WAVELET-BASED MULTI-LEVEL GANS FOR FACIAL ATTRIBUTES EDITING

Jun Shao

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Computer Science
Concordia University
Montréal, Québec, Canada

November 2021

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Jun Shao**

Entitled: **Wavelet-based Multi-level GANs for Facial Attributes Editing**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Ching Yee Suen	Chair
Dr. Sudhir Mudur	Examiner
Dr. Ching Yee Suen	Examiner
Dr. Tien D. Bui	Supervisor
Dr. Adam Krzyzak	Co-Supervisor

Approved _____
Dr. Hovhannes Harutyunyan
Chair of Department or Graduate Program Director

_____ 20 _____ _____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Wavelet-based Multi-level GANs for Facial Attributes Editing

Jun Shao

Recently, both face aging and expression translation have received increasing attention from the computer vision community due to their wide applications in the real world.

For face aging, age accuracy and identity preserving are two important indicators. Previous works usually rely on an extra pre-trained module for identity preserving and multi-level discriminators for fine-grained features extraction. In this work, we propose a cycle-consistent loss based method for face aging with wavelet-based multi-level facial attributes extraction from both generator and discriminators. The proposed model consists of one generator with three-level encoders and three levels of discriminators with an age and a gender classifier on top of each discriminator. Experiment results on both MORPH and CACD show that the application of multi-level generator can improve the identity preserving effects in face aging and reduce the training time significantly by eliminating the rely of an identity preserving module. Our model can outperform most of the existing approaches including the state-of-the-art techniques on two benchmark aging databases in terms of both aging accuracy and identity verification confidence, demonstrating the effectiveness and superiority of our method.

In real world, expression synthesis is hard due to the non-linear properties of facial skin and muscle caused by different expressions. A recent study showed that the practice of using the same generator for both forward prediction and backward reconstruction as in current conditional GANs would force the generator to leave a potential "noise" in the generated images, therefore hindering the use of the images for further tasks. To eliminate the interference and break the unwanted link between the first and second translation, we design a parallel training mechanism with two generators that perform the same first translation but work as a reconstruction model for each other. Additionally, inspired by the successful application of wavelet-based multi-level Generative Adversarial Networks(GANs) in face aging and progressive

training in geometric conversion, we further design a novel wavelet-based multi-level Generative Adversarial Network (WP2-GAN) for expression translation with a large gap based on a progressive and parallel training strategy. Extensive experiments show the effectiveness of our approach for expression translation compared with the state-of-the-art models by synthesizing photo-realistic images with high fidelity and vivid expression effect.

Acknowledgments

I would like to thank the following people, without whom I would not have been able to complete this research.

First, I would like to thank my supervisor Dr. Tien D. Bui for his great support not only in finance and academic research but also in my life. Thanks for his selfless help through the period of my master program and his kind encouragement each time when I was confronted with frustration and depression.

Then, I'd like to thank my second supervisor Dr. Adam Krzyzak for his guidance in my academic research and thesis preparation.

I also want to thank Duong Hai Nguyen for his great support during the early stage of my research. His wide knowledge and experience in hardware, software and paper writing give me a great help. Thanks to Zhenfei and all other teammates for their selfless supports during my masters life.

Thanks to the Natural Sciences and Engineering Research Council of Canada. This work was supported in part by the Discovery Grant Program of the council to Dr. Bui.

Thanks to the Arbour Foundation which provided me an Arbour scholarship that released my economic pressure effectively during my program.

Finally, I would like to thank my family for all the support and understanding they have shown me through this research.

Related Publications

Here are my publications which are closely related to this thesis:

- **Jun Shao**, Tien D. Bui. Wavelet-based Multi-level GAN for Face Aging. Submitted to Computer Vision and Image Understanding (CVIU), 2021.

- **Jun Shao**, Tien D. Bui. WP2-GAN: Wavelet-based Multi-level GAN for Progressive Facial Expression Translation with Parallel Generators. British Machine Vision Conference (BMVC), 2021.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Facial attributes editing technologies including face aging, facial expression translation, gender translation and so on have a wide application in read world.

Face aging originated from the need of finding missing children has shown significance for cross-age recognition and recreation applications. However, face aging is an intractable task owing to the lack of image set of the same person over a long age span as well as the variants of face poses, the change of illumination, and the existence of occlusion [43].

Although, expression translation shows wide applications to photography technologies, human-computer-interaction and animation movies, facial expression manipulation is challenging owing to the non-linear facial geometric variation caused by different expressions.

Recently, generative models based on wavelet-based multi-level features extractions [44] have shown superiority in synthesizing subtle features. It is meaningful to apply the function of wavelet-based multi-level features extraction to our facial attributes editing tasks.

1.2 Thesis Outline

The outline of the thesis is as follows:

Chapter 2 refers to the work related the wavelet-based multi-level GAN for face

aging. The proposed model consists of a wavelet-based multi-level generator and three wavelet-based multi-level discriminators with an age and a gender classifier on top of each discriminator. Extensive experiments show that utilization of multi-level generator combined with wavelet transform decomposition can improve the identity verification confidence in face aging, and significantly reduce the time for model training by eliminating the use of an identity preserving module. The related paper:

- **Jun Shao**, Tien D. Bui. Wavelet-based Multi-level GAN for Face Aging. Submitted to Computer Vision and Image Understanding (CVIU), 2021.

Chapter 3 mentions the work related to a wavelet-based multi-level GAN for progressive facial expression translation with parallel generators. Two parallel generators were introduced to the facial expression translation task and to eliminate the interference existed in previous methods that is caused by using one single generator for both forward and backward translation. Additionally, we designed a novel progressive training strategy based on the parallel generators, combined with wavelet-based multi-level discriminators to improve the quality of expression translation. Extensive experiments showed that our method outperformed the current state-of-the-art models in terms of both expression translation accuracy and image quality. The related paper:

- **Jun Shao**, Tien D. Bui. WP2-GAN: Wavelet-based Multi-level GAN for Progressive Facial Expression Translation with Parallel Generators. British Machine Vision Conference (BMVC), 2021.

Chapter 4 summarizes the main contributions of this work as well as the ideas for future work.

Chapter 2

Wavelet-based Multi-level GANs for Face Aging [58]

2.1 Introduction

Face aging including progression and regression has attracted much attention from the community of computer vision in last decade. Face aging originated from the need of finding missing children has shown significance for cross-age recognition and recreation applications. However, face aging is an intractable task owing to the lack of image set of the same person over a long age span as well as the variants of face poses, the change of illumination, and the existence of occlusion [43]. Although difficult it is, face aging has achieved great progress owing to the rapid development of deep neural networks. Especially, the recent advent of Generative Adversarial Networks (GANs), which have obtained an amazing achievement in generating photorealistic face images [2, 24, 28, 54, 78], has opened a new door to the face manipulating technologies.

In the area of face aging, most of the GAN-based methods [19, 37, 38, 44, 68, 71] adopt a pretrained neural network for identity preserving. However, empirical experiments show that adding a pretrained deep neural network (e.g. a VGG-Face descriptor [50]) will increase the training time dramatically. In this work we consider face aging as a multi-domain translation task and adopt a CycleGAN-based [79] method for identity preserving. Although effective in simulating both face progression and regression with a single model, the current CycleGAN-based aging methods [60,

62] suffer from generating blurring results, less fine-grained details and even artifacts.

Recently, the work related to multi-level GAN [44, 71] shows that facial features extracted from images at multiple scales by discriminators can force the generator to synthesize vivid aging effects by back-propagation of the adversarial/condition loss from the discriminators to the generator. Intuitively, feeding multi-level features of original inputs to multi-level generator should be able to directly improve the performance of the model in face aging but few works dig in this way. GLCA-GAN [37] leveraged one global and three local generators to capture facial features at different scales for face aging. However, this approach can only provide facial attributes at no more than two levels.

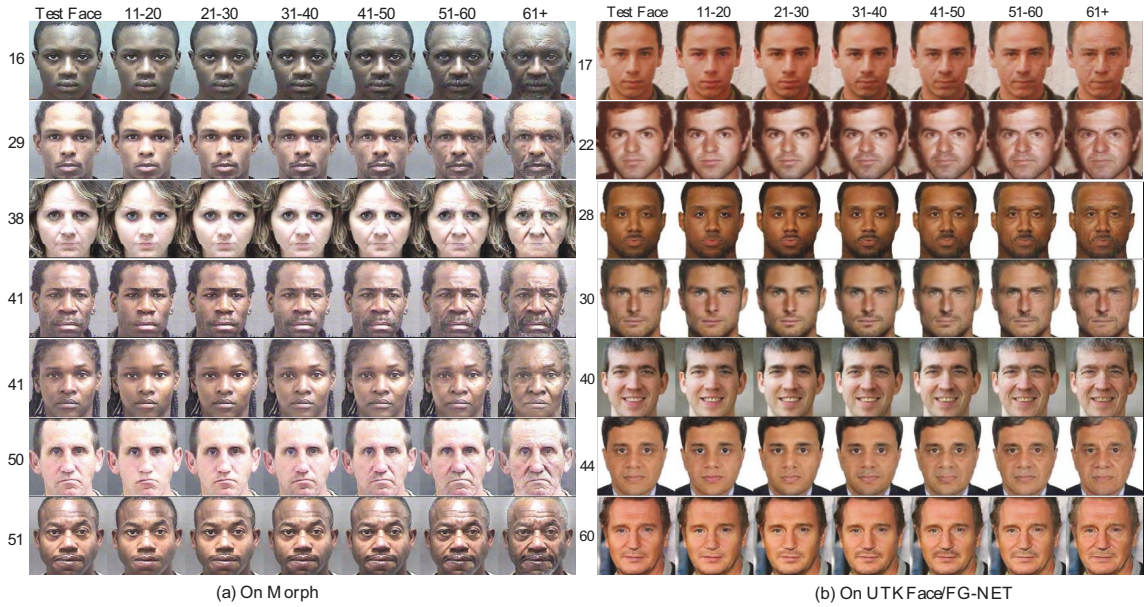


Figure 1: Sampled results of face aging (a) on MORPH and (b) on UTKFace/FG-NET with an age span of 10. The first column contains the input faces, followed by the synthesized images in age group 11-20, 21-30, 31-40, 41-50, 51-60 and 61+.

In our approach, we design a novel multi-level generator combined with a wavelet packet transform (WPT) module for facial features extraction. The original inputs are decomposed into multi-level wavelet coefficients that are used as inputs to the generator with multi-level encoders. Fusing information in different frequency bands with multi-level resolution will provide more identity-related information to the generator. Compared with the image pyramid used in [62], multi-level wavelet coefficients

possess more information in different frequency.

Similar to Wavelet-GAN [44], our approach also performs wavelet-based multi-level features extraction in the discriminators but we adopt three parallel discriminators instead of concatenating the results of three discriminators as one tensor at the end. By this way, we can add a gradient penalty loss [25] to each discriminator to stabilize the training of our model.

In this work, our model consists of a wavelet-based multi-level generator (G) and three wavelet-based multi-level discriminators (D1, D2 and D3) with an age and a gender classifier on top of each discriminator. The generator consists of three encoders (G1, G2, G3) with a gradual decrease of depth. It takes as input the translated multi-level wavelet coefficients and a target condition ($\mathbf{y}_g$) to generate an image with the target attributes. To enable the model to distinguish age and gender features, we concatenate the age group label with the gender label as a combined condition. Thus, our well trained model is not only effective for face aging but also for gender conversion. Three level discriminators are used to force the generator to synthesize indistinguishable images. To force the generated image to fall into the target age and gender group, two classification (age and gender) losses are added to the total loss function of our model. Finally, a cycle-consistent loss and a pixel loss are introduced to the total loss function to preserve the identity-level and pixel-level consistency. Unlike [38, 44, 68, 71], our model does not rely on an additional module for identity preserving, which can reduce the time for training significantly.

