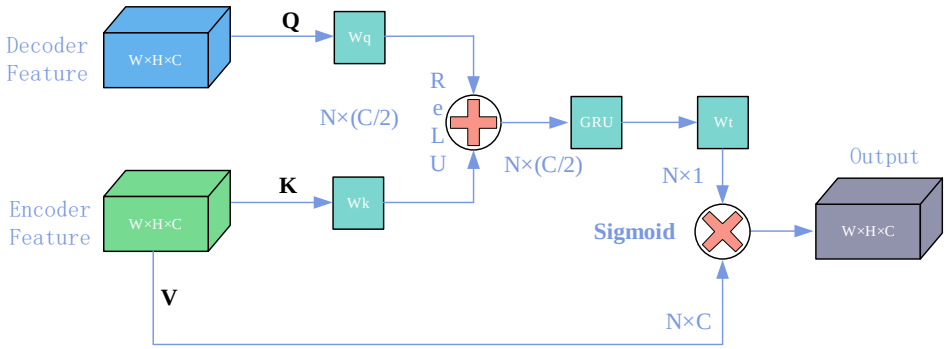

Figure 4. Elaborate details of the structural unit combining additive attention mechanism and GRU. Q is linearly transformed from the decoded features, both K and V are linearly transformed from the encoding features.

Taking the encoder–decoder layer l as an example, we begin by mapping the image information representation E^l / D^l from the preceding encoder/decoder layers through their respective linear transformations, transforming into two feature spaces, q and $K \in R^{C \times N}$, denoted as W_q and W_k. Here, N = W × H represents the format reshaped into a vector (W × H) × (C/2). Let i represent the i-th position in the vector. Linear transformation is achieved through the utilization of a 1 × 1 convolution, whereby the number of channels is reduced by half, precisely to c/2, mirroring the input size.

$$q(D_i^l) = W_q^T D_i^l, k(E_i^l) = W_k^T e_i^l \tag{1}$$

The sum of $q(D_i^l)$ and $k(E_i^l)$ yields the similarity α_i^l, which is further processed through the activation functions ReLU and Sigmoid to compute the attention map α and another transformation block W_t, denoted as:

$$\alpha_i^l = RELU(q(D_i^l) + k(E_i^l)) \tag{2}$$

$$\alpha_i = \frac{1}{(1 + \exp(-W_t^T b_i^l))} \tag{3}$$

Among these, the attention coefficient $\alpha_i \in [0, 1]$ is employed to identify salient regions in the image and trim the representation, allowing only the activations that do not contain attribute-related information to persist. The output of AGU is the element-wise multiplication of the encoder representation E_i^l and attention coefficient α_i, denoted as:

$$\hat{E}^l = \sum_{i=1}^{N} \alpha_i^l E_i^l \tag{4}$$

The AGU merges the encoder representation $\hat{E}^l$ with the decoder representation D_i^l in a cascaded manner, both possessing identical dimensions, forming the input for the subsequent upsampling. The complementary nature of the decoder representation is bolstered through AGU transmission, compensating for the information loss caused by convolutional downsampling and enriching the intricacies of the image.

A facial attribute model based on GANs, constructed with convolutional layers: Due to the limited receptive field of the convolutional kernels, these models can only process information from neighboring pixel regions in the images. As a result, many GAN models built with CNNs face a common challenge of inadequately meeting global geometric constraints. Although MUGAN utilizes a self-attention mechanism as a complement to the convolutional layers in G, effectively modeling dependencies across long-range, spatially disjoint regions (as seen in the second row of Figure 1, where facial hair is preserved during the "to female" transformation, unlike in STGAN), it comes with the drawback of quadratic complexity. This leads to a significant increase in model parameters and training time. To address this, we introduce external attention as a supplement to the generator's convolutional layers. Not only does it reduce training time, but it also achieves the same effects as self-attention. The details of external attention are illustrated in Figure 5.

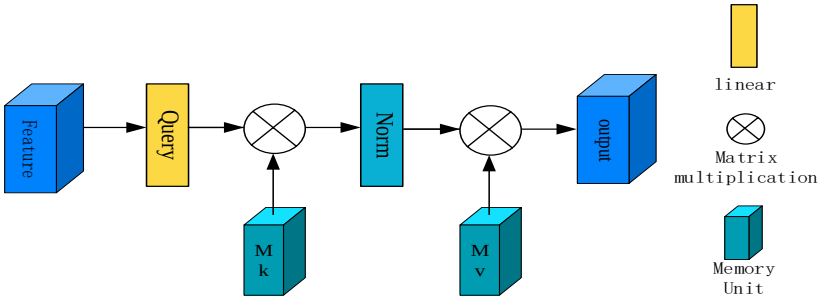

Figure 5. The structure of external attention.

The external attention module employs a concatenation of two cascaded linear layers and normalization layers to achieve its purpose. In essence, it computes the attention between the input and external memory unit, denoted as memory $M \in R^{S \times d}$. By utilizing the feature input extracted by preceding CNN layers as $G \in R^{C \times W \times H}$, the formulation can be expressed as follows:

$$A = \alpha_{i,j} = Norm(GM^T) \tag{5}$$

$$G_{out} = AM \tag{6}$$

In this context, the symbol $\alpha_{(i,j)}$ denotes the similarity between the i-th element and the jth row of matrix M. The memory unit represents a parameter independent of the input, serving as the collective memory for the entire training dataset. Firstly, M is shared, allowing for implicit consideration of the interrelationships among different samples. Secondly, owing to the flexibility of S, we can control its magnitude to make the entire external attention mechanism adaptable, transforming attention into a complexity of N.

3.2. Discriminator

The discriminator D is composed of two main components, as depicted in Figure 6: the adversarial discriminator D_{adv} and the facial attribute classifier D_c. The discriminator takes both real and fake images as input. What sets our discriminator apart from others is the application of our proposed discriminative attention mechanism (DAM) to the intermediate feature layers. This mechanism preserves information in terms of channels and spatial aspects, enhancing the importance of inter-dimensional interactions. By reducing information reduction and amplifying global interactive representations, the performance of the deep neural network is improved, allowing the model to achieve superior attribute-editing capabilities. The main network architecture consists of a feature extractor and two

independent fully connected layers, denoted as D_{adv} and D_c, as depicted in the diagram. The feature extractor, composed of three stacked convolutional layers, serves to extract informative features from the input image. Subsequently, a discriminative attention mechanism is employed to capture crucial features across all three dimensions, yielding the output features. Instance normalization and leaky ReLU activation functions are applied to all convolutional layers. Finally, the convolutional neural network backbone is bifurcated into two branches, each connected to the independent fully connected layers D_{adv} and D_c. The D_{adv} branch discriminates between genuine and counterfeit images, while the D_c branch verifies the presence of the desired facial attributes in the input image, thus imposing constraints on the generated facial attributes.

