Article

Semi-Supervised FaceGAN for Face-Age Progression and Regression with Synthesized Paired Images

Quang T. M. Pham ¹ , Janghoon Yang ² and Jitae Shin ^{1,*}

¹ Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, Korea; quangpham@skku.edu

² Department of New Media, Seoul Media Institute of Technology, Seoul 07590, Korea; jhyang@smit.ac.kr

* Correspondence: jtshin@skku.edu

Received: 16 March 2020; Accepted: 30 March 2020; Published: 1 April 2020

Abstract: The performance of existing face age progression or regression methods is often limited by the lack of sufficient data to train the model. To deal with this problem, we introduce a novel framework that exploits synthesized images to improve the performance. A conditional generative adversarial network (GAN) is first developed to generate facial images with targeted ages. The semi-supervised GAN, called SS-FaceGAN, is proposed. This approach considers synthesized images with a target age and the face images from the real data so that age and identity features can be explicitly utilized in the objective function of the network. We analyze the performance of our method over previous studies qualitatively and quantitatively. The experimental results show that the SS-FaceGAN model can produce realistic human faces in terms of both identity preservation and age preservation with the quantitative results of a decent face detection rate of 97% and similarity score of 0.30 on average.

Keywords: deep learning; generative adversarial networks; face aging; age progression; age regression; synthesized images

1. Introduction

Age progression and age regression have appealed to the research community for a long time. Age regression is the attempt to represent the preceding face of a person. On the contrary, age progression or face aging is the task of presenting the synthesized future-look of a face given an input face (see Figure 1). Both tasks are important tasks because of their numerous applications [1,2]. One can use this technique to find missing persons or to alter the faces of actors virtually according to the age of the character in a movie. The two main factors in this research are identity preservation and age preservation. That is, the input and output faces need to look like they come from the same person, and the generated face should be in accordance with the target age group.

Figure 1. Face aging task.

The previous methods require paired face images [3,4]. However, it is difficult to collect such labeled datasets. There are many public face datasets with age labels for each image. However, most

of them do not include serial face images of the same person at different timelines (e.g., UTKFace [2], IMDB-WIKI [5,6], APPA-REAL [7]). Because of this, previous methods usually consider face aging as an unpaired image-to-image translation problem [2,8,9]. The dataset should be split into several age groups (called domains). The goal now is to translate images from a source domain to the target domain.

Many researchers have tried to apply generative adversarial networks (GANs) to face age translation [9–11]. This technique is known as a powerful tool for generating realistic images [12,13]. Basically, they exploited conditional GANs to transform the given face to the synthesized face at the target age. In this task, it is ideal for learning the aging transformation features using samples from the same person at many different times. However, most of the public datasets do not have such paired images. The limitations of such datasets cause several problems. For example, the model may not learn the aging transformation features in enough detail. Different people have different age features (e.g., facial structure, beard). Moreover, there is a trade-off between identity preservation and aging transformation learning. If a model cannot determine the aging features and the identity features properly, it may produce low-quality images in terms of age accuracy while trying to maintain the person's identity and vice versa. Although there are many studies related to face aging, not many attempts have been made regarding face regression [2].

To solve these issues, we propose a framework with two generative adversarial networks (GANs). The first GAN is an age-conditional GAN that focuses on learning aging features. The other GAN learns both aging features and identity features from a real dataset and the output of the first GAN. After training, the first GAN generates the synthesized data, which is exploited for training the second GAN. The synthesized data from the first GAN includes paired images (same person at different ages). Therefore, this approach helps overcome the limitations of real datasets, which do not have paired data.

The proposed model is distinct from the existing works [2,10] in the following way. While previous studies do not provide architecture that can be trained by real images and synthesized images efficiently, the proposed model provides a structure that exploits real images and synthesized images to improve performance. In addition, the proposed model has a totally different network architecture from the baseline model [2], where Unet is used as a generator and a pair of synthesized images can improve training through implicit supervision.

Overall, the contributions of this paper are as follows:

1. We proposed a novel framework for age progression and regression including two GAN models. By using an additional GAN, we can train the model with a semi-supervised approach with synthesized paired images, which avoids the limitations of real datasets.
2. We introduced a new way of training that separates the aging features and identity features so that we can better train our model. With our proposed method, we can use a Unet-based model as a generator, which can overcome the bottleneck limitation of auto-encoder. This helps our model to produce more detailed images.

2. Related Works

Before deep learning, the two main approaches for age progression were physical model approaches and prototype approaches. The physical model approaches focused on the change of physical factors (e.g., hair, wrinkles, mouth) over time [14,15]. Those approaches were complicated and required a large amount of paired data. The prototype approaches learned the features by averaging the faces of people in the same age group [3,16,17]. The aging features can be presented differently for each group. However, this method results in smoothed face images that lose information regarding identity.