The contributions of this work are summarized as follows.

(1) Wavelet Packet Transform module and multi-level encoders are applied to the generator of GANs for face aging for the first time.

(2) Application of multi-level generator combined with wavelet transform decomposition can improve the identity verification confidence in face aging, and significantly reduce the time for model training by eliminating the use of an identity preserving module.

(3) Extensive experiments demonstrate the superiority and effectiveness of our method by synthesizing vivid aging effects and outperforming the existing state-of-the-art models in both face aging accuracy and identity verification confidence.

(4) Experiments for continuous face aging show that our model can generate images with a continuous and smooth increase of age.

2.2 RELATED WORKS

2.2.1 Face Aging

Traditional approaches for face aging can be categorized into two classes: prototype-based methods [21, 30] and physical model-based methods [63, 64]. Physical models usually focus on the change of facial texture, shape, and other physical measures [68]. But these models are normally complex and require a large number of data. Prototype-based approaches leverage the differences between the average face information of different age groups for age pattern transfer [21, 30]. These methods neglect the differences between different person, thus are likely to synthesize unlike face images.

Over the past decade, deep neural networks have achieved great success in representation learning and have also been widely applied to face aging. Wang et al. [66] proposed a recurrent face aging framework that is able to smooth the aging process, but less identity information can be maintained. Duong et al. successfully applied Deep Restricted Boltzmann Machines [12, 13, 14] and Deep Reinforcement Learning [15, 16] to face aging and have achieved an impressive performance. However, these approaches still need paired images with different ages of the same person for model training.

The advent of GANs ushered a vigorous development of face aging. Antipov et al. [1] first deployed the conditional GAN for age transformation. Later, Zhange et al. [75] proposed a conditional adversarial auto-encoder (CAAE) that leverages the high-level features of input images for the generation of target images. But their synthesized faces show little change on aging effect for different age conditions.

Many existing studies [29, 37, 44, 68, 71] in face aging adopted multiple GANs to model age progression or regression, thus limited the efficiency of the methods. For example, Yang et al. [71] proposed PAG-GAN which takes faces below 30 years old as inputs. Different models need to be trained for different target age groups. Huang et al. [29] released the limits to some extent by assigning a different model for each pair of adjacent age groups and trained the whole system in a progressive way, but multiple models were still needed. Additionally, all of these methods rely on a pretrained deep neural network for identity preserving, increased the time of model training.

Another line of research in face aging is based on CycleGAN, which models the age progression and regression with one or two models. For example, Song et al. [60] proposed to use two conditional GANs (Dual cGANs) to model face progression and regression respectively. A reconstruction loss was adopted to maintain the identity of inputs. Sun et al. [62] proposed Ranking GAN which considers both age progression and regression as a multi-domain translation task based on Cycle-GAN. To model the inter-relationship of age patterns among different age groups, Fang et al. [19] designed Triple-GANs learned with a triple translation loss. But a pretrained model for identity-related features extraction was still adopted.

To capture fine details of facial attributes, features at multi-level resolution were widely applied in literature. For instance, Li et al. [37] proposed GLCA-GAN, using one global generator and three local generators to capture both global and local changes during face aging process. Additionally, multi-level discriminators [44, 62, 71] were also popularly utilized to extract features at multiple scale. Liu et al. [44] combined multi-level discriminator with a wavelet packet transform (WPT) module for high-level age-related features extraction to improve the visual fidelity of generated images. Generally, our approach adopts the cycle-GAN based strategy for both face progression and regression with a single model. To improve the face aging effect and fidelity of generated images, we successfully applied wavelet-based multi-level feature extraction to both generator and discriminators. While most of previous multi-level aging models [44, 62, 71] only adopt multi-level discriminators. Besides, our model does not extract facial features of different part physically like GLCA-GAN [37] but leverages multi-level encoders to extract identity related features automatically, thus improving the effectiveness.

All of the methods mentioned above consider face aging in terms of age group. Two recent work [38, 48] designed their aging models on disentangled or learned identity and age embedding and are able for long lifespan face aging, i.e. from children to aged adults. However, both of them suffers from details or background information maintaining during aging process. Additionally, LifeSpan-GAN [48] is sensitive to the background information and shows limits in special cases such as with extreme pose, glasses and occluded face. Other related work includes [76, 77] that proposed a deep architecture AIM unifying face aging and age recognition in a mutual boosting way.

In this work, our model only focuses on adult aging. The well-trained model

can be used to generate fine-grained aging and rejuvenating results for adults with accurate aging effects and high identity preserving confidence.

2.2.2 Multi-level Feature Extraction

It is a common practice to synthesize photo-realistic images with fine-grained attributes by providing multi-level features from inputs to models. Both [37] and [70] adopted one global and three local generators to capture both global and local facial features for face attributes editing.

Li et al. [35, 36] proposed integrated face analytic networks for multi-task face synthesis and analysis. Their work shows that features learned from different tasks can boost the performance of each task.

Besides providing more information to generators, another line of multi-level approach focused on supervising the predictive results of the generator. For example, [71] and [62] adopted multi-level discriminators to supervise the quality and accuracy of the synthetic images. Recently, wavelet based decomposition has been successfully applied to image classification and restoration [39, 42] by providing multi-level features in the frequency space. Liu et al. [44] incorporated a Wavelet packet transform (WPT) module to the multi-level discriminator of Wavelet-GAN to capture age-related features at multiple scales in frequency space.

Different from [37] and [70] that physically assigned global image and local patches to different generators or [71] and [62] that only adopt multi-level discriminators, we integrate a WPT-based image decomposition into both discriminators and generator.

In addition to imposing supervision on the last layer of the generator, a few previous work [8, 9, 49] focused on the control of last several latent layers of the generator, as deep supervision on the hidden layers of neural networks can improve the performance of the models [34]. Since using a wavelet-based decomposition for multi-level features extracting in the frequency space, we only adopt the supervision on the final result of the generator for compactness.

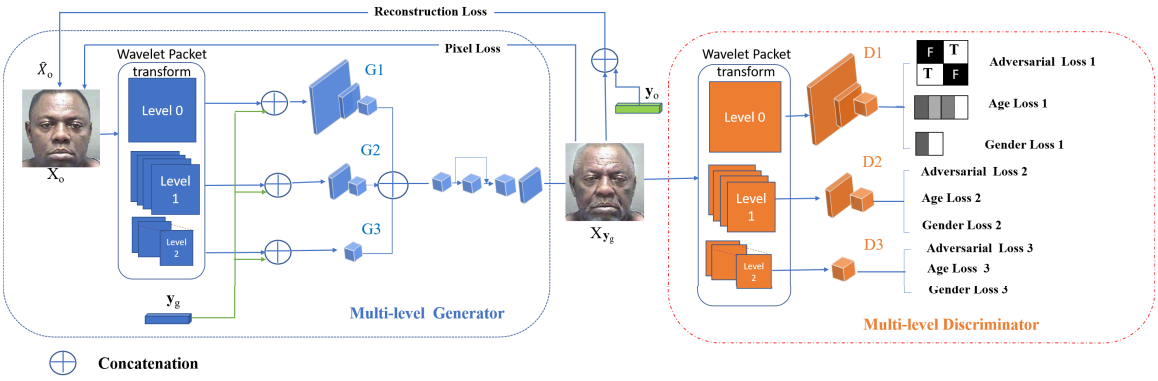


Figure 2: An overview of the wavelet-based multi-level Generative Adversarial Network framework. Input image is converted to three levels of wavelet coefficients by a wavelet packet transform (WPT) module before entering the three level encoders of the generator G, which takes as input the multi-level wavelet coefficients and the target condition y_g to synthesize a photo-realistic image X_{y_g}. The multi-level discriminators take each level of the wavelet coefficients as input and evaluate the realism of given images as well as the condition loss. The generator G is called twice to reconstruct the original image and a reconstruction loss (or cycle-consistent loss) is applied to preserve the identity of input image.

2.3 PROPOSED METHOD

2.3.1 Overview

Considering age effects as a special facial style, we propose a novel conditional GAN with multi-level path in both generator and discriminator for face progression/regression based on an image-to-image translation method. Our model consists mainly of two blocks: a multi-way generator $G(X_o|\mathbf{y}_g)$ transforming the aging effects in the original image X_o to the desired condition $\mathbf{y}_g$, and three-level discriminators (D_1, D_2 and D_3) distinguishing the photo-realism as well as the facial attributes such as age group and gender.

To assist the multi-level generator and discriminators to capture identity-related and age-related texture details, a wavelet packet transform (WPT) module is performed to decompose the given image into high-frequency and low-frequency coefficients at multiple scales (as shown in Fig. 3). The low-frequency components preserve global information of the face while high-frequency components preserve the local details.The wavelet coefficients at each decomposing level concatenated with other

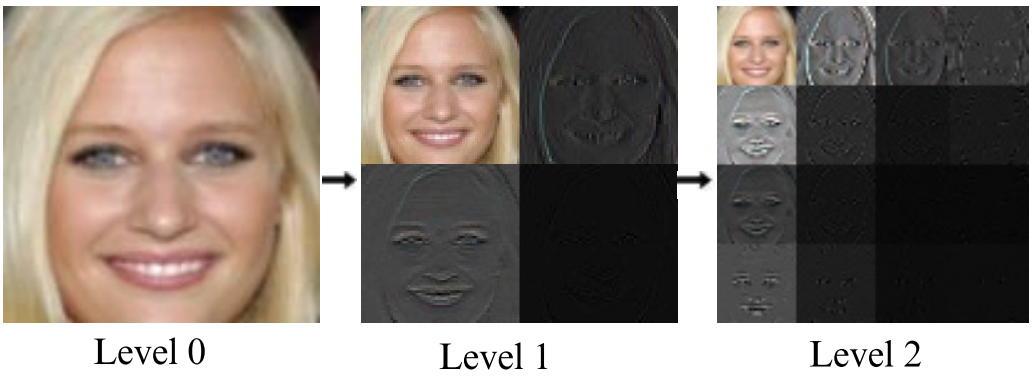


Figure 3: A sample image with multi-level wavelet coefficients decomposed by wavelet packet transform.

conditions are fed into a corresponding encoder of the generator or discriminator.

Considering that face aging task on age groups with a small age span is challenging than on age groups with a large span [62], we train and test our approach on age groups with two different age spans (i.e. 5 and 10 years).

We combine the gender information with age group label as the final condition. Gender information here can not only help the model to peel off gender attributes from aging effects but also endow our model the capability of gender translation.

To get rid of image pairs of the same person, the generator is called twice, first from the original image to synthesize the desired image then reconstructs the original one from the generated face. The re-construction loss helps to preserve the identity of the outcome.

The overview of our proposed framework is displayed in Fig. 2.

2.3.2 Wavelet-based Multi-level Generator

Although multi-level discriminators have been successfully applied to face synthesis, few direct applications of multi-level features extracted from generator were reported in literature. Previous works usually utilize an auto-encoder based generator (with one encoder and one decoder) for image synthesizing. Intuitively, application of multi-level features extracting in generator can also improve the performance of the model in face aging by providing multi-level facial attributes to the decoder of the generator. Although the condition loss calculated after the discriminators can be back-propagated to the generator through the discriminator, the effect is indirect and may

Table 1: Neural Network Architecture of Multi-level Generator ('K','S' and 'P' denote 'Kernel size','stride' and 'padding').