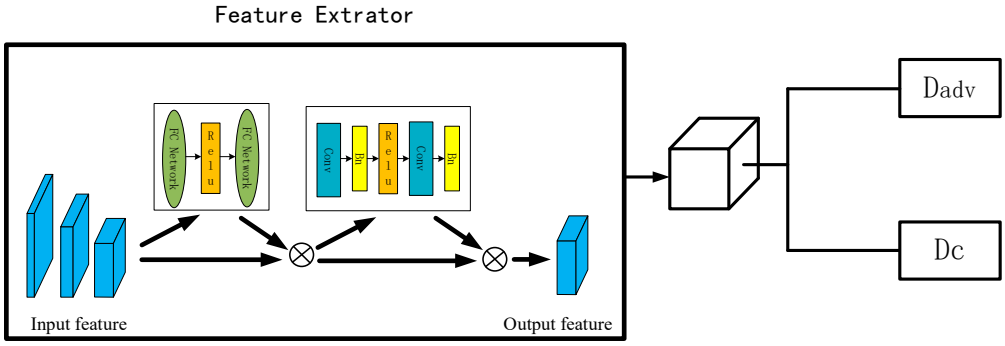

Figure 6. Elaborated details of the discriminator's DAM structure at the intermediate feature layer.

3.3. Loss Functions

The generator of MAgan is composed of two components, G_{enc} and G_{dec}. G_{enc} encodes the input image into a latent representation, while G_{dec} generates images with the desired attributes. Given a facial image x^{a} with n binary attribute labels a, the encoder G_{enc} is used to encode x^{a} into a latent representation. G_{enc} employs three convolutional layers to extract its latent representation, which is defined as:

$$F_e = G_{enc}(x^a) \tag{7}$$

$$F_e = f_e^1 f_e^5 \tag{8}$$

where F_e represents the output of the encoder.

Taking the i-th layer of the encoder–decoder as an example, the outputs extracted by the encoder/decoder at the i-th layer are denoted as f_E^i / f_D^i, while f_{in}^i represents the input representation of the i-th layer in the decoder. Subsequently, we concatenate the innermost encoder representation f_E^5 with the target attribute vector b and pass it to the decoder. Guided by the attribute vector b, we employ AGU to transfer the encoder representations to each decoder layer.

$$\hat{f}_E^{(i-1)} = AGU(f_D^{(i-1)}, f_E^{i-1}) \tag{9}$$

$$\hat{f}_{in}^{(i-1)} = C(f_D^{(i-1)}, \hat{f}_E^{i-1}) \tag{10}$$

$$f_D^i = D({f_{in}}^{i-1}) \tag{11}$$

Among these, C and D respectively denote the channel concatenation function and the deconvolutional layer. $f_E^{(i-1)}$ selectively transfers information from the encoder by incorporating an attention mechanism. This information is then combined with $f_D^{(i-1)}$ to form the input representation, denoted as ${f_{in}}^{i-1}$, which is used for the subsequent

transpose convolutional layer. Finally, x_α is transformed into a new image x_b, with the desired attributes through the use of G_{enc} and G_{dec}.

$$x^{\hat{b}} = G_{dec}(G_{enc}(x^a), b) = G(x^a, b) \tag{12}$$

Aims to minimize the disparity between generated images and real images, bringing their distributions closer to that of real images: The introduction of adversarial learning in our proposed method has significantly enhanced the visual realism of generated images. WGAN [34] employs the Wasserstein distance as a metric to measure the distance between the generated image distribution and the real image distribution, thereby enhancing the quality of image generation. In comparison to conventional GANs, WGAN exhibits greater stability and the ability to generate higher-quality images. Building upon WGAN, we represent the adversarial loss between G and Dadv as follows:

$$\min_{D_{adv}} L_{adv} = -E_{x^a \sim I_r} D(x^a) + E_{x^{\hat{b}} \sim I_f} D(x^{\hat{b}}) \tag{13}$$

$$\min_{\mathrm{G}} \mathrm{L'_{adv}} = -\mathrm{E}_{\mathrm{x}^{\hat{\mathrm{b}}} \sim \mathrm{I_r}} \mathrm{D}(\mathrm{x^a}) + \mathrm{E}_{\mathrm{x}^{\hat{\mathrm{b}}} \sim \mathrm{I_f}} \mathrm{D}(\mathrm{x}^{\hat{\mathrm{b}}}) \tag{14}$$

In the equation, I_r / I_f represents the distributions of real and fake image samples, respectively.

In order to meticulously transform the image $\mathrm{x^a}$ into an image $\mathrm{x^b}$ with the desired facial attribute b, it is imperative to employ the facial attribute classifier Dc to classify the facial attributes. Subsequently, this classification guides the generator G to enforce attribute constraints, thus generating images that possess the accurate facial attributes. The attribute classification loss in this context is defined as follows:

$$\min_{D_c} L_{cls} = E_{x^a \sim I_r}[L(x^a, a)] \tag{15}$$

$$L(x^a, a) = -\sum_{i=1}^{n} \left[a_i \log\left(D_a^i(x^a) \right) \right] + (1 - a_i) \log(1 - D_c^i(x^a)) \tag{16}$$

The value of G is:

$$\min_{\mathrm{D_c}} \mathrm{L'_{cls}} = \mathrm{E}_{\mathrm{x}^{\hat{\mathrm{b}}} \sim \mathrm{I_f}}[\mathrm{L}(\mathrm{x}^{\hat{\mathrm{b}}}, \mathrm{b})] \tag{17}$$

$$L(x^{\hat{b}}, b) = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \left[\mathrm{b_i} \log\left(\mathrm{D_c^i}(\mathrm{x}^{\hat{\mathrm{b}}}) \right) \right] + (1 - \mathrm{b_i}) \log(1 - \mathrm{D_c^i}(\mathrm{x}^{\hat{\mathrm{b}}})) \tag{18}$$

Here, n represents the number of attribute categories and D_c^i denotes the predicted label for the i-th attribute in D_c. L, on the other hand, is the summation of binary cross-entropy losses for all attributes.

By employing the adversarial and classification loss methods, it fails to guarantee that only attribute-relevant regions are altered while preserving the intricacies of the attributes. Hence, G_{dec} needs to learn, under the condition of the original attribute label "a", how to reconstruct the image x^a from the latent representation F_{enc} of G_{enc}. Simultaneously, the introduction of the L1 norm serves as a measure to assess the similarity between the generated image x^a and the original image x^a, thus defining the reconstruction loss.

$$L_{rec} = E_{x_a \sim I_r} ||x^a - G(x^a, a)||_1 \tag{19}$$

The subscript 1 in this context refers to the L_1 norm, which, compared with the L_2 norm, has a greater ability to effectively suppress the blurring effect of an image.