Nowadays, deep convolution networks are widely used in image processing. Many studies have tried to apply GANs to face aging since it often generates realistic images. The basic GAN includes two networks, the generator and discriminator, to learn and mimic a real data distribution [18]. During training, the discriminator tries to distinguish whether the input image is from real data or the output

of the generator, while the generator tries to confuse the discriminator by generating images that make the discriminator predict real images. In the last several years, a lot of improvements have been made to keep the training process more stable and generate more high-quality images [12,19,20].

On the other hand, one variant of GAN, called conditional GAN [21], is more applicable than the original GAN. Instead of generating images from random noise, conditional GAN generates an image from a given label [22] or image [23,24]. The conditional GAN is good at image-to-image translation tasks such as style transfer [25], super-resolution [26], and colorization [27]. Therefore, many studies exploited conditional GAN to synthesize aged faces. For example, acGAN is an age-conditional GAN model for age progression [8]. For better quality, S. Liu et al. [11] introduced an extra module to take advantage of cross-age transition patterns while a pyramid-structured discriminator is used for learning both global and local features in [28]. Moreover, we can improve performance by exploiting additional information/loss functions such as facial attributions [29], perceptual loss, age classification loss [10], bias loss [30], and regularization on latent space [2]. Other studies tried to apply reinforcement learning to face aging [31] or explain why GANs are good at face aging tasks [32]. To overcome the limitations of real datasets, reconstruction loss and identity loss are applied in References [2,33] while the CycleGAN-based model [24] is used in References [30,34]. However, those techniques are not good enough to deal with a wide range of ages, where we have to handle significant changes between images. Consequently, age regression is not handled well because it requires the ability to learn global facial changes.

3. Method

3.1. Baseline Method

The baseline network is taken from the Conditional Adversarial Autoencoder (CAAE) model [2] (see Figure 2). The CAAE model includes one generator and two discriminators. The generator is an auto-encoder. The encoder part extracts features and produces an encoded z given the input image I_1 . The decoder takes the encoded z and the target age information t_2 , then generates the output image I_2 corresponding to the age group t_2 . The main discriminator D_{img} distinguishes the real/fake images based on the output of the generator and the age group information t_2 . The additional discriminator D_z forces the encoded z 's distribution as a prior distribution (e.g., uniform distribution). Besides the common loss functions of GAN, the objective function also includes the loss function of the discriminator D_z , the reconstruction loss L_2 norm between input and output image (for identity preserving) and the total variation loss (for removing artifact ghosting).

Figure 2. The baseline model based on Conditional Adversarial Autoencoder (CAAE).

Although the CAAE model can capture aging features over time, the image quality is not good. The CAAE model generates the output images based on the encoded vector z . However, the encoder

encodes a high-dimensional input image into a low-dimensional vector z for extracting high-level facial features [2]. Therefore, vector z cannot capture all the information from the input images because of dimensionality reduction [35]. The reconstruction loss used in auto-encoder also cannot deal with multi-modal data [36]. Therefore, the output of CAAE is blurry. In addition, the reconstruction loss they used also has other issues. Let us denote E and G as the encoder and decoder of the generator, respectively. In the CAAE framework, the authors use the L_2 norm for reconstruction loss.

$$\min_{E,G} L_2(I_1, G(E(I_1), t_2)). \quad (1)$$

In Equation (1), the reconstruction loss between the output and input image for keeping the identity could be harmful to learning aging features, since the model should generate the output image at the target age group. However, the input image comes from a different age group. Therefore, forcing the model to generate the output image close to the input image actually makes the output face appear as though it comes from the source age group, not the target age group. This conflict also results in blurry output images.

3.2. Proposed Model

To solve the problem of image quality degradation in the output, we apply the Unet architecture [37] to replace the auto-encoder. The Unet architecture includes skip-connections, which may help to generate more detailed output images. Moreover, skip-connections also improve the gradient flow so the model can learn better than in the case of an auto-encoder.

However, once we apply the Unet architecture to the generator, the CAAE model cannot learn effectively due to reconstruction loss. During the training phase, the reconstruction loss overwhelms the conditional adversarial loss. The Unet model allows the layers close to an output layer to receive information from the primary layers of the encoder. Therefore, to minimize the reconstruction loss, the Unet generator can simply emphasize the input information on the output and ignore the high-level information from intermediate layers. In that case, the generator cannot learn the age features and the reconstruction loss can reduce to 0 quickly. As a result, the output images look the same as the input images (to keep the L_2 norm low) while the generative loss is high (see Figure 6). On the other hand, if we reduce the impact of L_2 loss, the output will become unrealistic.