	Layer	Output Size			Details
		Enc1	Enc2	Enc3	
Input	ConvI	$128 \times 128 \times 64$	$64 \times 64 \times 64$	$32 \times 32 \times 128$	K7, S1, P3
Down-sampling	ConVD1	$64 \times 64 \times 128$	$32 \times 32 \times 128$	$16 \times 16 \times 256$	K4, S2, P1
	ConVD2	$32 \times 32 \times 256$	$16 \times 16 \times 256$	—	K4, S2, P1
	ConVD3	$16 \times 16 \times 512$	—	—	K4, S2, P1
Residual Blocks	ConvR1-1	$16 \times 16 \times 512$	$16 \times 16 \times 256$	$16 \times 16 \times 256$	K3, S1, P1
	ConvR1-2	$16 \times 16 \times 512$	$16 \times 16 \times 256$	$16 \times 16 \times 256$	K3, S1, P1
	...	...	...		
	ConvR6-1	$16 \times 16 \times 512$	$16 \times 16 \times 256$	$16 \times 16 \times 256$	K3, S1, P1
	ConvR6-2	$16 \times 16 \times 512$	$16 \times 16 \times 256$	$16 \times 16 \times 256$	K3, S1, P1
Up-sampling	ConvU1		$32 \times 32 \times 256$		K4, S2, P1
	ConvU2		$64 \times 64 \times 128$		K4, S2, P1
	ConvU3		$128 \times 128 \times 64$		K4, S2, P1
Output	ConvO		$128 \times 128 \times 3$		K7, S1, P3

Note: Before entering the up-sampling layers, tensors of three encoders are concatenated in channel.

be impacted by the vanishing gradient problem caused by long-range dependencies.

In this work, we directly apply multi-level features extracting to image synthesizing and form a generator consisting of three encoders (G1, G2, G3) and one decoder. To make sure the encoders can receive both low-frequency information (approximation) and high-frequency information (details) as inputs, at first, the input image is transferred to multi-level wavelet coefficients by a Wavelet Packet Transform (WPT)[1] [57] module. Then, concatenated with the target condition, multi-level coefficients are fed into the corresponding pathway of the generator. Considering the gradually shrinking size of the coefficients, we design multi-level encoders with a decreasing number of down-sampling layers.

Mapping features (with the same height and width) from multi-level encoders are concatenated in depth as one tensor, which is then fed to the up-sampling convolutional layers until reconstructing an image with the size of $128 \times 128 \times 3$.

The work has the most similar idea to us is [37], but the approach there needs four separate generators to learn a global face and three local patches, then a merging network to fine-tune the final result. While our method only needs a multi-level generator possessing three encoders, incorporated with wavelet-based transformation, to capture multi-level features from the inputs, thus more effective. The structure of the multi-level generator is shown in Table 1.

[1]https://pytorch-wavelets.readthedocs.io/en/latest/readme.html.

Table 2: Neural Network Architecture of Multi-level Discriminators ('K','S' and 'P' denote 'Kernel size','stride' and 'padding').

	Layer	Output Size			Details
		D1	D2	D3	
Input	ConvI	$64 \times 64 \times 64$	$32 \times 32 \times 64$	$16 \times 16 \times 128$	K4, S2, P1
Down-sampling	ConVD1	$32 \times 32 \times 128$	$16 \times 16 \times 128$	$8 \times 8 \times 128$	K4, S2, P1
	ConVD2	$16 \times 16 \times 256$	$8 \times 8 \times 256$	$4 \times 4 \times 256$	K4, S2, P1
	ConVD3	$8 \times 8 \times 512$	$4 \times 4 \times 512$	$2 \times 2 \times 512$	K4, S2, P1
	ConVD4	$4 \times 4 \times 1024$	$2 \times 2 \times 1024$	—	K4, S2, P1
	ConVD5	$2 \times 2 \times 2048$	—	—	K4, S2, P1
Output	ConvO-Adv	$2 \times 2 \times 1$	$2 \times 2 \times 1$	$2 \times 2 \times 1$	K3, S1, P1
	ConvO-Age	$1 \times 1 \times GP$	$1 \times 1 \times GP$	$1 \times 1 \times GP$	K2, S1, P0
	ConvO-gender	$1 \times 1 \times 2$	$1 \times 1 \times 2$	$1 \times 1 \times 2$	K2, S1, P0

Note: GP is the number of age groups the aging data set is divided, i.e., GP=4 or GP=7.

2.3.3 Wavelet-based Multi-level Discriminators

To capture age-related attributes such as wrinkles and eye bags, a Wavelet Packet Transform (WPT) module is also incorporated into the multi-level discriminators.

In this work, we adopt three levels of discriminators and each discriminator has a gradually decreasing number of convolutional layers so that three levels of wavelet coefficients can be encoded into three matrices with the same size $Y_{D_i} \in R^{H/2^6 \times W/2^6}$, where $i = \{1, 2, 3\}$, H and W are the height and width of the input image. Each element of Y_{Di} represents the probability of the corresponding patch to be real.

To stabilize the training of the adversarial learning system, we add a penalty loss [25] to the gradient norm of each critic. Thus, we do not concatenate the outcomes of three discriminators as one tensor as [44] did. Empirical experiments show that separate discriminators do not cost much time than combined discriminators during model training. While the parameters of separate discriminators can be well regularized by their corresponding gradient norm loss functions.

Unlike [44, 60] that directly adding semantic attributes to the discriminator side, we adopt a multi-task learning framework to introduce conditioning of the discriminators by minimising the condition cross entropy loss. Besides photo-realism, three discriminators are also responsible for estimating the age group and gender of the given image. To reduce the number of parameters, we add auxiliary classifiers for age and gender classification, on the last second layer of each discriminator.

The structure of three discriminators are shown in Table 2 .

2.3.4 Overall Objective Functions

To make the generated images indistinguishable from real images, we adopt the adversarial loss proposed by WGAN-GP [25]. The average adversarial loss of this work is defined as:

$$\mathcal{L}_{avd} = \frac{1}{3}\sum_{i=1}^{3}\left\{\mathbb{E}_{X_o\sim\mathbb{P}}\left[D_i\left(G\right)\right] - \mathbb{E}_{X_o\sim\mathbb{P}}\left[D_i\left(X_o\right)\right] + \lambda_{gp}\mathbb{E}_{\widetilde{X}\sim\widetilde{\mathbb{P}}}\left(\parallel \nabla_{\widetilde{X}} D_i\left(\widetilde{X}\right) \parallel_2 -1\right)^2\right\}, \tag{1}$$

where $G = G\left(X_o|\mathbf{y}_g\right)$, $\mathbb{P}$ is the data distribution of input image X_0, $\widetilde{X}$ represents the random interpolated image by input image X_0 and its generated face $X_{\mathbf{y}_g}$, $\widetilde{\mathbb{P}}$ stands for the uniform interpolation distribution and λ_{gp} is a penalty coefficient.

Besides the adversarial loss, the generator and the discriminators also have to reduce the errors produced by the age and gender classifiers imposed on top of D_i. By this way, the model can not only generate photo-realistic images but also force the synthesized images to have desired aging effect and gender. During the training, the condition losses can be decomposed into two parts. On one hand, the condition losses associated with fake images are deployed to optimize the generator G. On the other hand, the condition losses for true images are used to improve the discriminator D_i. The whole condition losses can be written as:

$$\mathcal{L}_{age} = \frac{1}{3}\sum_{i=1}^{3}\left\{\mathbb{E}_{X_o\sim\mathbb{P}}\left[\parallel D_i(G) - \mathbf{y}_{g(C_{age})} \parallel_2^2\right] + \mathbb{E}_{X_o\sim\mathbb{P}}\left[\parallel D_i(X_o) - \mathbf{y}_{o(C_{age})} \parallel_2^2\right]\right\}, \tag{2}$$

$$\mathcal{L}_{gd} = \frac{1}{3}\sum_{i=1}^{3}\left\{\mathbb{E}_{X_o\sim\mathbb{P}}\left[\parallel D_i(G) - \mathbf{y}_{g(C_{gd})} \parallel_2^2\right] + \mathbb{E}_{X_o\sim\mathbb{P}}\left[\parallel D_i(X_o) - \mathbf{y}_{o(C_{gd})} \parallel_2^2\right]\right\}, \tag{3}$$

where $G = G\left(X_o|\mathbf{y}_g\right)$, $\mathbf{y}_{g(C_{age})}$, $\mathbf{y}_{g(C_{gd})}$ are target age group and gender label, while $\mathbf{y}_{o(C_{age})}$, $\mathbf{y}_{o(C_{gd})}$ are original age group and gender label, respectively,.

In addition, a cycle-consistent loss [79] and a pixel loss are utilized by the generator to preserve the identity-level and image-level consistency. We adopt the l_1 norm, which helps to capture features associated with low-frequencies. The cycle-consistent loss and pixel loss are formulated as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{X_o\sim\mathbb{P}}\left[\parallel G(G(X_o|\mathbf{y}_g)|\mathbf{y}_o) - X_o \parallel_1\right], \tag{4}$$

$$\mathcal{L}_{pix} = \mathbb{E}_{X_o \sim \mathbb{P}} \left[\| \, G(X_o|\mathbf{y}_g) - X_o \, \|_1 \right], \tag{5}$$

Finally, we can organize the loss functions for G and D as:

$$\begin{aligned} \mathcal{L}_D =& \mathcal{L}_{adv} + \lambda_{age}\mathcal{L}_{age} + \lambda_{gd}\mathcal{L}_{gd} \\ \mathcal{L}_G =& \mathcal{L}_{adv} + \lambda_{age}\mathcal{L}_{age} + \lambda_{gd}\mathcal{L}_{gd} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{pix}\mathcal{L}_{pix}, \end{aligned} \tag{6}$$

where λ_{age}, λ_{gd}, λ_{cyc} and λ_{pix} are hyper-parameters that control the relative importance of conditional loss, cycle-consistent loss and pixel loss.

2.4 EXPERIMENTS

2.4.1 Dataset

We train and validate our model on four popular aging databases. MORPH [55] is a well-known benchmark for age estimation and face aging, which contains 55,000 color images of more than 13,000 individuals. The ages of the subjects range from 16 to 77 years old. Similar to [62], we at first consider ages from 16 to 50 and divide the data set (51,700 images) into 7 age groups with an age span of 5, i.e., 16-20, 21-25, 26-30, 31-35, 36-40, 41-45 and 46-50. We randomly choose 4/5 subjects for training and the remaining unseen subjects for test. Following that, 41,396 and 10,304 images are collected for training and testing, respectively.

CACD [6] contains 163,446 face images of 2,000 celebrities with age ranging from 14 to 62. CACD is a challenging dataset for face aging as CACD contains a large number of images with large variants in head pose, expression, illumination and occlusion. Moreover, CACD is collected via the Google Image Search, making the age labelling error-prone. To reduce age label errors, we compare the existing age labels of images in CACD with predictive age labels from [18] and screen out images with an age gap bigger than 20 years. We divide CACD dataset into 7 age groups from 16 to 50 with an age span of 5. We randomly select 1/10 images for model test and the rest for training.

To compare our model with the baselines PAG-GAN, GLCA-GAN and Wavelet-GAN, we also divide MORPH and CACD dataset into 4 age groups as 30-, 31-40,

41-50 and 50+. CACD dataset does not provide the gender label, we leverage Face++ to generate gender label for the dataset.

UTKFace [75] is another large-scale face database containing over 20,000 face images in the wild. Considering the change of hair colour usually occurs from 50 to 60 years old, we divide both Morph and UTKFace into to 6 age groups as 11-20, 21-30, 31-40, 41-50, 51-60 and 61+ to show an obvious effect of aging hair. FG-NET [33] contains 1,002 face images of 82 subjects. We leverage it as the testing set to evaluate the generalization of our model trained on UTKFace.

Finally, all images are aligned, cropped and resized to the size of 128×128 by MTCNN [74].