Drawing upon all the aforementioned loss functions, our approach exhibits remarkable performance in both attribute editing and detail preservation. The objective of G can be summarized as follows:

$$\min_{D} L = L_{adv} + \lambda_1 L_{cls} \tag{20}$$

As for G, it represents:

$$\min_{G} L' = L'_{adv} + \lambda_2 L'_{cls} + \lambda_3 L_{rec} \tag{21}$$

where λ_1–λ_3 denote the hyperparameters of the loss function.

4. Experiments

4.1. Dataset and Preprocessing

We employed the CelebA dataset [35] to evaluate the proposed MAGAN. The CelebA dataset possesses numerous instances, wide diversity, and rich annotations, comprising 202,599 facial images from 10,177 distinct identities. Each image is additionally annotated with 40 binary attributes. The 12 attributes including MUGAN, AttGAN, and STGAN, including bald, bangs, black hair, blond hair, brown hair, bushy eyebrows, eyeglasses, gender, open mouth, mustache, pale skin, and age were selected in this experiment. We centered and resized the CelebA source images, originally 178 × 218 in size, to 128 × 128 dimensions. The CelebA dataset was divided into training and testing sets, with the training set consisting of 182,637 images, and the testing set containing 19,962 images. We also used the face dataset in LFW to verify the generalization ability of the model. LFW (Labelled Faces in the Wild) is a public benchmark dataset for face recognition tasks. The dataset contains 13,233 portrait images from different sources on the Internet.

4.2. Implementation Details

The proposed method is compared with AttGAN, STGAN, and MUGAN, all of which were trained and evaluated using multi-attribute models under the same setting. The models involved in the experiment were trained on a workstation equipped with an RTX2080ti GPU. All experiments were conducted in the PyTorch 1.10.1 environment, with CUDA 10.2 and CuDNN 8.2.2. The number of training iterations, or epochs, was set to 200. The model employed the Adam optimizer ($\beta1 = 0.5$, $\beta2 = 0.999$) for optimization, with an initial rate of 0.0002. The learning rate was reduced to 1/10 of its value every 33 epochs. During training, for each generator update, we performed 10 discriminator updates. The weights of the objective function were set as follows: $\lambda_1 = 3$, $\lambda_2 = 10$, $\lambda_3 = 100$.

4.3. Results

We compared the proposed method with two state-of-the-art methods and their variants, (i.e., AttGAN, STGAN, and MUGAN). As depicted in Figure 7, it is shown that the proposed method achieved superior enhancement of geometric constraints when manipulating visually prominent attributes such as bangs, eyeglasses, and mustaches.

Furthermore, as shown in Figure 8, each column represents a different attribute operation, totaling nine items. Each row showcases the qualitative outcomes of the comparative methods, with the source image positioned at the far left of each row. It can be observed that AttGAN, STGAN, and MUGAN achieved reasonable performance in attributes such as eyebrows and pale. However, when it comes to action of "adding bangs", these methods presented blurred and hazy states, failing to effectively incorporate the fringe attribute. STGAN attempts to enhance its performance by modifying the structure of the generator, yet it occasionally lacks clarity when adding the eyeglasses attribute. AttGAN, on the other hand, introduces some artifacts in certain attributes like "bangs". MUGAN struggles to accurately edit the gender attribute, as its result do not distinctly differentiate between genders. In contrast, the proposed method successfully transformed local attributes such as eyeglasses and global attributes like gender and age.

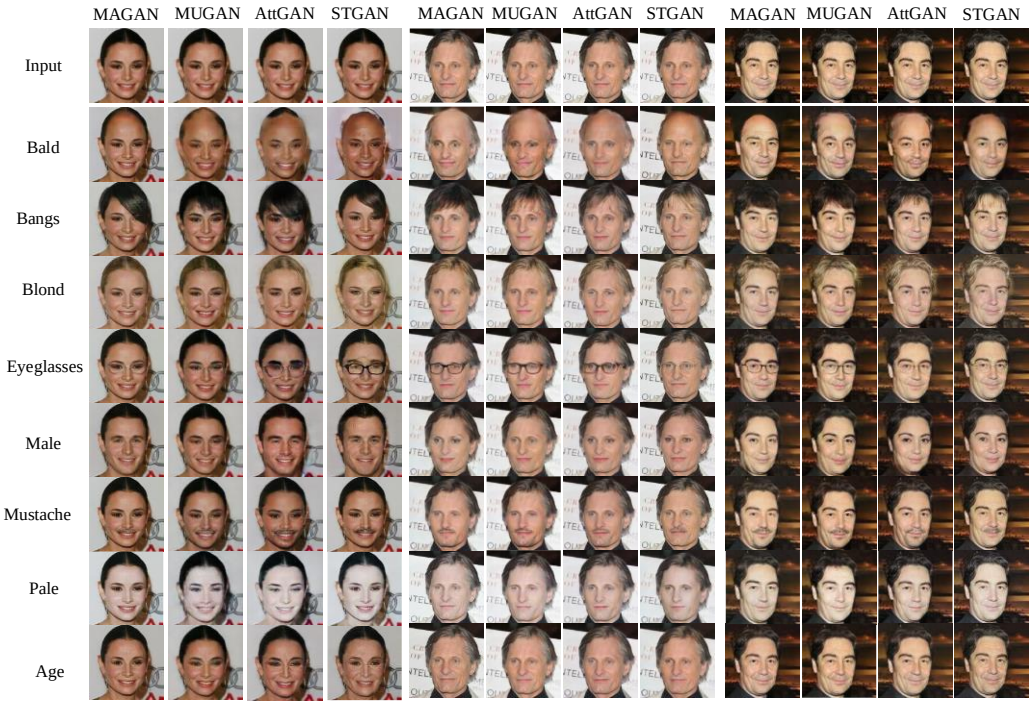

Figure 7. A comparative analysis of attribute-specific editing between AttGAN, STGAN, and MUGAN.