The reason for using this kind of reconstruction loss in the CAAE model is due to the limitations of real datasets. To deal with this problem, we propose a way to generate the labeled data, including sequence images, from the same person at different ages. Thus, instead of using the input image for comparison, we actually compare the output image with the synthesized image. This semi-supervised learning method helps to overcome not only the original problem of reconstruction loss, but also the problem of trade-off between identity preservation and aging translation during training. Besides identity preservation, the reconstruction loss now also helps to learn aging features. We use an additional GANs model for synthesizing the labeled dataset.

Overall, our proposed model combines two GANs. The first GAN is the main model with the Unet architecture as the generator, which we named FaceGAN. FaceGAN takes the face image as input and produces the face image of the same person in accordance with a target age. The second GAN generates face images from random vectors. We named this network conditional StyleGAN (cStyleGAN). The cStyleGAN is used for making the synthesized labeled dataset. We use the second GAN as a supervisor of our main GAN model. The whole framework is called semi-supervised FaceGAN (SS-FaceGAN) and described in Figure 3.

Figure 3. The proposed SS-FaceGAN network.

3.3. The Conditional StyleGAN

The cStyleGAN is used for making the synthesized dataset. It receives a random vector z and age information t as input and produces a face image with corresponding age group t . We can think of t as the conditional age information, which controls the aging features of the output image. The random vector z now only encodes the identity features. By fixing z and changing t , we are able to generate multiple face images that have the same identity features and different aging features. This helps the cStyleGAN learn the aging features and identity features separately without requiring sequence samples from the same person at different ages.

For implementation, we decide to use the StyleGAN [13]. The StyleGAN model is the state-of-the-art model for face generation, and can generate realistic images at high resolutions. The input of StyleGAN is fed directly to every layer via adaptive instance normalization operations [38] so that features in every layer can be generated directly from the input. We want to first train StyleGAN to generate face images in different age groups. To do this, we modify the original StyleGAN model to conditional StyleGAN by adding the age information as conditional information.

We assume the image x and its label t come from a real data distribution P_{data} and the random vector z is sampled from a prior distribution P_z . The objective function of cStyleGAN can be described as:

$$\min_{G_{style}} \max_{D_{style}} \mathbb{E}_{x,t \sim P_{data}, z \sim P_z} [\log(1 - D_{style}(G_{style}(z, t))) + \log(D_{style}(x, t))]. \quad (2)$$

3.4. The FaceGAN

Once we get the trained cStyleGAN model, we generate synthesized face images for the same person in different age groups by simply changing the age input for fixed z . Then, we use these images as input for training the FaceGAN. During this phase, we only train FaceGAN while freezing cStyleGAN.

For training FaceGAN, we use both the real dataset and the synthesized dataset. For the real dataset, we also apply the conventional objective function for GAN:

$$L_{G_{real}} = \mathbb{E}_{x,t \sim P_{data}} [\log(1 - D_{face}(G_{face}(x, t), t))] \quad (3)$$

$$L_{D_{real}} = \mathbb{E}_{x,t \sim P_{data}} [\log(D_{face}(G_{face}(x, t), t)) + \log(1 - D_{face}(x, t))], \quad (4)$$

where x is the real face image at age t in Equations (3) and (4). To train with synthesized data, we first generate them using cStyleGAN. Given a random vector z (where $z \sim P_z$) and any two random different age labels t_1 and t_2 ($t_1, t_2 \sim P_t$, P_t is a prior distribution), we use cStyleGAN to produce two synthesized face images i_1 and i_2 that look like they are from the same person but with a different age:

$$i_1 = G_{style}(z, t_1); i_2 = G_{style}(z, t_2). \quad (5)$$

After that, the generator of FaceGAN is trained with the objective function as below:

$$L_{G_{fake}} = \mathbb{E}_{z \sim P_z; t_1, t_2 \sim P_t} [\log(1 - D_{face}(G_{face}(i_1, t_2), t_2))]. \quad (6)$$

Since i_1 and i_2 in Equation (5) are not the real images, we ignore the following term for training the discriminator:

$$\min_{D_{face}} \mathbb{E}_{z \sim P_z; t_1, t_2 \sim P_t} [\log(1 - D_{face}(i_2, t_2)) + \log(D_{face}(G_{face}(i_1, t_2), t_2))]. \quad (7)$$

The reconstruction loss of our method is the L_1 norm between the output image and the synthesized image i_2 . We notice that using L_1 instead of L_2 leads to a better quality output. We train the generator of FaceGAN with the reconstruction loss in Equation (8).

$$L_{L1} = \mathbb{E}_{z \sim P_z; t_1, t_2 \sim P_t} [L_1(G_{face}(i_1, t_2), i_2)]. \quad (8)$$