2.4.2 Implementation Details

We utilize Adam [31] for model optimization with the following hyper-parameters: learning rate = 0.0001, beta1=0.5, beta2=0.999, and batch size=25. The generator is optimized once after five times optimization of the discriminator. On MORPH, we train the model for 150 epochs and linearly decay the learning rate to zero over the last 50 epochs. On CACD, our model is trained for 70 epochs and the learning rate decays for the last 20 epochs. The weight coefficients for the loss functions are set to $\lambda_{gp} = 10$, $\lambda_{age} = 10$, $\lambda_{gd} = 5$ and $\lambda_{cyc} = 10$ for all experiments. The parameter of λ_{pix} is set to be 0.05, 0.01 and 0.02 for experiments on 7 age groups of MORPH, 4 age groups of MORPH and all experiments on CACD, respectively. Our model is trained for 22h on MORPH and 28h on CACD, respectively, with a single Tesla V100 GPU.

2.4.3 Qualitative Results of Face Aging

Fig. 4 display samples of age transformation on MORPH and CACD dataset with an age span of 5 years. We can observe that the proposed model is competent for a smoothing change of aging effects, synthesizing faces in different age groups with high visual fidelity. Besides, our approach is robust to variations of gender, skin color, head pose and expression.

We also divide both Morph and CACD into 4 age groups (i.e.30-, 31-40, 41-50 and 51+) as a tradition. We can observe a consistent change of aging hair, headline, eye bags, mouth and laugh lines as age increased in Fig. 5 and Fig. 7.

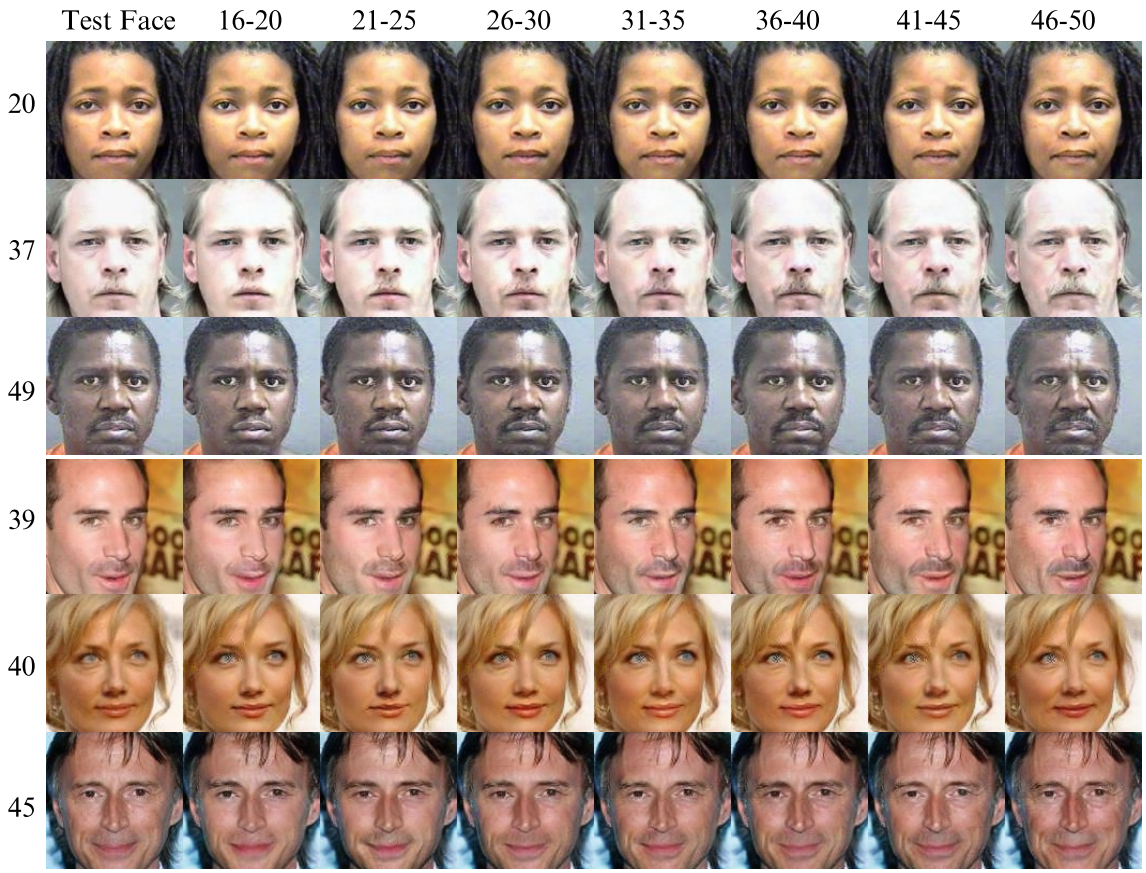
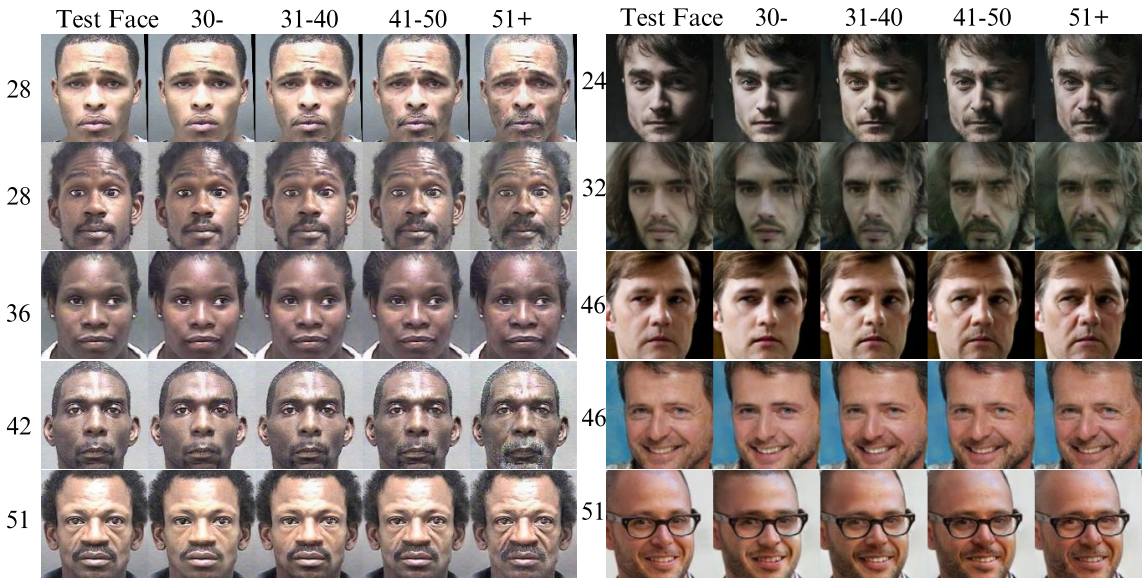


Figure 4: Sample results of face aging on MORPH (the first three rows) and CACD (the last three rows). The first column represents the input faces for testing, followed by the synthesized images in age group 16-20, 21-25, 26-30,31-35, 36-40, 41-45 and 46-50.

Considering the change of hair colour usually happens after 50 years old and the limited samples in aged group (61+) of CACD, we divide Morph and UTKFace into 6 age groups as 11-20, 21-30, 31-40, 41-50, 51-60 and 61+ to display the synthesis of aging hair. As shown in Fig. 1, we can perceive an obvious change of hair/beard colour and texture besides adding of wrinkle on faces of Morph and UTKFace/FG-Net after entering the age groups of 51-60 and 61+.

Performance comparison with previous models is shown in Fig. 6. To compare with IPCGAN, Ranking GAN and DAAE, we train our models on both MORPH and CACD with an age span of 5. To compare with PAG-GAN, Wavelet-GAN and Triple-GAN, we divide both MORPH and CACD databases into 4 age groups i.e.,



(a) On Morph (b) On CACD

Figure 5: Sampled results of face aging (a) on MORPH and (b) on CACD. The first column contains the input faces, followed by the synthesized images in age group 30-, 31-40, 41-50 and 51+.

30-, 31-40, 41-50 and 51+, and train our model on these two databases separately. As the code of most previous models is unavailable, we leverage the generated images from [38, 44, 62, 71] for comparison.

As shown in Fig. 6, IPCGAN has difficulty in predicting valid face in the old age group 45-50, by generating blurring face and misplacing moustache to a female face. Ranking GAN shows capability in both generating detailed aging effects and identity preserving. However, the outcomes of Ranking GAN seem to be blurring.

PAG-GAN, Wavelet-GAN and Triple-GAN can synthesize vivid aging effects and preserve the identity to some extent. However, PAG-GAN and Triple-GAN fail to maintain the illumination, leading to a distinguishable change in skin color and background pixel between inputs and the outcomes. Wavelet-GAN shows superiority in keeping the pixel of background and maintaining the illumination, but has difficulty in preserving some subtle facial attributes such as the pimple on the right cheek of the first sample and the right eye of the second sample. Moreover, Wavelet-GAN generates extra hair behind the neck for the third sample. Compared with PAG-GAN, Triple-GAN and Wavelet-GAN, our method shows obvious supremacy in identity

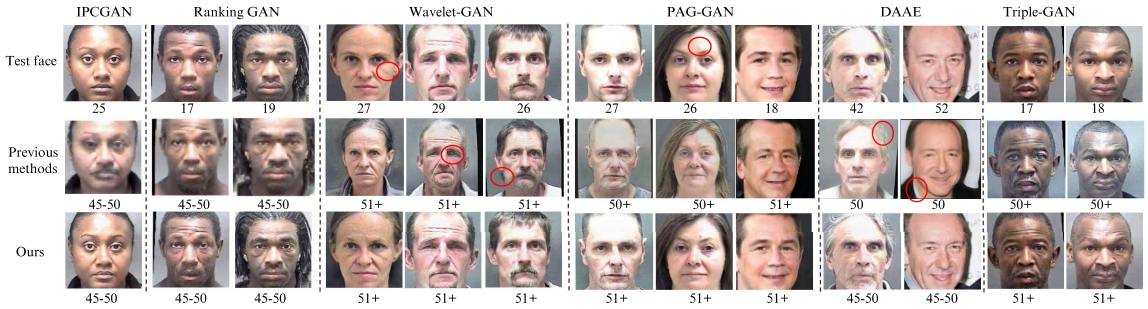


Figure 6: Fourteen sample results compared with IPCGAN [68], Ranking GAN [62], Wavelet-GAN [44], PAG-GAN [71], DAAE [38] and Triple-GAN [19]. From top to bottom are inputs, images generated by previous works and by our model (zoom in for a better view).

Figure 7: Illustration of aging consistency (zoom in for a better view). (1) Aging hair; (2) headline; (3) eyes; (4) moustache and beard; (5) laugh lines.

preserving.

Although DAAE is designed for continuous face aging competent for dataset with a long-tail distribution, it failed to preserve some background information and synthesize vivid facial features. For instance, DAAE failed to preserve the hair texture for the first sample. In the second example, DAAE failed to keep the letters in the background and transformed the blue collar to a while one. At last, the outcomes of DAAE are a little blurring compared to ours.

2.4.4 Aging Accuracy

In this section, we compare the performance of our model with the state of the arts on both MORPH and CACD databases in terms of aging accuracy, which is measured as the mean age difference between training data and the synthetic images. Age estimation was performed by [18] for all synthesized images and all original images

utilized for model training. The age estimation results on MORPH and CACD with an age span of 5 or 10 are shown in Table 3 - Table 5.