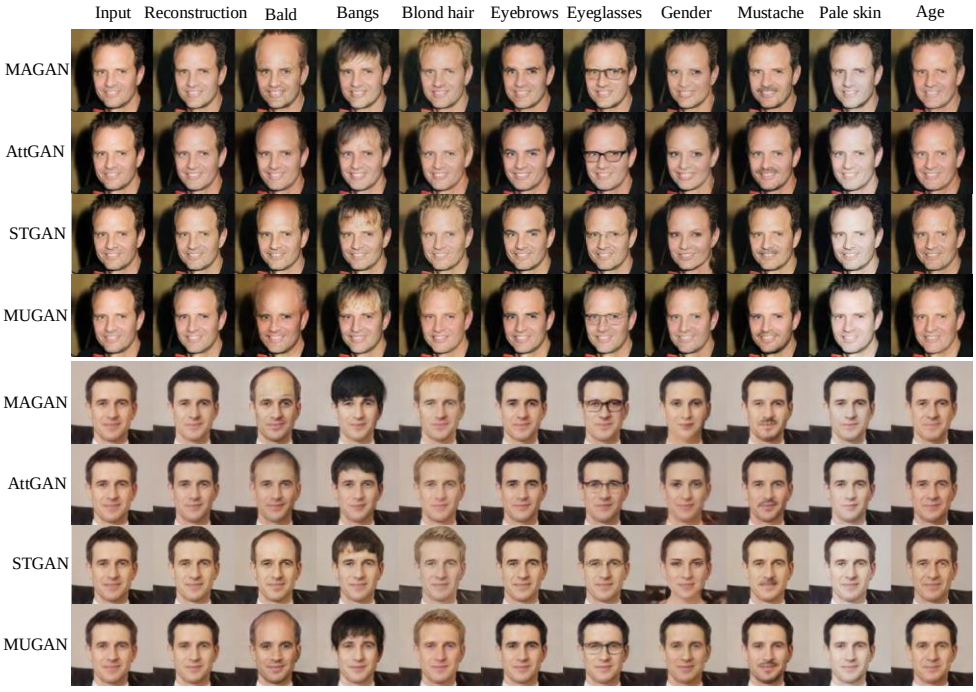

Figure 8. Facial attribute editing results on the CelebA dataset, where the rows from top to bottom correspond to image generation by MAGAN, AttGAN, StarGAN, and MUGAN.

For the attributes “blond” and ”bald”, AttGAN, STGAN, and MUGAN tended to manipulate irrelevant regions, resulting in image blurring. As illustrated in Figure 9, the first column representing the transformation to bald and the last column depicting the hair transformation to blond showed that the proposed model achieved more natural and well-edited images compared with other methods.

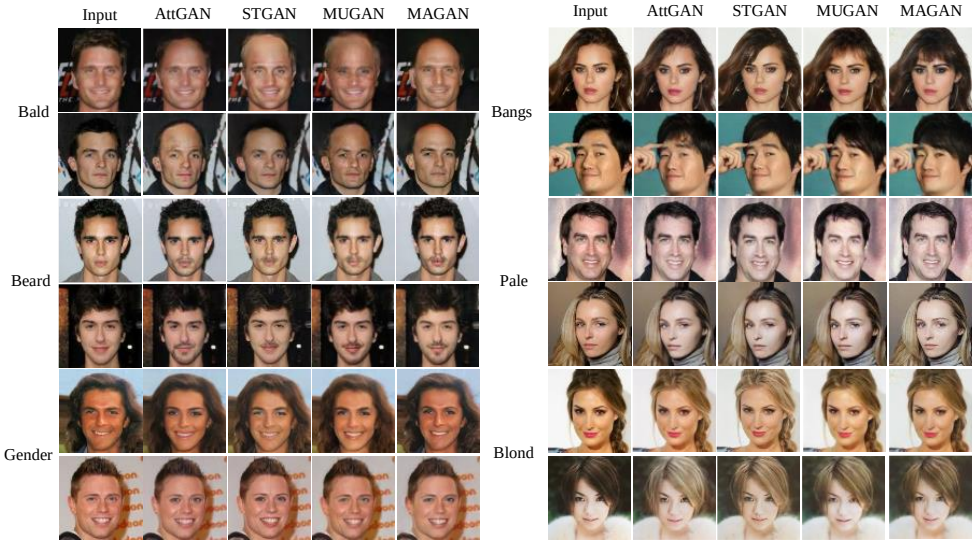

Figure 9. Comparative editing results across various attributes with respect to competing methods.

Finally, in order to help evaluate the robustness, generalization ability, and applicability of the model, and to understand the performance and limitations of the model more comprehensively, we use different datasets to test the model, using the LFW dataset to test our model and other comparative models. As shown in Figure 10, we can see that our face editing effect is still better than that of the other models. In the figure, AttGAN and MUGAN perform poorly in the bald attributes, with artifacts and distortions appearing. AttGAN and STGAN have insignificant editing effects in some attributes, such as bangs and beard. When editing global attributes, such as gender and age, our method generated images of significantly better quality than the others.

Figure 10. In the LFW dataset, the editing results of different attributes using different methods.

4.4. Evaluation

In the task of facial attribute editing, the primary evaluation focuses on the authenticity of the generated images and the accuracy of facial attribute manipulation. To assess the transformation from the source domain to the target domain, we employ the accuracy of attribute operation as a metric. The peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are introduced to evaluate the quality of reconstructed images. Two metrics, FID [36] and image fusion quality AG [37] are introduced to evaluate the quality of image generation. To evaluate attribute operation accuracy, we train a deep attribute classification model using STGAN on the training set of the CelebA dataset. The trained model achieves an accuracy of 94.7% when tested on the 12 attributes of the test set. It is employed to measure the classification accuracy of the edited attributes, as depicted in Figure 11. It can be observed that the attribute classification accuracy (except the "male") surpasses that of other models. Notably, there is a significant improvement in the generation accuracy for attributes such as bangs, black hair, blond hair, and mustache.

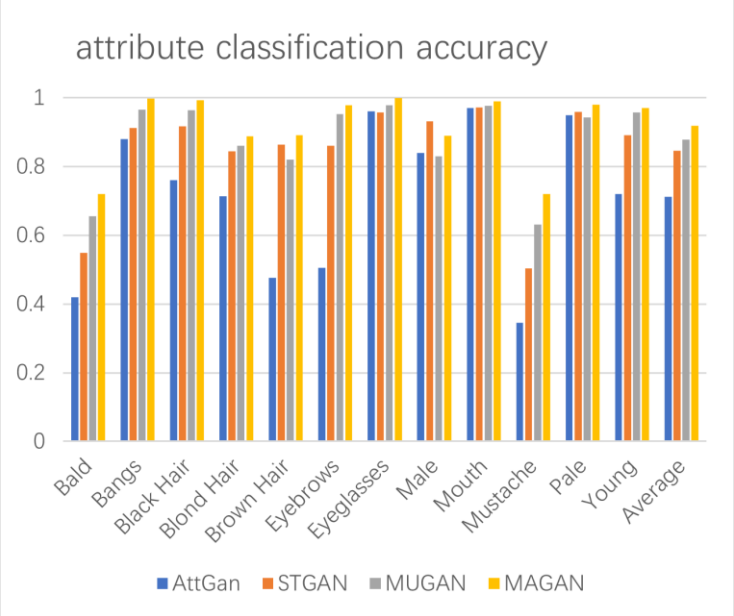

Figure 11. AttGAN, STGAN, MUGAN, and MAGAN's precision in attribute generation, encompassing a total of 12 facial attributes.

As shown in Table 1, the average accuracy of attribute classification stands at 91.83%, achieving a substantial improvement compared with AttGAN's 71.16%, STGAN's 84.67%, and MUGAN's 87.78%. This signifies a remarkable enhancement in the model's ability to manipulate attributes upon the incorporation of AGU and the discriminative attention mechanism.