We also apply total variation loss (denoted as $TV(\cdot)$ in Equation (9)) for reducing artifacts when training the FaceGAN generator with real data:

$$L_{TV} = \mathbb{E}_{x, t \sim P_{data}} [TV(G_{face}(x, t))]. \quad (9)$$

Overall, the total loss of the generator is:

$$L_G = \alpha_1 L_{G_{real}} + \alpha_2 L_{G_{fake}} + \alpha_3 L_{L1} + \alpha_4 L_{TV}. \quad (10)$$

In Equation (10), α_1 , α_2 , α_3 , and α_4 are the weights for each loss. We define the loss function discriminator as:

$$L_D = \gamma L_{D_{real}}, \quad (11)$$

where γ is the weight for $L_{D_{real}}$ in Equation (11).

Our new reconstruction loss shows several advantages. First, it can deal with the lack of paired samples in real datasets. The reconstruction loss in the CAAE framework gives a negative impact on capturing the aging features. On the contrary, the proposed model can also learn the aging features by comparing output faces with synthesized faces. The proposed method enables us to apply Unet as the generator of FaceGAN, which captures local features better than the auto-encoder. As a result, the output can be less blurry.

4. Experiments

4.1. Dataset

We use the UTKFace dataset [2] for training. The UTKFace dataset consists of over 20,000 face images along with information about age (from 0 to 116 years old), gender, and ethnicity. Those face images are aligned and cropped so that this dataset is suitable for many tasks including face aging. For a fair comparison, we only use age information for training and ignore gender and ethnicity labels. Differently from the CAAE experiment, we change the number of age groups from 10 to 6: 0–10, 11–20, 21–30, 31–40, 41–50, and 50+. For evaluation, we use The Face and Gesture Recognition Research Network (FG-NET) aging database, which is usually used for facial aging studies [39]. The FG-NET

includes 1002 images from 82 subjects (from 0 to 69 years old). The images from FG-NET are also aligned and cropped for fair evaluation.

4.2. Implementation Details

For cStyleGAN, we train it with images of which the maximum size is 128×128 . The input of the generator is random noise and age information, while the discriminator takes image and age information as the input. The random noise is a 500-dimensional vector and the age information is a one-hot vector with 6 dimensions (corresponding to the 6 age groups). We repeat the age information two times and concatenate it to the random noise to make a 512-dimensional vector to be the input of cStyleGAN. For faster training, we reduce the number of channels in both the generator and discriminator. A more detailed description of cStyleGAN is shown in Appendix A.

For the generator of FaceGAN, we use the auto-encoder and add skip-connections to build up the Unet model. The input of the generator of FaceGAN is the face image and the age information. The resolution of the input image is 128×128 . The age information is first encoded as a one-hot vector (6 dimensions). Then, it is resized to a tensor ($128 \times 128 \times 6$) where the entries corresponding to the target channel are filled with ones while the other channels are filled with zeros. The face image and that tensor are concatenated before feeding to the generator. The input of the cStyleGAN discriminator is also the concatenation of face image and age tensor. For the FaceGAN discriminator, the age information is fed to an upsampling block (using the transposed convolution layer) before concatenating the main flow. The network architecture of FaceGAN is described in Figure 4. Both cStyleGAN and FaceGAN are trained on the UTKFace dataset.

Figure 4. The architecture of the FaceGAN generator and discriminator where k is the kernel size, n is the number of channels and s is the stride for each convolutional layer.

For training FaceGAN, we set the learning rate to 0.0002 for both the generator and discriminator. We use the Adam optimizer [40] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The number of epochs is 120. We set the weight coefficients $\alpha_1 = 0.2, \alpha_2 = 0.2, \alpha_3 = 500, \alpha_4 = 1, \gamma = 1$.

For comparison, we implemented the CAAE model [2] and Identity-Preserved Conditional Generative Adversarial Networks (IPCGANs) model [10]. In the CAAE model, we modify the auto-encoder such that its inputs are image and age information only for a fair comparison. For training IPCGANs, we use the UTKFace dataset to first train the age classification module and then train the main model. Both compared models are trained with 128×128 images and 6 age groups. Each method is trained to make a single model for both age progression and age regression. In the inference phase, the inputs are arbitrary human face images without any additional information (age, gender, etc.).

4.3. The Qualitative Results

First of all, we train the cStyleGAN model on the UTKFace dataset. The synthesized outputs of cStyleGAN are shown in Figure 5. Those images are generated by the random vector z and the age information t . As we can see, by fixing the random noise z and changing the conditional age t , the model can present aging features well while maintaining the other facial features for identity preservation. This shows that the model is able to learn those two types of features separately.