On seven age groups of MORPH with an age span of 5, our model outperforms Ranking GAN in 5/7 groups in terms of mean age difference (Table 3). Moreover, Ranking GAN obtains mean age gaps more than 1.0 in three age groups, while our model only has one bigger mean age gap (1.27) in age group 16-20. Compared with our method and Ranking GAN, IPCGAN has a poor performance in most of the age groups in terms of mean age gap (with three age groups more than 2.0 and two age groups more than 1.0).

Table 3: Age estimation and identity verification results on MORPH with an age span of 5, compared with IPCGAN and Ranking GAN (differences of mean ages are measured in absolute value.)

Age group	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Estimated Age Distributions							
Original	23.54	27.56	31.56	36.61	41.28	46.01	51.44
Synthetic	24.81 ± 6.08	28.40 ± 6.58	31.42 ± 6.48	36.40 ± 7.08	41.01 ± 7.09	46.07 ± 6.84	51.35 ± 7.05
Difference of Mean Ages							
IPCGAN	0.31	0.94	1.79	2.10	2.54	1.11	2.14
Ranking GAN	**0.19**	0.87	0.75	1.19	**0.23**	1.67	1.78
Ours	1.27	**0.84**	**0.14**	**0.21**	0.26	**0.06**	**0.09**
Identity verification Confidence							
IPCGAN	94.29 ± 1.57	95.12 ± 1.08	95.41 ± 0.84	95.61 ± 0.59	95.46 ± 0.66	94.86 ± 1.22	93.88 ± 2.01
Ranking GAN	95.58 ± 1.04	95.97 ± 0.68	96.00 ± 0.60	96.08 ± 0.51	95.95 ± 0.62	95.75 ± 0.81	94.93 ± 1.14
Ours	$\mathbf{95.98 \pm 0.83}$	**96.20**± 0.60	$\mathbf{96.15 \pm 0.55}$	$\mathbf{96.21 \pm 0.49}$	$\mathbf{96.05 \pm 0.59}$	$\mathbf{95.82 \pm 0.77}$	$\mathbf{95.29 \pm 0.96}$

Table 4: Age estimation and Face verification results on CACD with an age span of 5

Age group	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Original age	26.34	28.14	31.48	35.58	40.04	44.89	49.90
Synthetic	24.93 ± 5.18	26.94 ± 5.51	30.79 ± 6.34	35.70 ± 7.29	40.93 ± 7.50	44.54 ± 7.09	49.53 ± 6.48
Identity Confidence	91.91 ± 2.82	93.23 ± 2.18	93.51 ± 1.98	93.35 ± 2.26	93.29 ± 2.33	93.18 ± 2.36	93.32 ± 2.28

We validate the performance of our model on the in-the-wild database CACD. The results (Table 4) show that our model can accurately predict face aging effects on age groups of CACD with an age span of 5. Our approach achieves a mean age gap smaller than 1.0 in 5/7 age groups and a mean age gap no more than 2.0 in the other two age groups.

To verify the superiority of our model on age groups with large age span, we compare our model to the state-of-the-art models GLCA-GAN, PAG-GAN, Wavelet-GAN and DAAE on both MORPH and CACD from age group 30- to 51+. Table 5 shows that our model consistently outperforms the state-of-the-arts in all three age

groups of MORPH. On CACD dataset, our model achieves the lowest age gap in age group 31-40 and 41-50 while seeing a performance drop in the eldest age group that may be caused by the age labelling error. DAAE achieves the same aging accuracy as our model in age group 31-40, but inferior results in two other groups.

All experiments above prove the superiority of our model in terms of aging accuracy on both MORPH and CACD.

2.4.5 Identity Preservation Evaluation

Besides aging accuracy, identity verification confidence is another important indicator to evaluate the performance of face aging models. Face verification experiments have been conducted through [18] to evaluate the identity preserving effect of our models on both MORPH and CACD. Like previous works, we compare each image in the test set of MORPH/CACD with the synthetic images of the same person in different age groups.

As shown in Table 3, our model achieves an average verification confidence ranged from 95.29% to 96.21% on MORPH (7 age groups) and consistently outperforms Ranking GAN in identity preserving.

Table 5: Age estimation and identity verification results on MORPH and CACD (differences of mean ages are measured in absolute value)

	MORPH				CACD			
Age group	30-	31-40	41-50	51+	30-	31-40	41-50	51+
Estimated Age Distributions								
Original	27.08	38.94	48.21	58.32	29.23	38.01	47.27	55.98
Synthetic	27.86 ± 7.01	38.88 ± 7.10	48.24 ± 7.08	58.33 ± 7.50	27.57 ± 5.85	38.14 ± 7.34	47.49 ± 6.60	57.66 ± 7.19
Difference of Mean Ages								
GLCA-GAN	—	0.23	3.61	8.61	—	1.72	2.07	2.85
PAG-GAN	—	0.38	0.52	1.48	—	0.70	**0.22**	**0.57**
Wavelet-GAN	—	0.13	0.19	0.68	—	0.37	0.58	0.66
DAAE	—	1.43	1.30	1.45	—	**0.13**	0.68	2.02
Ours	**0.74**	**0.06**	**0.03**	**0.01**	**1.62**	**0.13**	**0.22**	1.68
Identity Verification Confidence								
PAG-GAN	-	94.64 ± 0.03	91.46 ± 0.08	85.87 ± 0.25	-	94.13 ± 0.04	91.96 ± 0.12	88.60 ± 0.15
Wavelet-GAN	-	95.77	94.64	87.53	-	93.67	91.54	90.32
Ours	$\mathbf{96.47 \pm 0.86}$	$\mathbf{96.50 \pm 0.45}$	$\mathbf{95.98 \pm 0.75}$	$\mathbf{94.37 \pm 1.36}$	$\mathbf{94.60 \pm 1.73}$	$\mathbf{94.58 \pm 1.81}$	$\mathbf{94.41 \pm 1.87}$	$\mathbf{93.99 \pm 1.96}$
Identity Verification Rate(%)								
GLCA-GAN	-	97.66	96.67	91.85	-	97.72	94.18	92.29
PAG-GAN	-	**100.00**	98.91	93.09	-	**99.99**	99.81	98.28
Wavelet-GAN	-	**100.00**	**100.00**	98.26	-	99.76	98.74	98.44
DAAE	-	99.48	99.36	99.36	-	99.24	99.19	99.19
Ours	**100.00**	**100.00**	**100.00**	**100.00**	**99.98**	99.96	**99.97**	**99.96**

Compared with PAG-GAN and Wavelet-GAN on both MORPH and CACD, our model exceeds the state-of-the-arts with a large gap in identity verification confidence, especially in the two eldest age groups. As shown in Table 5, our model achieves an identity verification confidence ranging from 94.37% to 96.50% on MORPH, reaping a

rise ranging from 0.73% to 6.84% in identity verification confidence compared to PAG-GAN and a rise ranging from 1.86% to 8.6% compared to Wavelet-GAN. Likewise, our model exceeds PAG-GAN on CACD with a gap ranging from 0.45% to 5.39%, while exceeds Wavelet-GAN with a gap ranging from 0.91% to 3.67%. Results of face verification experiments show the superiority of our model in identity preserving compared to Ranking GAN, PAG-GAN and Wavelet-GAN. Table 5 also shows that our model consistently outperforms DAAE in terms of identity verification rate on both MORPH and CACD.

2.4.6 Ablation Study

In this section, we perform an ablation study to inspect the role of different components of our model for face aging and identity preserving. We evaluate the performance of models including/excluding multi-level generator, multi-level discriminators and gender condition. In this work, model without multi-level generator and without multi-level discriminators refer to model with only one pathway in the generator and model with one level discriminator, respectively. We only perform ablation study on seven age groups of MORPH with an age span of 5 because of the noisy labels of CACD.

As shown in Fig. 8, compared with the proposed model, the model without multi-generator fails to preserve some subtle facial attributes such as the scar on the second sample. We can also observe a slight skin color change between test face and generated images for the first sample. Compared with other models, the model without gender information synthesizes much younger faces in age group 46-50, as fewer wrinkles are observed in the generated images.

Quantitative comparison is shown in Table 6. Compared to the proposed model, variants without multi-level discriminator or gender condition have a performance drop on face aging accuracy, as there are big discrepancies between the mean age of synthetic images and original images in several age groups. This means multi-level discriminators and the assistance of gender attribute are necessary for guaranteeing the accuracy of face aging. Although model without multi-level generator has no obvious performance drop in face aging accuracy on MORPH database, the identity preserving confidence turns out to be the lowest compared to other models. It means

Figure 8: Ablation study results on MORPH. The first column contains test faces. From the second column to the most right are outcomes in age group 46-50 generated by the model without multi-level generator(No Multi-G), model without discriminator(No Multi-D), model without gender condition(No Gender) and the proposed model.

Table 6: Comparison of face aging accuracy and identity verification confidence on MORPH between variants of the proposed model (differences of mean ages are measured in absolute value).

Age group	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Deviation of Estimated Ages							
woMG	**0.16**	**0.45**	0.91	**0.06**	0.53	0.49	1.61
woMD	2.10	1.55	0.15	0.25	0.97	0.58	2.19
woGender	2.33	1.39	**0.03**	5.25	1.52	4.01	0.38
Proposed	1.27	0.84	0.14	0.21	**0.26**	**0.06**	**0.09**
Identity Verification Confidence							
woMG	95.33	95.43	95.35	95.32	95.04	94.88	94.29
woMD	96.26	96.43	96.47	96.50	96.41	96.15	95.95
woGender	96.89	96.82	96.86	96.89	96.84	96.89	96.89
Proposed	95.98	96.20	96.15	96.21	96.05	95.82	95.29

Note: woMG represents model without multi-level generator; woMD: without multi-level discriminators.

Table 7: Comparison of time costing for model training per epoch and inference among variants of the proposed model (150 epochs on MORPH).

Model	woMG	woMD	woGender	Proposed	Proposed+IP
Training Time (min/epoch)	7	6	8	9	31
Inference Speed (frames/sec)	34.28	11.47	11.77	11.47	11.47

that cycle-consistent loss alone is not enough to maintain the identity, while multi-level generator is another key component for identity preserving. The effectiveness of the multi-level generator in identity preserving may caused by the multi-level features in different frequency bands extracted by the multi-level encoders in the generator G.

The model without multi-level discriminators achieves the lowest face aging accuracy, proving the importance of multi-level discriminators in face aging.

Compared to the proposed model, the model without gender condition has a better effect in identity preserving but bad face aging accuracy, which means that to preserve identity effects, some identity related features have been merged into aging attributes. Considering the contribution of gender label in face aging accuracy, we choose the model with gender attribute as the final model.

Finally, we compare the time for model training among variants of the proposed model. We also train a proposed model combined with an identity preserving module [50], similar to PAG-GAN and Wavelet-GAN. The identity preserving module is called once when the generator is trained. The training time for each epoch is shown in Table 7. The proposed model costs 9 minutes for each epoch of training and about 22.5h for the whole process of training. Although model without multi-level discriminators (woMD) and model without multi-level generator (woMG) achieve a shorter training time (6 and 7 minutes per epoch, respectively) than the proposed model due to partial lack of multi-level features extraction, the difference is slight. However, adding a pre-trained deep neural network for identity preserving will increase the time for training significantly by 3.4 times. Due to the application of multi-level encoders, the proposed model achieves a lower inference speed (11.47 frames/sec) than the variant without multi-level generator. But this sacrifice is worth considering the performance improvement in face aging and identity preserving effect.