Table 1. Comparative analysis showcasing the average accuracy for 12 facial attributes and the results of reconstruction quality.

Method	Average Accuracy	PSNR/SSIM
AttGAN	71.16%	20.65/0.801
STGAN	84.67%	30.67/0.927
MUGAN	87.78%	31.58/0.934
MAGAN	91.83%	32.52/0.957

For evaluating the reconstruction results, we employ peak signal-to-noise ratio and structural similarity (PSNR/SSIM) as evaluation metrics. PSNR is a metric that assesses the quality of reconstructed images by calculating the mean squared error between the original and reconstructed images, expressed in a logarithmic scale. A higher value indicates a smaller disparity between the reconstructed and original images, indicating higher image quality. On the other hand, SSIM comprehensively considers the structural and pixel-level similarity of images, yielding a result between 0 and 1. A value closer to 1 indicates a higher similarity between the reconstructed and original images, indicating better image quality. As shown in Table 1, we observed that the generator, incorporating external attention, preserved more image information. The proposed model achieved higher quality image reconstruction compared with STGAN, AttGAN, and MUGAN. Table 2 shows the average accuracy and reconstruction quality results of 12 facial attributes generated by each method in the LFW dataset. It can be seen that in the LFW dataset that both the average accuracy of facial attributes and the reconstruction quality results are better than the others. The performance of the CelebA dataset is poor. This is due to the quality of the dataset. Most of the faces in LFW pictures are not frontal, and the picture pixels are blurred. However, even so, our method still has the best effect compared with other methods.

Table 2. Comparative analysis showcasing the average accuracy for 12 facial attributes and the results of reconstruction quality in the LFW dataset.

Method	Average Accuracy	PSNR/SSIM
AttGAN	70.59%	19.35/0.755
STGAN	82.84%	29.95/0.903
MUGAN	85.98%	30.37/0.916
MAGAN	89.32%	31.28/0.933

In terms of visual quality and preventing undesired variations, we employed the Fréchet Inception Distance (FID), which is a metric to assess the quality of images generated by Generative Adversarial Networks (GANs). It is used to measure the discrepancy between generated and original images, aiding in the refinement of GAN models and producing generated images that closely resemble the original images. Initially, we divided the CelebA test set into two subsets, Test set A and Test set B. Test set A serves as the source images for attribute editing, while Test set B contains a set of distinct source images. Subsequently, we compared the two image groups (test set A0, derived from test set A's synthetic image set, and test set B, the original image set) to mitigate the impact of input-output similarity when measuring FID scores. Note that we made some modifications to each image (e.g., male to female/female to male), thereby making the evaluation based on FID scores meaningful. The results are depicted in Figure 12. The FID metric assesses the similarity between two sets of images based on the statistical similarity of their computer vision features. A lower score indicates a greater similarity in statistical properties between the image sets. The optimal score is 0.0, indicating complete identity between the two. It can be observed that, except for the attributes of bald, bangs, and mustache, the proposed model achieved lower FID scores compared with the other models.

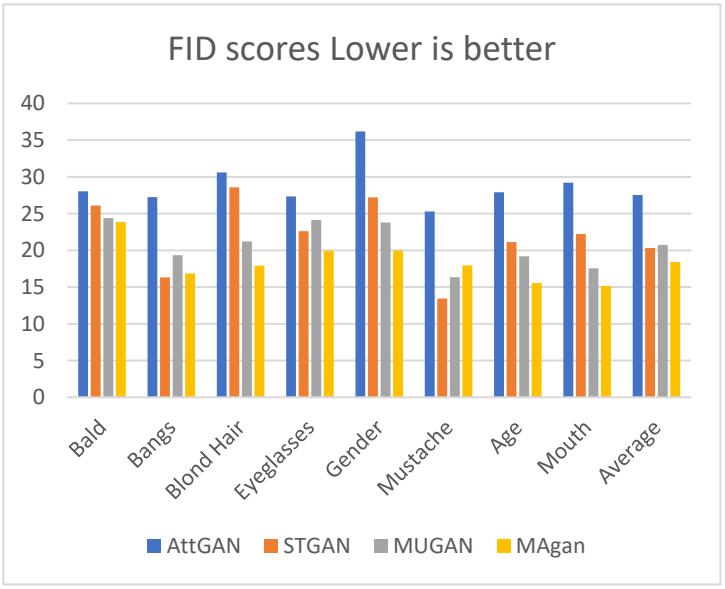

Figure 12. FID scores depicting the image quality of attribute generation for AttGAN, STGAN, MUGAN, and MAGAN.

The FID metric quantifies the generated images by measuring the distance between feature vectors of the generated images and original images. However, when it comes to facial attribute editing, there are limitations in using the FID metric for evaluation since there is no reference to original edited images. The AG metric, on the other hand, is employed to assess the clarity of fused images based on the average gradient value. A higher average gradient indicates higher image clarity and better fusion quality. Evaluating 12 attributes, as depicted in Figure 13, it shows that the AG scores of the proposed model

are higher than those of other models. This indicates that the proposed model is capable of generating higher-quality edited images.

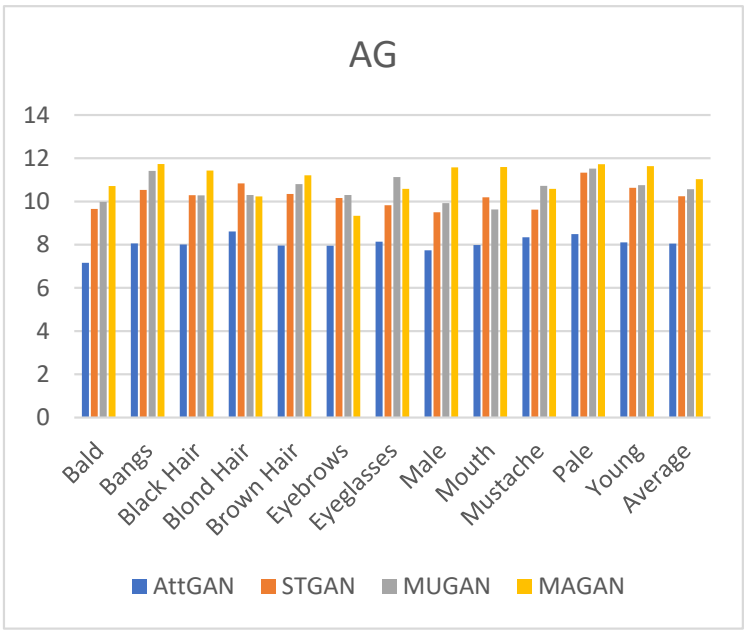

Figure 13. We evaluated the AG metrics for twelve attributes, including AttGAN, STGAN, MUGAN, and MAGAN.