Figure 5. The results from cstyleGAN (time flows from top to bottom rows).

To verify the domination problem of reconstruction loss, which can overwhelm the adversarial loss, we conduct an experiment using the baseline CAAE model with Unet as the generator. We check the performance of the new model (where Unet is a generator) with the original objective function of CAAE. As we expected, in the training phase, the reconstruction loss decreased to 0 quickly and the model could not learn anything. It can be seen from Figure 6 that if we replace the auto-encoder generator with Unet, all the output images look the same as the input ones. This indicates that our hypothesis in Section 3.2 is true.

The results of the three methods are shown in Figure 7. It is easy to see that the result images of CAAE are blurry and the effect of aging features is not obvious. Since the reconstruction loss of CAAE cannot make sufficient differences in the input and output, it fails to learn the global changes. Meanwhile, our method can capture aging features well, for example, the shape of heads or the presence of beards from the young group to the old group. Additionally, our method provides images with more details than CAAE.

Figure 6. The results based on CAAE with and without the Unet generator. With the impact of reconstruction loss, the output images look exactly the same as the input images.

Figure 7. The results with three methods (in each box, from top to bottom: Identity-Preserved Conditional Generative Adversarial Networks (IPCGANs), CAAE, SS-FaceGAN and the real images). For every person, the real images (the top-left image and the images in 4th row) and the synthesized images for 6 age groups are shown.

Compared with IPCGANs, the SS-FaceGAN method can also learn global changes better. Our method results in more realistic synthesized young faces than IPCGANs. In the case of the old group, IPCGANs learn the aging features directly from the trained age classifier. However, it only presents the local aging features (e.g., wrinkles). It sometimes results in unrealistic faces in a wide range of translational cases. For example: when predicting old faces of babyfaces, there are several dark areas and the shape of the head still remains the same.

In terms of age regression, SS-FaceGAN shows the best performance. The aging change is not evident in the case of CAAE. The synthesized young face images of IPCGANs come with artifacts, such as abnormal eyes and mouths. Moreover, the IPCGANs could not remove beards and glasses completely in young face images. In contrast, our method is able to generate realistic baby face images and show clear aging features.

Comparisons of the synthesized images with real images show that SS-FaceGAN outperforms the existing methods in expressing aging dependent features while maintaining the identity. For the input image of age 2, IPCGANs fail to generate normal human faces while CAAE also falls short of expressing aging dependent features. SS-FaceGAN is observed to express a slim face well. The synthesized image generated by IPCGANs and CAAE is found to be blurred significantly for the input image of age 31, possibly due to the low brightness. SS-FaceGAN is shown to generate a synthesized image with better quality. It is noted that the synthesized face images of SS-FaceGAN do not have glasses on the face for the young age image while they are present for the old age image, which implies that SS-FaceGAN might be able to learn when people wear glasses or not, which can depend on age. Similarly, the synthesized images of young faces generated by SS-FaceGAN do not have mustaches while other methods may make such considerations. SS-FaceGAN is also shown to generate synthesized images with aging features most accurately for the input images of ages 54 and 69.

4.4. The Quantitative Results

To make data for evaluation, we use 1002 images from FG-NET, which includes all 6 age groups. For every image, we generate 6 synthesized images corresponding to the 6 age groups. Thus, we have 6012 generated images for each method. This is different from the way that the IPCGANs method generates their evaluation dataset. In the IPCGANs method, the authors use only real images from the young group (11–20 years old) to assess the age progression task, while we wish to examine the performance for both age progression and age regression. However, due to missing data, we can only compare 3440 synthesized images to ground truth images.

We examine how realistic the face images generated are by a face detection model, which allows identifying human faces in images more objectively free from human bias. The Multi-task Cascaded Convolutional Networks (MTCNN) framework [41], which detects faces using confidence scores, is exploited. The number of detected faces and associated detection rates are as follows: for a total of 6012 images, the number of detected face images of CAAE, IPCGANs and our method are 5895 (98%), 5227 (87%) and 5834 (97%), respectively. The proposed method outperforms IPCGANs, which often fails to generate realistic faces, while it performs as well as CAAE, which generates face images with minor changes from the original image.

The quality of the detected face can be compared using the confidence score, which is the probability that the detected region in an image is a human face. We ignore all the samples generated by any of the three methods that cannot be detected by MTCNN. The confidence score is the output of the face classifier model, which indicates the probability of detecting a human face in an image. Figure 8 shows the average confidence score for each age group, which is calculated over the face images detected by all three methods for a fair comparison. The average confidence scores of SS-FaceGAN are larger than those of IPCGANs for all age groups. They are even slightly larger than those of CAAE, which generates the face images with minor local changes from the original face image. Even though CAAE quantitatively provides a larger confidence score for the groups of 0–10 and 10–20 years old, the generated image with the target age falls short of belonging to those age groups, as shown in Figure 7.