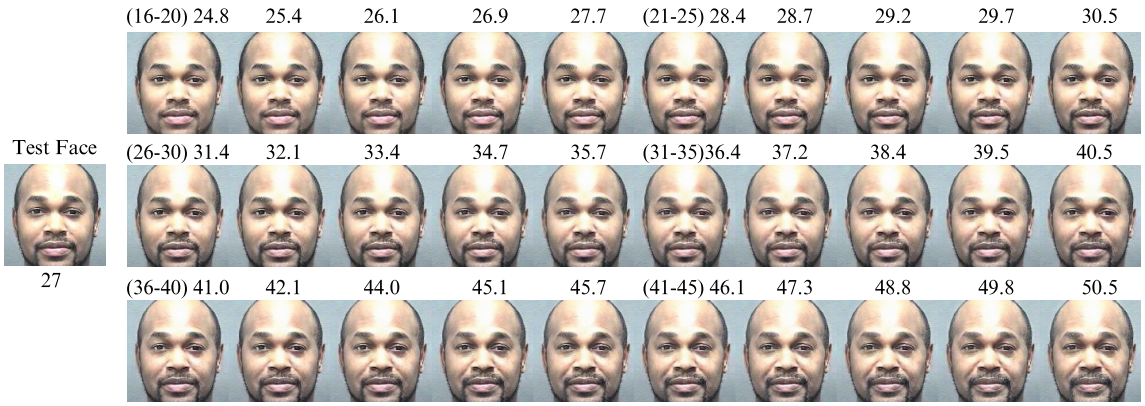


Figure 9: A sample of continuous face aging on MORPH. The first column contains the test faces. The second and seventh columns are synthetic images corresponding to discrete age groups. Other columns are interpolated results. The numbers above all synthetic images are the mean age corresponding to the target age.

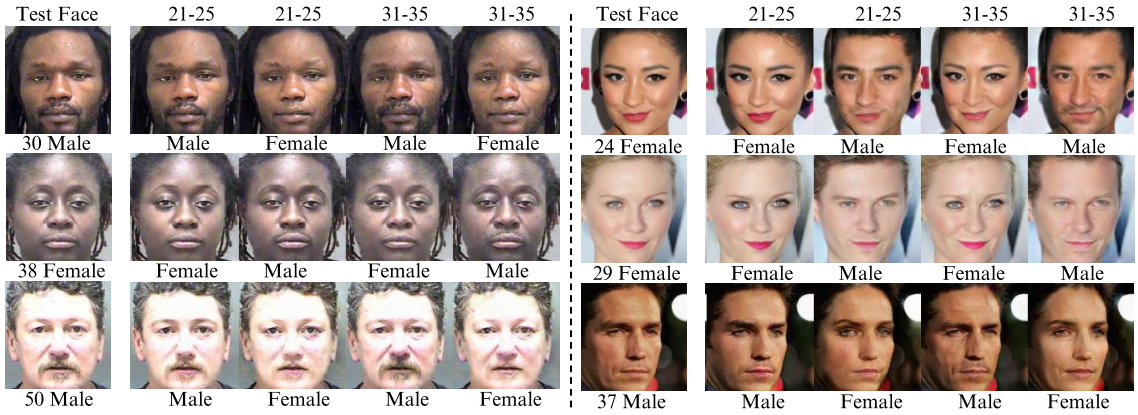
2.4.7 Continuous Face Aging

Motivated by [72], we evaluate the performance of our model qualitatively and quantitatively for continuous face aging on MORPH. As the aging datasets are divided into 7 age groups with an age span of 5, we bridge the gap between two adjacent age groups with a linear formula: $C_\alpha = C_{age_m} + \alpha \times (C_{age_{m+1}} - C_{age_m})$ with $\alpha = \{0, 0.35, 0.45, 0.55, 0.7\}$ and m is the index for the age groups. To save the time for testing, we randomly test 2,000 images for each interpolated age condition.

As shown in Fig. 9, the left most column is the test face. The second and seventh columns are images corresponding to discrete age groups with the age group label and the mean age on the top, while other columns are interpolated outcomes. Above each interpolated image is the mean age of the generated faces belonging to the target age condition. The age labels are estimated by Face++ API. We can find that between the discrete age groups, our model can generate images with a smooth increase of age, which shows superiority of model even on continuous face aging.

2.4.8 Gender Translation Combined with Aging

Combining gender information with age group label as the final condition can help the model to distinguish aging attributes from gender features. In addition, integrating gender information into target condition enables our model to perform aging and



(a) On MORPH (b) On CACD

Figure 10: Gender conversion combined with face aging (a)on MORPH and (b) on CACD. The first column is the test face. The remaining four columns are generated images in age group 21-25 and 31-35 with the same or opposite gender of the test face.

gender transformation simultaneously. As shown in Fig. 10, gender attributes are well separated from face aging and transformed correctly according to the gender label.

2.4.9 Face aging for high resolution images

Our approach can be easily extended for image with high resolution. For images with a size of 256×256, we just need to increase one down-sampling convolutional layer (K4, S2, P1) to each of the multi-level discriminators. We train the extensive model on MORPH, the results are shown in Fig. 11.

2.4.10 Limits of This Work

While our work can generate vivid aging effects for young (30-) and adult age group(31-40, 41-50) with high identity verification rate, there are many failure cases in age group 51+ of CACD to produce natural aging hair (as shown in Figure 12(left)). This may be caused by the limited aged examples with obvious aging hair in the age group 51+ of CACD. Moreover, the diversity of hair colors (black, blond, brown and gray) in the young and adult age groups of CACD makes the simulating of aging hair harder for age group 51+. Our model also shows slight change of aging effects when the

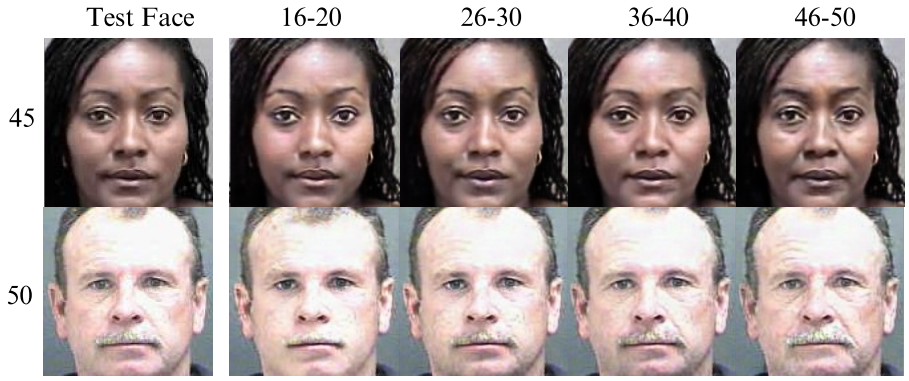


Figure 11: Sampled results of face aging on MORPH with a resolution of 256. The first column contains the input faces, followed by synthesized images in age group 16-20, 26-30, 36-40 and 46-50.

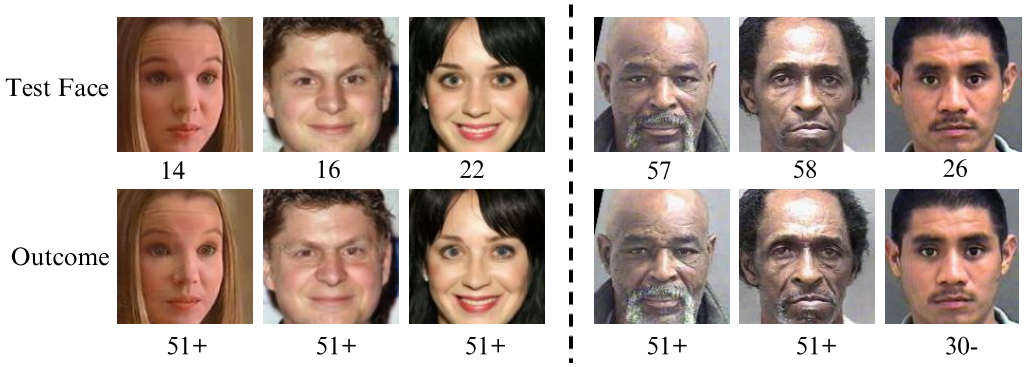


Figure 12: Failure cases. (1)The left columns show the failure of generating aging hair for age group 51+ on CACD. (2) Right columns show subtle change of aging effects when predicting age group same as the original age group. The first row contains the input faces, followed by synthesized images in the second row.

target age group is same as the source age group. As shown in Figure 12(right), the first two examples display much denser white beard on synthetic faces belonging to age group 51+ than the test faces which are also in age group 51+. While the third example shows that the generated young face on age group 30- has denser black hair than the test face with an age of 26.

In the next stage, we will consider of adopting relative condition of age group and a self reconstruction loss to guarantee that the model can take into account the original age of the inputs and synthesize the same images as inputs when the target age group is same as the original age group. To prevent the negligence of aging hair synthesis for the aged group caused by the unbalanced distribution of gray hair in the current database, we will add hair colour information (black, blond, brown and gray) generated by a classifier learned from CelebA [45]. Likewise, We can add more other attributes such as baldness and beard information to the generative model. During inference for aged faces, the intensity of aging hair can be manipulated by setting an interpolate coefficient between original hair color condition and gray hair condition.

2.5 Conclusion

In this work, we consider the process of face aging as a multi-domain image-to-image translation and propose wavelet-based multi-level Generative Adversarial Networks for both age progression and regression. The wavelet packet transform based image decomposition into multi-level coefficients in the frequency space is integrated into the multi-level features extraction of both the multi-level generator and three discriminators.

Extensive experiments show that the wavelet-based multi-level generator can effectively preserve the identity of input images. Combined with the wavelet-based multi-level discriminators, our model outperforms the state-of-the-art models in all or most of the age groups in aging accuracy and identity preserving effects on both MORPH and CACD. Moreover, our approach reduces the time for training significantly by getting rid of an identity preserving module. Finally, with interpolated condition between discrete age groups, our model can perform continuous face aging.

Chapter 3

WP2-GAN: Wavelet-based Multi-level GAN for Progressive Facial Expression Translation with Parallel Generators [59]

3.1 Introduction

Recently, expression synthesis has attracted much attention from the community of computer vision because of its wide applications to photography technologies, human-computer-interaction and animation movies. However, facial expression manipulation is challenging owing to the non-linear facial geometric variation caused by different expressions.

Although difficult it is, expression translation has achieved great progress due to the rapid development of deep neural networks. Especially, the advent and development of Generative Adversarial Networks (GANs) [2, 24, 28, 54, 78] have opened a new door to the face manipulating technologies [7, 10, 44, 52, 71]. The advent of Condition GAN (cGAN) [46] and Cycle-GAN [79] made the attributes editing on the same subject possible without paired images belonging to the same subject. Many recent models [7, 52, 62] applied the principle of cGAN and Cycle-GAN for facial expression translation. Specifically, one generator is called twice to perform expression translation and reconstruction by conditioning on different expression domain

(i.e. expression label or Action Units (AUs) code [20]). However, this manner will force the generator to leave an unseen "noise" to the generated image for a convenient reconstruction in the second step. Based on the facial attributes editing task performed by StarGAN [7], Sanchez et al. [56] found that the second translation of the generator based on the outcome of the first translation will produce results almost the same as the input images no matter what conditions were adopted. The footprint left in the outcomes hampered the reuse of these images for further task. We infer the interference may be caused by the tight linkage between the forward prediction and backward reconstruction by using the same generator, resulting in a defective generator leaving a footprint in the outcome. To eliminate the unwanted interference, we propose a parallel training system consisting of two generators with equal importance. The generators are trained simultaneously for the same forward prediction but then act as the reconstruction model for each other. Our method can break the unwanted link between the first and second translation (as shown in Figure 14).

An intuitive application of our unbound generators is to equip them for progressive training. Previous end-to-end models for expression editing usually generate artifacts or blurs around the expression-rich areas such as the forehead, eyes and mouth. Inspired by the successful application of progressive training in geometric conversion [41, 70], we propose a novel progressive training framework based on our parallel training scheme.