4.5. Ablation Study

To analyze the effects of these modules, we constructed several variants of MAGAN under the similar experimental settings but with following updates: (1) Change0–based on our proposed model; (2) Change1–excluding the external attention mechanism while retaining the AGU module and DAM; (3) Change2–removing the discriminative attention mechanism (DAM) while keeping the others intact; and (4) Change3–eliminating the AGU module while retaining the other two components. These variants were trained and tested on the same CelebA dataset, and the same evaluation metrics were used. Figure 14 shows the attribute classification accuracy of different variants of MAGAN.

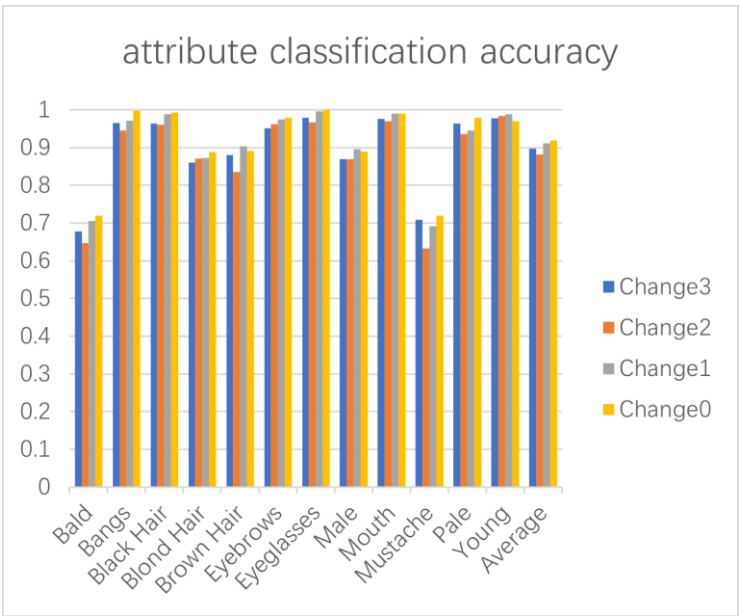

Figure 14. Accuracy of attribute generation in the variant of MAGAN.

4.5.1. Effect of AGU Structure

Firstly, the editing results between our complete model, Change0, and the model without AGU, Change3, are depicted in Figure 15. Comparing the generated results of Change0 and Change3, the images generated by Change3 exhibit a greater degree of blurriness, particularly in the bald attribute. One possible reason for this phenomenon is that the model fails to capture the contextual relationships among facial features during the generation process, resulting in a confusion between skin tone and background images, thus causing image blurring. The inclusion of AGU enables a focus on crucial facial features, distinguishing between the face and the background, thereby achieving a more natural appearance in the generated faces.

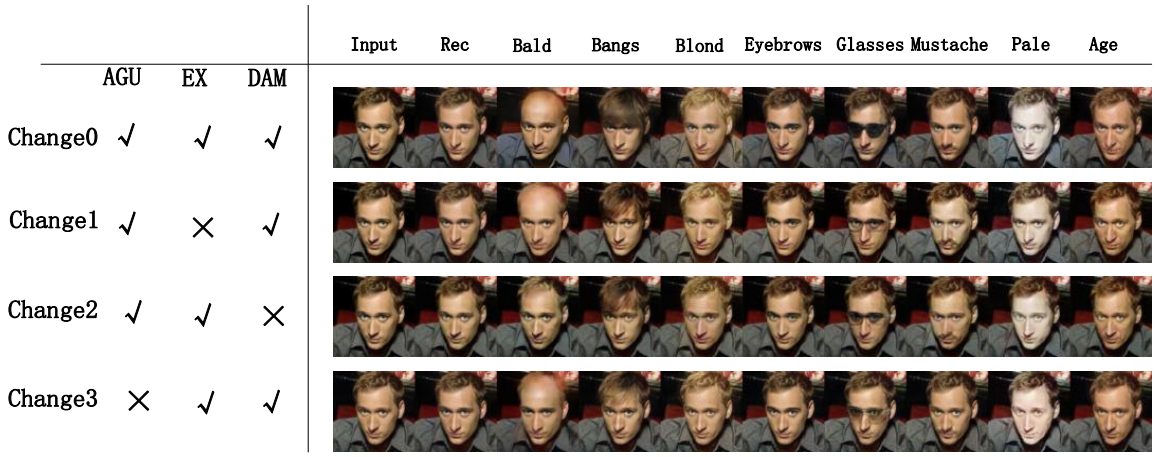

Figure 15. The generated outcomes for various combinations of the model.

4.5.2. Effect of an External Attention Mechanism

As shown in Figures 15 and 16 and Table 3, the inclusion of the external attention mechanism enhances the attribute classification accuracy. Figure 17 and Table 4 showed that the generator, equipped with external attention, achieved a superior perceptual quality in reconstructing images, as evidenced by the improved PSNR and SSIM values. It indicates that FID scores are consistently higher when the external attention mechanism is absent, suggesting a slight degradation in the quality of generated images. Moreover, the average AG results for Change0 surpass those of Change1 AG, underscoring that the introduction of external attention leads to improved clarity and fusion quality. The incorporation of external attention into the generator not only reduces model parameters but also elevates the overall image generation quality of the model.

Table 3. The average accuracy of twelve facial attributes in the MAGAN variants.

Method	Average Accuracy
Change3	89.78%
Change2	88.17%
Change1	91.05%
Change0	91.83%

Table 4. Mean attribute generation and image reconstruction results for different variants.

Method	Average AG	PSNR/SSIM
Change3	10.84	31.74/0.946
Change2	10.68	31.58/0.934
Change1	10.89	31.942/0.951
Change0	11.03	32.52/0.957

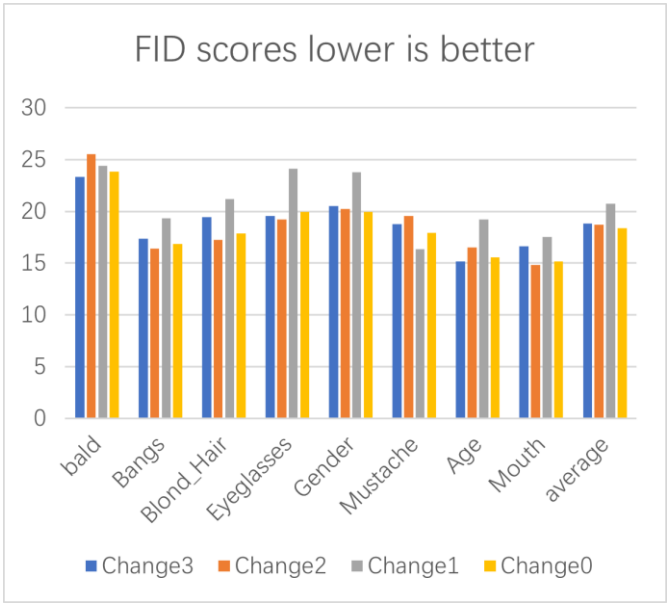

Figure 16. FID score results for images generated by different variants.