The images generated by CAAE do not have obvious changes compared to the input images. As a result, the output is still realistic, but cannot interpret the correct age features.

Figure 8. The performance on face recognition.

Secondly, we compare the identity preserving ability and the age transformation ability. The output image should preserve the identity of the face despite the age translation, while the age features should be presented appropriately according to the desired age group. In order to do this, we use a Resnet-50 model trained on VGGFace2 [42]. This is the pre-trained model on a large-scale face dataset. We exploit the pre-trained model as the feature extraction and compute the cosine similarity of the vector features of the synthesized output images and the real images. This means that for each subject in the FG-NET dataset (as described in Figure 7), we compare the synthesized images in each column to the ground truth images of the same column in the 4th row. We verify that a small cosine similarity implies the images represent the same person, as opposed to a large cosine similarity. Therefore, by obtaining the similarity scores between the output images and the real images, we can evaluate the ability of a model to maintain the identity information and learn the aging features correctly for the synthesized outputs. The lower similarity scores mean the model preserves the identity information and performs the aging features better.

The similarity scores are computed for each age group. The results in Table 1 show that our method achieves the best scores for all groups. Our method can accurately reflect the aging factor (the appearance of glasses and beard growth in Figure 7) while maintaining the identity information. The poor similarity score of CAAE also implies that this method may not result in significant changes between the input and output. Therefore, the CAAE method is not good at learning aging factors.

Table 1. The cosine similarity scores with different methods.

	IPCGANs	CAAE	SS-FaceGAN
0–10 years old	0.33	0.34	0.32
11–20	0.35	0.37	0.32
21–30	0.32	0.37	0.29
31–40	0.33	0.37	0.28
41–50	0.36	0.36	0.28
51+	0.34	0.35	0.29

5. Conclusions

In this paper, we introduced a novel semi-supervised learning method for age progression and regression. To make a synthesized paired dataset, a conditional styleGAN is trained to learn the aging and identity features first. The additional styleGAN helps to overcome the limitation of

datasets. Then, we use this module to train FaceGAN to generate synthesized face images from input images. This leads to better learning for both aging features and identity features. We also exploit the Unet architecture for the generator of FaceGAN to improve the output quality. Both qualitative and quantitative experiments prove that our method is better than the other considered methods, especially regarding the age regression task. The confidence score shows that our model can produce realistic human faces. Our model also performs well at maintaining the identity information and presenting aging factors with a low cosine similarity score.

However, there is still a lot of room for improvement. For the age progression task, our model does not perform well when learning local aging features, such as skin irregularities or wrinkles, in the eldest group. Although the synthesized data from cStyleGAN can improve the results of the FaceGAN model, we still cannot solve the multi-modal problem with reconstruction loss [36]. In the future, we may add an age classification module [10] to help the model capture more aging features at a high level. To deal with the multi-modal problem, we may apply a multi-modal framework [43] or change the objective function [44]. Moreover, we can exploit other information of a person, such as gender or ethnicity, as conditional information in our model to improve the performance. We can also improve the way to use the synthesized face data for training the main model. An articulated definition of loss function using the synthesized data and an advanced architecture can be another direction to improve the performance of the proposed method. The performance of existing networks for face-age progression/regression is often limited by a lack of sufficient data over a wide range of ages. Thus, a method of collecting proper data in a new way to overcome the limitation of the existing research needs to be paid attention to.

Author Contributions: Conceptualization, Q.T.M.P. and J.S.; methodology, Q.T.M.P. and J.Y.; formal analysis, Q.T.M.P., J.Y. and J.S.; software, validation, writing—original draft preparation, Q.T.M.P.; writing—review and editing, Q.T.M.P., J.Y. and J.S.; supervision, project administration, and funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03031752) and was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2018-0-01798) supervised by the IITP (Institute for Information & communications Technology Promotion).

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

Appendix A. The cStyleGAN Architecture

The cStyleGAN architecture is described in Figure A1 and Tables A1 and A2. The cStyleGAN follows the original architecture of StyleGAN [13], but we reduce the number of channels for several layers. The parameters of the generator and discriminator are presented in Tables A1 and A2. Note that the input vector of the generator is fed into the mapping network, while the output of the mapping network is the style vector. Then, the main generator (in Table A1) runs the style vector through Adaptive Instance Normalization Layers [38]. For the discriminator, the image and age information are concatenated (the same setting as the input for the generator of FaceGAN) before feeding it into the network.

Figure A1. The StyledConvBlock with L output channels in the generator. In each block, the parameters of the convolution layer are: k is the kernel size, n is the number of channels and s is the stride for each convolutional layer.