Besides efficient geometric translation, identity preserving with fine-grained facial features is another important task of facial expression editing. Recent research [44] showed that multi-level discriminators integrated with wavelet-based information decomposition can help to extract features related to identity and age for face aging. Considering facial expression translation also involves identity preserving and the synthesis of local expression-related features such as forehead wrinkles and smiling lines, it is intuitive to apply the wavelet-based multi-level discriminators to facial expression translation.

In this work, we propose a novel WP2-GAN for continuous expression translation. The model consists of two parallel generators and a set of wavelet-based multi-level discriminators. All the modules are trained and updated progressively hence we can effectively reduce the computing resource for model training. We adopt an attention mechanism like [52] to each of the generators so that two generators can mainly

focus on the active areas for expression conversion. To maintain the background information of the input image after several progressive translations, we take the original image as the source to calculate the background information of the generated image. Wavelet-based multi-level discriminators are employed to extract expression-related features at multiple scales from the given images, enforcing the generators to synthesize photo-realistic images with vivid expressions.

Our main contribution is to introduce two parallel generators to the facial expression translation task and to eliminate the interference existed in previous methods that is caused by using one single generator for both forward and backward translation. Additionally, we design a novel progressive training strategy based on the parallel generators, combined with wavelet-based multi-level discriminators to improve the quality of expression translation. Extensive experiments illustrate the effectiveness of our method for expression translation with a large gap.

3.2 RELATED WORKS

3.2.1 GAN

Generative Adversarial Networks (GANs) [24] were first proposed to generate images based on minimax game theory, then were improved by many other works [2, 25, 28]. Later, Mirza and Osindero [46] proposed a conditional GAN (cGAN) that embeds prior information into image generation. Cycle-GAN [79] was proposed to perform image-to-image translation without paired images through a cycle-consistent loss. Soon after, models combined with cGAN and Cycle-GAN were widely applied to cross-domain translation [7, 52, 62, 73]. Most of these works only adopt one generator for target features translation and then the reconstruction of the input image. Sanchez et al. [56] mentioned that using one generator for both prediction and reconstruction would leave a "noise" to the outcome, therefore hindering the further application of the generated images. The authors proposed a recurrent cycle-consistency loss to replace the original loss. However, their approach needs paired images with the same identity, thus loses the advantage of Cycle-GAN for unpaired images translation. In this work, we propose to use two parallel generators to conduct the forward translation but served as the reconstruction model for each other. Empirical experiments show that our method can overcome the drawback of previous methods (shown in

Figure 14).

3.2.2 Facial Expression Translation

Current methods for facial expression translation can be generally categorized into two classes. The first class resorts to a 3D model for expression editing. Blanz and Vetter [4] proposed the first 3D Morphable model for 3D face reconstruction. Vlasic et al. [65] presented a multilinear model of 3D face meshes for expression translation. Cao et al. [5] introduced a method for facial image animation based on the 3D face mesh. Geng et al. [23] proposed a 3D-guided generative model for continuous expressions editing. paGAN [47] can perform fine-grained expression translation by conditioning on multiple conditions such as the desired blendshape expression and viewpoint generated by a 3D fitting model. Facial expression translation methods using a 3D model usually require efforts for complex parametric fitting, thus are computing resource demanding.

The second category of methods for facial expression synthesis leverages deep generative models. Many previous works [22, 53, 61] performed discrete or continuous facial expression by conditioning on facial landmarks. ExprGAN [10] can control the intensity of expression by conditioning on an embedding generated from expression labels. LEED [69] realized label-free expression translation by disentangling the expression-related features from identity. But a pre-trained GAN for neutral expression synthesis is still needed to extract the identity related features. StarGAN [7] achieved multi-task translation among different domains with one model. But this model can only generate limited and discrete emotion expressions. Pumarola et al. [52] proposed GANimation with an attention mechanism to predict continuous expression translation by conditioning on AUs [20]. However, this model still generates some artifacts for expression translations with a large gap. Many other works [49, 62] leverage multi-level discriminators to extract expression-related features during model training.

Different from [22, 53, 61], our approach can perform continuous expression editing by conditioning on AUs code which can be extracted by Openface [3] conveniently. Unlike [7, 49, 51, 62], which only utilize one generator for both forward prediction and reconstruction, our method adopts two parallel generators to alleviate the interference mentioned by [56]. Besides using multi-level discriminators like [62], we integrated the

wavelet-based image decomposition at multiple scales in frequency space to promote the expression-related features extraction in discriminators.

The most recent work that also adopted a progressive training strategy for expression editing is Cascade EF-GAN [70]. Different from Cascade EF-GAN that adopts three local sub-networks to synthesize three local patches (i.e. eyes, nose and mouth) and one global network to predict a whole face, our approach leverages wavelet-based multi-level discriminators to extract multi-level facial features automatically without physical concatenation. Furthermore, we design a new progressive training method based on two parallel generators, that can be updated gradually instead of stacking all well-trained modules and optimizing them at one time. Hence our strategy can simplify model training and reduce the computing memory needed.

3.3 PROPOSED METHOD

3.3.1 Problem Formulation

Let X and Y represent the source facial image and expression domains, respectively. Given an original face $x_o \in X$ with an expression $y_o \in Y$ and a different target expression $y_g \in Y$, our goal is to learn a transformation that can generate the facial image $x_g \in X$ with the same identity as x_o but with the desired expression y_g.

As we mainly consider the problem of continuous expression translation, the continuous Action Units (AUs) intensity [20] is adopted as AUs code, which can be extracted by OpenFace [3]. Given the AUs code of a target expression, we can obtain the intermediate condition for progressive training according to the interpolation formula: $y_t = y_o + \alpha * (y_g - y_o)$, where $\alpha \in \{0.3, 0.6, 1.0\}$, is a hyper-parameter to control the intensity of each step of progressive training. An overview of our architecture is given in Figure 13.

3.3.2 Parallel and Progressive Training Mechanism

Motivated by the discovery in [56], we design a parallel training mechanism for the AUs code conditioned expression translation. During the training process, two generator G_A and G_B both take as input the original image x_o and the target condition y_g to synthesize the image x_A and x_B, respectively. Then, the generator G_B (G_A) takes

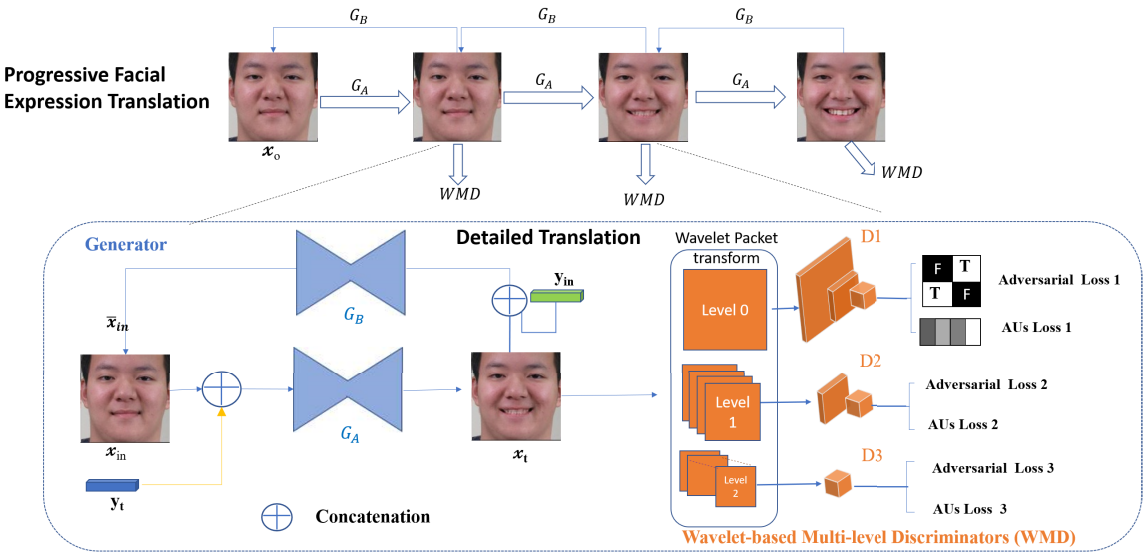


Figure 13: An overview of the WP2-GAN framework. The workflow of the progressive training is shown on the top, while the details of each step are shown in the zoom-in area. As two generators perform a similar forward translation and work as reconstruction model for each other, we only show one stream of the translations. In each progressive step, one generator G_A takes as input the image x_{in} and the target expression y_t to synthesize the image x_t. Then the other generator G_B works as a reconstruction model to restore the input image x_{in}. A cycle-consistent loss is calculated by comparing x_{in} with $\bar{x}_{in}$ to preserve the identity of the input image. A similarity loss imposed on the outcomes of two forward translations is adopted to force two generators proceed in the same direction. Wavelet-based multi-level discriminators(WMD) take as input different levels of wavelet coefficients generated from the synthetic image x_t or the original image x_o and evaluate the realism of given images as well as the AUs code translation accuracy.

as input x_A (x_B) and the original expression y_o to reconstruct the original image. As each generator leverages the outcome of another generator to reconstruct the input image, it removes the potential "short-cut" in the model to memorize the input image for the second translation. Our generators are auto-encoder based networks adopted from [52]. The structure of the generator is shown in Table 8. It is worth noting that there are InstanceNorm and ReLU layers between the convolutional layers in the generator. A Tanh layer is posed on the outcome of ConVO1 to generate the color map, while a Sigmoid layer is leveraged to produce mask map from the outcome of ConVO2.

Inspired by the impressive success of progressive training methods [41, 70] in

Table 8: Neural network architecture of generator (G_A/G_B) ('K','S' and 'P' denote 'Kernel size','stride' and 'padding' of convolutional layers).

	Layer	Output Size	Details
Input	ConvI	$128 \times 128 \times 64$	K7, S1, P3
Down-sampling	ConVD1	$64 \times 64 \times 128$	K4, S2, P1
	ConVD2	$32 \times 32 \times 256$	K4, S2, P1
	ConVD3	$16 \times 16 \times 512$	K4, S2, P1
Residual Blocks	ConvR1-1	$16 \times 16 \times 512$	K3, S1, P1
	ConvR1-2	$16 \times 16 \times 512$	K3, S1, P1
	...	...	...
	ConvR6-1	$16 \times 16 \times 512$	K3, S1, P1
	ConvR6-2	$16 \times 16 \times 512$	K3, S1, P1
Up-sampling	ConvU1	$32 \times 32 \times 256$	K4, S2, P1
	ConvU2	$64 \times 64 \times 128$	K4, S2, P1
	ConvU3	$128 \times 128 \times 64$	K4, S2, P1
Output	ConVO1	$128 \times 128 \times 3$	K7, S1, P3
	ConVO2	$128 \times 128 \times 1$	K7, S1, P3

geometric conversion, we design a novel progressive learning strategy for our task based on the parallel training mechanism. As shown in Figure 13, we decompose the previous end-to-end translation into three progressive steps. Especially, in each progressive training step, the forward generator G_A takes as input the interpolated condition y_t and the image x_{in}, which can be the original image x_o or an intermediate result of last step of translation. The condition y_{in} corresponding to image x_{in} can be the original expression y_o or an interpolated condition.

Different from [70], our progressive training is based on two parallel generators instead of one single generator. Thus we can avoid the accumulation of interference as mentioned before. Besides, our approach does not stack multiple pre-trained generators together and update all networks at final step but trains and updates the neural networks by each progressive step, thus reducing the computing memories needed. The work with the most similar idea to ours is [41]. But it is designed for unsupervised image-to-image translation instead of semi-supervised facial attributes editing. The intermediate results of progressive translation are shown in Figure 15.