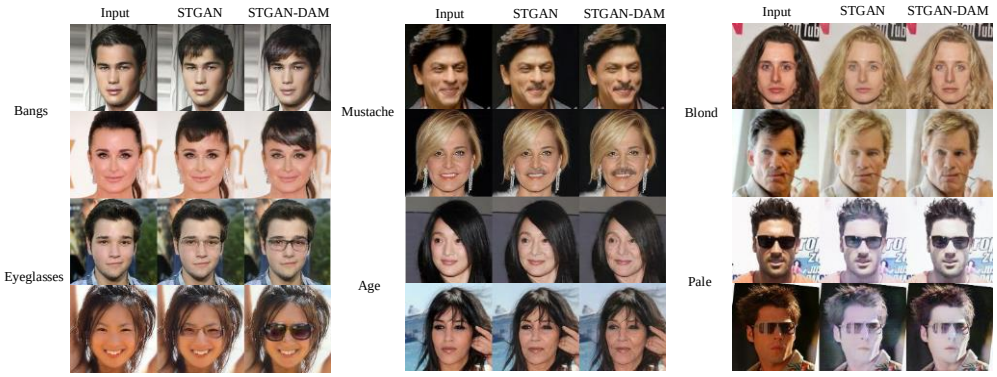

Figure 17. A comparison of STGAN and STGAN+DAM in terms of the mustache, age, blond, and pale attributes.

4.5.3. Effect of DAM

As depicted in Figures 14 and 16, the proposed model suffered from a noticeable decline in its facial editing attribute capability, particularly in the bald and eyeglasses attributes, without DAM. PSNR/SSIM in Table 4 indicate that the inclusion of DAM yielded improvements in the model's reconstruction performance. The average AG score increases after integrating DAM (as shown in Figure 16). Change0 achieved a lower score than Change2, suggesting that the incorporation of DAM strengthens the discriminator, elevating the model's demand for image authenticity. Furthermore, to demonstrate the generalizability of our proposed attention mechanism, we have applied it to AttGAN and STGAN, as shown in Figures 17 and 18. When AttGAN and STGAN incorporate our proposed attention mechanism, their attribute editing capabilities are significantly enhanced. By comparing the visibly perceptible attribute changes, it is evident that the models exhibit stronger attribute editing capabilities. As depicted in Figure 19, the DAM enhances the model's constraint capability in attribute classification. Hence, our proposed discriminator attention mechanism has the potential to improve the model's attribute editing capability.

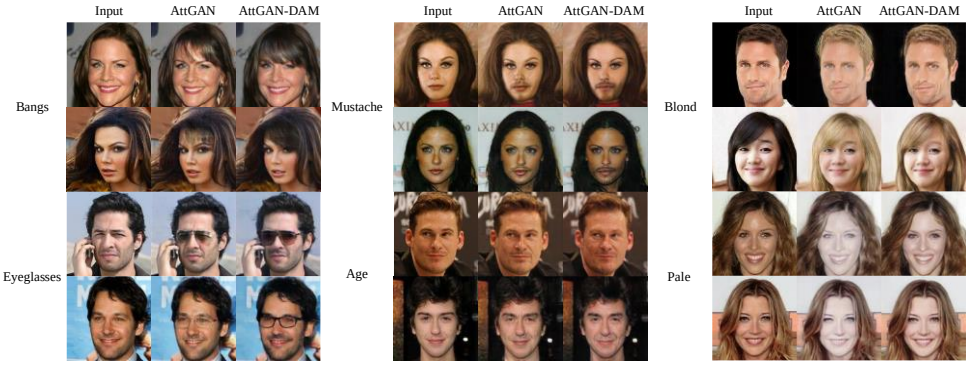

Figure 18. A comparison between AttGAN and AttGAN + DAM regarding the attributes of mustache, age, blond, and pale.

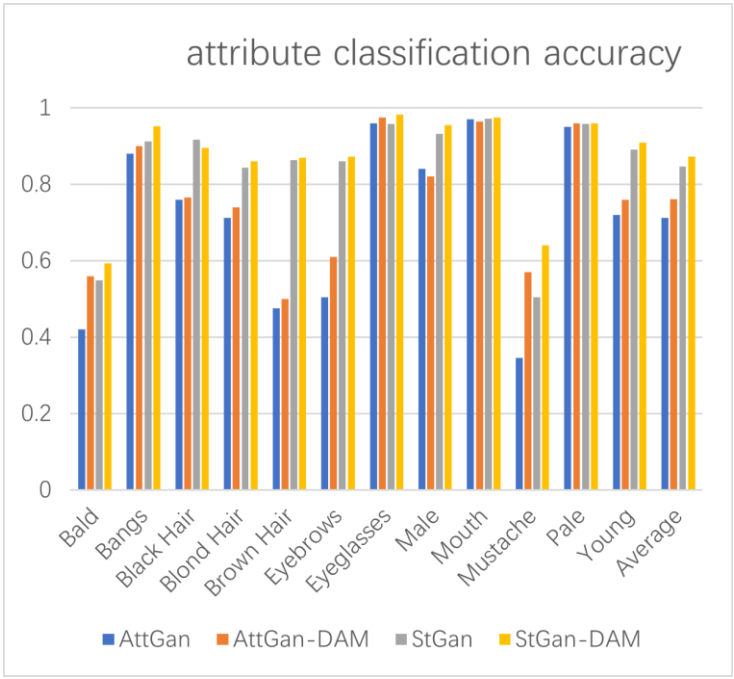

Figure 19. The attribute generation accuracy of AttGAN, STGAN, AttGAN-DAM, and STGAN-DAM.