Table A1. The cStyleGAN generator.

Layer	Kernel Size; Padding	Output Shape
StyledConvBlock with ConstantInput	$3 \times 3; 1$	$512 \times 8 \times 8$
UpSample	-	$512 \times 16 \times 16$
StyledConvBlock	$3 \times 3; 1$	$512 \times 16 \times 16$
UpSample	-	$512 \times 32 \times 32$
StyledConvBlock	$3 \times 3; 1$	$512 \times 32 \times 32$
UpSample	-	$512 \times 64 \times 64$
StyledConvBlock	$3 \times 3; 1$	$512 \times 64 \times 64$
UpSample	-	$512 \times 128 \times 128$
StyledConvBlock	$3 \times 3; 1$	$512 \times 128 \times 128$

Table A2. The cStyleGAN discriminator.

Layer	Kernel Size; Padding	Activation	Output Shape
Input Image	-	-	$9 \times 128 \times 128$
Conv (from_rgb)	$1 \times 1; 1$	-	$128 \times 128 \times 128$
Conv	$3 \times 3; 1$	LeakyReLU	$128 \times 128 \times 128$
Conv	$3 \times 3; 1$	LeakyReLU	$128 \times 128 \times 128$
DownSample	-	-	$128 \times 64 \times 64$
Conv	$3 \times 3; 1$	LeakyReLU	$256 \times 64 \times 64$
Conv	$3 \times 3; 1$	LeakyReLU	$256 \times 64 \times 64$
DownSample	-	-	$256 \times 32 \times 32$
Conv	$3 \times 3; 1$	LeakyReLU	$512 \times 32 \times 32$
Conv	$3 \times 3; 1$	LeakyReLU	$512 \times 32 \times 32$
DownSample	-	-	$512 \times 16 \times 16$
Conv	$3 \times 3; 1$	LeakyReLU	$512 \times 16 \times 16$
Conv	$3 \times 3; 1$	LeakyReLU	$512 \times 16 \times 16$
DownSample	-	-	$512 \times 8 \times 8$
Minibatch stddev	-	-	$513 \times 8 \times 8$
Conv	$3 \times 3; 1$	LeakyReLU	$512 \times 8 \times 8$
Conv	$8 \times 8; 0$	LeakyReLU	$512 \times 1 \times 1$
Fully-Connected	-	linear	$1 \times 1 \times 1$

For training, the number of images for each phase is 600k. We use the Adam optimizer, the learning rate is 0.0015, $\beta_1 = 0.0$ and $\beta_2 = 0.99$ for training cStyleGAN. The loss function is WGAN-GP loss [45]. The source code for cStyleGAN is available at <https://github.com/QuangBK/cStyleGAN>.

References

1. Fu, Y.; Guo, G.; Huang, T.S. Age Synthesis and Estimation via Faces: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. doi:10.1109/TPAMI.2010.36. [[CrossRef](#)] [[PubMed](#)]
2. Zhang, Z.; Song, Y.; Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
3. Shu, X.; Tang, J.; Lai, H.; Liu, L.; Yan, S. Personalized Age Progression with Aging Dictionary. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 3970–3978. doi:10.1109/ICCV.2015.452. [[CrossRef](#)]
4. Wang, W.; Cui, Z.; Yan, Y.; Feng, J.; Yan, S.; Shu, X.; Sebe, N. Recurrent Face Aging. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2378–2386. doi:10.1109/CVPR.2016.261. [[CrossRef](#)]
5. Rothe, R.; Timofte, R.; Gool, L.V. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.* **2016**. [[CrossRef](#)]
6. Rothe, R.; Timofte, R.; Gool, L.V. DEX: Deep EXpectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Santiago, Chile, 7–13 December 2015.
7. Agustsson, E.; Timofte, R.; Escalera, S.; Baró, X.; Guyon, I.; Rothe, R. Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. In Proceedings of the 12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017.
8. Antipov, G.; Baccouche, M.; Dugelay, J. Face aging with conditional generative adversarial networks. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2089–2093. doi:10.1109/ICIP.2017.8296650. [[CrossRef](#)]
9. Yang, H.; Huang, D.; Wang, Y.; Jain, A.K. Learning Face Age Progression: A Pyramid Architecture of GANs. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 31–39. doi:10.1109/CVPR.2018.00011. [[CrossRef](#)]
10. Wang, Z.; Tang, X.; Luo, W.; Gao, S. Face Aging with Identity-Preserved Conditional Generative Adversarial Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
11. Liu, S.; Sun, Y.; Zhu, D.; Bao, R.; Wang, W.; Shu, X.; Yan, S. Face Aging with Contextual Generative Adversarial Nets. In Proceedings of the ACM Multimedia, Mountain View, CA, USA, 23–27 October 2017.
12. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
13. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
14. Suo, J.; Zhu, S.; Shan, S.; Chen, X. A Compositional and Dynamic Model for Face Aging. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 385–401. doi:10.1109/TPAMI.2009.39. [[CrossRef](#)] [[PubMed](#)]
15. Shen, C.T.; Huang, F.; Lu, W.H.; Shih, S.W.; Liao, H.Y. 3D Age Progression Prediction in Children's Faces with a Small Exemplar-Image Set. *J. Inf. Sci. Eng.* **2014**, *30*, 1131–1148.
16. Rowland, D.A.; Perrett, D.I. Manipulating facial appearance through shape and color. *IEEE Comput. Graph. Appl.* **1995**, *15*, 70–76. doi:10.1109/38.403830. [[CrossRef](#)]
17. Kemelmacher-Shlizerman, I.; Suwajanakorn, S.; Seitz, S.M. Illumination-Aware Age Progression. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3334–3341. doi:10.1109/CVPR.2014.426. [[CrossRef](#)]
18. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Mandi, India, 16–19 December 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 2672–2680.

19. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
20. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
21. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
22. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. In Proceedings of the 34th International Conference on Machine Learning, ICML'17, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 2642–2651.
23. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
24. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
25. Chen, X.; Xu, C.; Yang, X.; Song, L.; Tao, D. Gated-GAN: Adversarial Gated Networks for Multi-Collection Style Transfer. *IEEE Trans. Image Process.* **2019**, *28*, 546–560. [[CrossRef](#)] [[PubMed](#)]
26. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114. doi:10.1109/CVPR.2017.19. [[CrossRef](#)]
27. Nazeri, K.; Ng, E.; Ebrahimi, M. Image Colorization Using Generative Adversarial Networks. In *International Conference on Articulated Motion and Deformable Objects*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 85–94.
28. Yang, H.; Huang, D.; Wang, Y.; Jain, A.K. Learning Continuous Face Age Progression: A Pyramid of GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, doi:10.1109/TPAMI.2019.2930985. [[CrossRef](#)] [[PubMed](#)]
29. Liu, Y.; Li, Q.; Sun, Z. Attribute Enhanced Face Aging with Wavelet-based Generative Adversarial Networks. *arXiv* **2018**, arXiv:1809.06647.
30. Wang, C.; Liu, H.; Pei, S.; Liu, K.; Liu, T. Face Aging on Realistic Photos by Generative Adversarial Networks. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Hokkaido, Japan, 26–29 May 2019; pp. 1–5. doi:10.1109/ISCAS.2019.8702303. [[CrossRef](#)]
31. Duong, C.N.; Luu, K.; Quach, K.G.; Nguyen, N.; Patterson, E.; Bui, T.D.; Le, N. Automatic Face Aging in Videos via Deep Reinforcement Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10005–10014. doi:10.1109/CVPR.2019.01025. [[CrossRef](#)]
32. Genovese, A.; Piuri, V.; Scotti, F. Towards Explainable Face Aging with Generative Adversarial Networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–29 September 2019; pp. 3806–3810. doi:10.1109/ICIP.2019.8803616. [[CrossRef](#)]
33. Li, P.; Hu, Y.; Li, Q.; He, R.; Sun, Z. Global and Local Consistent Age Generative Adversarial Networks. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1073–1078. doi:10.1109/ICPR.2018.8545119. [[CrossRef](#)]
34. Palsson, S.; Agustsson, E.; Timofte, R.; Van Gool, L. Generative Adversarial Style Transfer Networks for Face Aging. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2165–2168. doi:10.1109/CVPRW.2018.00282. [[CrossRef](#)]
35. Baldi, P. Autoencoders, Unsupervised Learning and Deep Architectures. *Proc. ICML Unsupervised Transf. Learn.* **2011**, *27*, 37–50.
36. Goodfellow, I.J. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv* **2017**, arXiv:1701.00160.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.

38. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1510–1519. doi:10.1109/ICCV.2017.167. [[CrossRef](#)]
39. Panis, G.; Lanitis, A.; Tsapatsoulis, N.; Cootes, T.F. Overview of research on facial ageing using the FG-NET ageing database. *IET Biom.* **2016**, *5*, 37–46. doi:10.1049/iet-bmt.2014.0053. [[CrossRef](#)]
40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
41. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. doi:10.1109/LSP.2016.2603342. [[CrossRef](#)]
42. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.
43. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-image Translation. In Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
44. Lee, S.; Ha, J.; Kim, G. Harmonizing Maximum Likelihood with GANs for Multimodal Conditional Generation. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
45. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5769–5779.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).