5. Conclusions

In this paper, we propose the integration of GRU and an additive attention mechanism within the generator of a facial attribute editing model, giving rise to a formidable skip-connection unit named AGU. Through experimental validation, this fusion model has showcased remarkable performance. The uniqueness of AGU lies in its ability to harness both the sequence modeling prowess of GRU and the crucial weight allocation characteristics of the additive attention mechanism. GRU effectively captures long-term dependencies in input sequences through its gating mechanism, making it adept at handling temporal data in the generator. On the other hand, the additive attention mechanism allows AGU to focus its attention on the most relevant parts of the input sequence, providing the generator with more accurate contextual information. Experimental results demonstrate a significant performance improvement by AGU in facial attribute editing tasks. Compared with traditional generators, AGU-generated images exhibit enhanced detail, clarity, and naturalness, ensuring facial integrity and consistency. This substantiates the effectiveness and practicality of AGU as a skip-connection unit in facial-attribute editing tasks. In addition, a discerning attention mechanism has been proposed to optimize the intermediate feature layers of the discriminator, enhancing its capacity to capture crucial features across three dimensions. This augmentation improves the discriminator's ability to locate spatial

regions and specifically identify areas associated with specified attributes. Consequently, it enables effective control over facial edits, restricting them to regions relevant to the designated target attribute. This enhancement elevates the facial attribute editing capabilities of the model while enhancing the performance of the discriminator, resulting in the generation of more authentic images. We have employed external attention on the generator, addressing the issues of excessive model parameters and prolonged training time, while also enhancing the quality of generated images. Through qualitative and quantitative analyses, we have compared the performance of STGAN, AttGAN, and MUGAN. The proposed method demonstrates superior performance in most target attributes compared with these three methods, achieving significant improvements in certain attributes. In terms of model performance, we use two datasets, CelebA and LFW, for comparison. It can be seen that our method performs well in different datasets, indicating that our model has high robustness and no overfitting, convergence, or underfitting.

Although our method has achieved good results in face editing, it has certain limitations. First, when the background color is similar to the skin color, when adding bangs, although the generated bangs have obvious features, the they are not realistic, as shown in Figure 8. This may be because the model does not have strong enough context awareness. Due to limitations of hardware resources, we currently only choose to add external attention in layers 2 and 4. Theoretically, it would be better if external attention were introduced to all layers. Our next work is to add different layers of external attention, and increase the number, to analyze and compare the effect on the model. In the future, as hardware and software resources become increasingly abundant, we will also consider using more complex and effective attention mechanisms to improve the face editing model. Finally, the method we proposed uses a variety of attention mechanisms to supplement convolution information. Although the external attention mechanism has greatly reduced the complexity of the model, the number of parameters of the entire model is still too large and the training time is too long. In the future, we will consider using some new convolutional modules to reduce redundant information between features in convolutional neural networks, thereby compressing the number of model parameters and improving its performance.

Author Contributions: Conceptualization, Z.L.; Methodology, W.X.; Formal analysis, X.M.; Investigation, W.X.; Resources, Z.L.; Writing—review & editing, H.X.; Visualization, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of Zhejiang Province (No. LY21F020005).

Data Availability Statement: We use the face public dataset CelebA. The data are not publicly available due to our laboratory's policy or confidentiality agreement. We have fully described the experimental design, analysis, and results, as well as the procedures for data analysis and processing. If editors and reviewers have questions about specific data, we will endeavor to provide more detailed explanations and clarifications.

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

References

1. Liu, X.; Wang, R.; Peng, H.; Yin, M.; Chen, C.F.; Li, X. Face beautification: Beyond makeup transfer. *Front. Comput. Sci.* **2022**, *4*, 910233. [CrossRef]
2. Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; Alahi, A. Social gan: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 2255–2264.
3. 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]
4. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *Stat* **2014**, *1050*, 1.
5. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
6. Kim, M.; Liu, F.; Jain, A.; Liu, X. DCFace: Synthetic Face Generation with Dual Condition Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12715–12725.

7. Huang, Z.; Chan, K.; Jiang, Y.; Liu, Z. Collaborative Diffusion for Multi-Modal Face Generation and Editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6080–6090.
8. He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.* **2019**, *28*, 5464–5478. [CrossRef] [PubMed]
9. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 8789–8797.
10. Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; Wen, S. Stgan: A unified selective transfer network for arbitrary image attribute editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3673–3682.
11. 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.
12. Almahairi, A.; Rajeshwar, S.; Sordoni, A.; Bachman, P.; Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 195–204.
13. Zhang, K.; Su, Y.; Guo, X.; Qi, L.; Zhao, Z. MU-GAN: Facial attribute editing based on multi-attention mechanism. *IEEE/CAA J. Autom. Sin.* **2020**, *8*, 1614–1626. [CrossRef]
14. Guo, M.H.; Liu, Z.N.; Mu, T.J.; Hu, S.M. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5436–5447. [CrossRef] [PubMed]
15. Hinton, G.E.; Zemel, R. Autoencoders, minimum description length and Helmholtz free energy. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 30 November–3 December 1992.
16. Nie, W.; Wang, Z.; Patel, A.B.; Baraniuk, R.G. An improved semi-supervised VAE for learning disentangled representations. *arXiv* **2020**, arXiv:2006.07460.
17. Huang, H.; He, R.; Sun, Z.; Tan, T. Introvae: Introspective variational autoencoders for photographic image synthesis. In Proceedings of the Advances in Neural Information Processing Systems 31, Montreal, QC, Canada, 3–8 December 2018.
18. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
19. 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, 15–21 June 2019; pp. 4401–4410.
20. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 852–863.
21. Salehi, P.; Chalechale, A. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In Proceedings of the 2020 IEEE International Conference on Machine Vision and Image Processing (MVIP), Qom, Iran, 18–20 February 2020; pp. 1–7.
22. Yang, G.; Fei, N.; Ding, M.; Liu, G.; Lu, Z.; Xiang, T. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2951–2960.
23. Usman, B.; Dufour, N.; Saenko, K.; Bregler, C. Puppetgan: Cross-domain image manipulation by demonstration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9450–9458.
24. Li, M.; Zuo, W.; Zhang, D. Convolutional network for attribute-driven and identity-preserving human face generation. *arXiv* **2016**, arXiv:1608.06434.
25. Upchurch, P.; Gardner, J.; Pleiss, G.; Pless, R.; Snavely, N.; Bala, K.; Weinberger, K. Deep feature interpolation for image content changes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7064–7073.
26. Li, M.; Zuo, W.; Zhang, D. Deep identity-aware transfer of facial attributes. *arXiv* **2016**, arXiv:1610.05586.
27. Shen, W.; Liu, R. Learning residual images for face attribute manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4030–4038.
28. Zhou, S.; Xiao, T.; Yang, Y.; Feng, D.; He, Q.; He, W. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv* **2017**, arXiv:1705.04932.
29. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1558–1566.
30. Perarnau, G.; Van De Weijer, J.; Raducanu, B.; Álvarez, J.M. Invertible conditional gans for image editing. *arXiv* **2016**, arXiv:1611.06355.
31. Lample, G.; Zeghidour, N.; Usunier, N.; Bordes, A.; Denoyer, L.; Ranzato, M.A. Fader networks: Manipulating images by sliding attributes. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.

32. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
33. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
34. Adler, J.; Lunz, S. Banach wasserstein gan. In Proceedings of the Advances in Neural Information Processing Systems 31, Montreal, QC, Canada, 3–8 December 2018.
35. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
36. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in neural information processing systems 30, Long Beach, CA, USA, 4–9 December 2017.
37. Schmidt, M.; Le Roux, N.; Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Program.* **2017**, *162*, 83–112. [CrossRef]