Similar to [52, 70], a visual attention mechanism is applied to the generators, enforcing the network to only focus on the active facial area rather than the periphery. To overcome the gradual loss of background information during progressive translation, we leverage the original input to compute the background information of the synthetic image in each progressive step. The image can be calculated by:

$$x_t = M_A \otimes x_o + (1 - M_A) \otimes M_C, \tag{7}$$

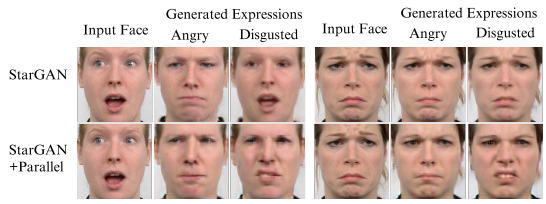


Figure 14: Comparison between StarGAN and modified StarGAN with parallel generators for expression translation. Each triplet contains the input face in the first column followed by outcome of the first translation (angry face) in the middle and then the result of second translation (disgusted face) based on the first outcome.

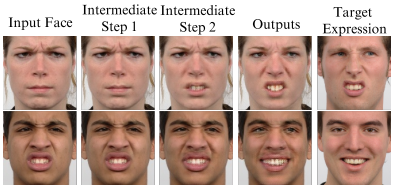


Figure 15: Display of progressive translation results by our WP2-GAN. The first column contains the input faces followed by two intermediate results and then the final outcomes. The last column shows images with target expressions. The progressive model provides a gradual transformation between expressions with a large gap.

where M_A and M_C denote the attention map and color map generated by the generator from the original or intermediate input. x_o represents the original input image instead of the intermediate input.$\otimes$ indicates the element-wise multiplication. This strategy enables the progressive model to preserve background and face pixel information located in inactive areas.

3.3.3 Wavelet-based Multi-level Discriminators

Recently, wavelet-based multi-level discriminators have been successfully applied to face aging [44**?**]. Wavelet Packet Transform (WPT) can decompose an image into multi-level wavelet coefficients which contain both texture and geometric information [44]. Considering expression translation involves changes in both shapes and texture, image decomposition at multiple scales by WPT could promote the performance of the system.

In this work, we adopt three levels of discriminators which have a gradually decreasing number of convolutional layers so that three levels of wavelet coefficients can be encoded into three matrices with the same size $Y_{D_i} \in R^{H/2^6 \times W/2^6}$, where $i = \{1, 2, 3\}$, H and W are the height and width of the input image. Each element of Y_{Di} represents the probability of the corresponding patch to be real. We do not concatenate the outcomes of three critics as one tensor as [44] did. Empirical studies show that separate discriminators do not cost much time than combined ones but can

stabilize the training process. As we adopt WGAN-GP [25] for stabilized adversarial training, a penalty loss is added to the gradient norm of each critic.

Besides photo-realism, three discriminators are also responsible for estimating the AUs code. To reduce the number of parameters, we add one regression layer for AUs regression, on the last second layer of each discriminator.

Table 9: Neural network architecture of multi-level discriminators ('K','S' and 'P' denote 'Kernel size', 'stride' and 'padding' of convolutional layers).

	Layer	Output Size			Details
		D1	D2	D3	
Input	ConvI	$64 \times 64 \times 64$	$32 \times 32 \times 64$	$16 \times 16 \times 128$	K4, S2, P1
Down-sampling	ConVD1	$32 \times 32 \times 128$	$16 \times 16 \times 128$	$8 \times 8 \times 128$	K4, S2, P1
	ConVD2	$16 \times 16 \times 256$	$8 \times 8 \times 256$	$4 \times 4 \times 256$	K4, S2, P1
	ConVD3	$8 \times 8 \times 512$	$4 \times 4 \times 512$	$2 \times 2 \times 512$	K4, S2, P1
	ConVD4	$4 \times 4 \times 1024$	$2 \times 2 \times 1024$	—	K4, S2, P1
	ConVD5	$2 \times 2 \times 2048$	—	—	K4, S2, P1
Output	ConvO-Adv	$2 \times 2 \times 1$	$2 \times 2 \times 1$	$2 \times 2 \times 1$	K3, S1, P1
	ConvO-Aus	$1 \times 1 \times 17$	$1 \times 1 \times 17$	$1 \times 1 \times 17$	K2, S1, P0

The structure of three discriminators are shown in Table 9. It is worth noting that there are InstanceNorm and LeakyReLU layers between the convolutional layers in the discriminators.

3.3.4 Loss Functions

To make the generated images indistinguishable from real images, we adopt the adversarial loss proposed by WGAN-GP [25] for each step of progressive training, which is defined as:

$$\mathcal{L}_{avd} = \frac{1}{3}\sum_{i=1}^{3}\left\{\frac{1}{2}\mathbb{E}_{\mathbf{x}_{in}\sim\bar{\mathbb{P}}}\left[D_i\left(G_A\right)+D_i\left(G_B\right)\right]-\mathbb{E}_{\mathbf{x}_o\sim\mathbb{P}}\left[D_i\left(\mathbf{x}_o\right)\right]+\lambda_{gp}\mathbb{E}_{\widetilde{\mathbf{x}}\sim\widetilde{\mathbb{P}}}\left(\parallel\nabla_{\widetilde{\mathbf{x}}}D_i\left(\widetilde{\mathbf{x}}\right)\parallel_2-1\right)^2\right\}, \tag{8}$$

where $G_m = G_m\left(\mathbf{x}_{in}|\mathbf{y}_t\right), m \in \{A, B\}$. $\mathbb{P}$ is the data distribution of original image $\mathbf{x}_0$, $\widetilde{\mathbf{x}}$ represents the random interpolated image by input image $\mathbf{x}_0$ and the generated face $\mathbf{x}_t$, $\widetilde{\mathbb{P}}$ stands for the uniform interpolation distribution, $\bar{\mathbb{P}}$ indicates the union of $\mathbb{P}$ and $\widetilde{\mathbb{P}}$, while λ_{gp} is a penalty coefficient.

According to [52], the attention mask A is easy to saturate to 1, making $G\left(\mathbf{x}_{in}|\mathbf{y}_t\right) = \mathbf{x}_{in}$. To prevent the saturation, l_2-weight penalty is added to the attention mask. A

Total Variation Regularization is imposed on A to enforce the generated images to be smooth. The attention loss can be defined as:

$$\mathcal{L}_A(G, \mathbf{x}_{in}, \mathbf{y}_t) = \frac{1}{2} \sum_{M \in \{M_A, M_B\}} \Big\{ \mathbb{E}_{\mathbf{x}_{in} \sim \bar{\mathbb{P}}} \left[\| M \|_2 \right] + \lambda_{TV} \mathbb{E}_{\mathbf{x}_{in} \sim \bar{\mathbb{P}}} \left[\sum_{i,j}^{H,W} (M_{i+1,j} - M_{i,j})^2 + (M_{i,j+1} - M_{i,j})^2 \right] \Big\}, \tag{9}$$

where M_A and M_B are masks generated by generator G_A and G_B, respectively. λ_{TV} is the penalty coefficient for mask smoothing.

Besides the adversarial loss and the attention loss, the generator and the discriminator also have to reduce the errors produced by the regression layer imposed on top of each critic. The whole condition losses can be written as:

$$\mathcal{L}_{cond} = \frac{1}{3} \sum_{i=1}^{3} \Big\{ \frac{1}{2} \sum_{m \in \{A,B\}} \mathbb{E}_{\mathbf{x}_{in} \sim \bar{\mathbb{P}}} \left[\| D_i(G_m) - \mathbf{y}_t \|_2^2 \right] + \mathbb{E}_{\mathbf{x}_o \sim \mathbb{P}} \left[\| D_i(\mathbf{x}_o) - \mathbf{y}_o \|_2^2 \right] \Big\}, \tag{10}$$

where $G_m = G_m\left(\mathbf{x}_{in} | \mathbf{y}_t\right), m \in \{A, B\}$ and *cond* is AUs code. $\mathbf{y}_t$ and $\mathbf{y}_{in}$ are target and input condition (expression).

To ensure two generators G_A and G_B proceed in the same direction, we impose a similarity loss to the forward outcomes of two generators. The similarity loss is formulated as:

$$\mathcal{L}_{sim} = \mathbb{E}_{\mathbf{x}_{in} \sim \bar{\mathbb{P}}} \left[\| G_A(\mathbf{x}_{in} | \mathbf{y}_t) - G_B(\mathbf{x}_{in} | \mathbf{y}_t) \|_1 \right], \tag{11}$$

Minimizing the loss functions above does not guarantee that the generated images keep the same identity with their input counterparts. A cycle-consistent loss [79] is utilized by the generators to preserve the identity-level consistency. We adopt the l_1 norm, which helps to capture features associated with low-frequencies. The cycle-consistent loss is formulated as:

$$\mathcal{L}_{cyc} = \frac{1}{2} \mathbb{E}_{\mathbf{x}_{in} \sim \bar{\mathbb{P}}} \Big[\| G_B(G_A(\mathbf{x}_{in} | \mathbf{y}_t) | \mathbf{y}_{in}) - \mathbf{x}_{in} \|_1 + \| G_A(G_B(\mathbf{x}_{in} | \mathbf{y}_t) | \mathbf{y}_{in}) - \mathbf{x}_{in} \|_1 \Big], \tag{12}$$

The overall loss functions for G and D can be formulated as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{cond} \mathcal{L}_{cond} + \lambda_{sim} \mathcal{L}_{sim} + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_A (\mathcal{L}_A(G, \mathbf{x}_{in}, \mathbf{y}_t) + \mathcal{L}_A(G, \mathbf{x}_t, \mathbf{y}_{in})), \tag{13}$$

where λ_{cond}, λ_{sim}, λ_{cyc} and λ_A are hyper-parameters that control the relative importance of conditional loss, similarity loss, cycle-consistent loss and attention loss, respectively.

3.4 EXPERIMENTS

3.4.1 Dataset

We train and test our model WP2-GAN on two public facial expression databases: RafD [32] and Compound Facial Expressions of Emotions Dataset (CFEED) [11]. RafD consists of 8,040 images of 73 subjects collected from different angles. We only adopt frontal images and collect 1,608 images for our experiments. CFEED consists of 5,060 compound expression images of 230 subjects. We randomly select 9/10 images of each database above for model training and the remaining for model testing.

In our experiments, all images are aligned, cropped and resized to the size of 128×128 by Openface [3]. We also leverage Openface to extract the AUs code for every image.

3.4.2 Implementation Details

Our approach adopts three steps of progressive training. We utilize Adam [31] for the model optimization with the following hyper-parameters: learning rate=0.00005, beta1=0.5, beta2 =0.999 and batch size=25. During progressive training, the generators and discriminators are optimized at the same frequency. The weight coefficients for the loss functions are set to $\lambda_A = 0.2$, $\lambda_{gp} = 10$, $\lambda_{cond} = 160$, $\lambda_{sim} = 1$ and $\lambda_{cyc} = 10$.

Our model is trained on RafD [32] and CFEED [11] for 200 epochs, respectively. The learning rate (lr) is linearly decayed to zero over the last 50 epochs of the training.

For the parallel only training experiments, the generator is optimized once after five times optimization of the discriminators, with an initial learning rate of 0.0001. The weight coefficient that is different from the progressive training is $\lambda_{sim} = 5$.

Our progressive training model is trained on RafD and CFEED for about 13h and 40h, respectively, with a single Tesla V100 GPU.

3.4.3 Qualitative Experimental Results

In this section, we test our approach on both RafD and CFEED and compare the results with three previous models: GANimation [51], UNet-MFS [40] and Cascade EF-GAN [70], all of which are conditioned on AUs code for continuous expression

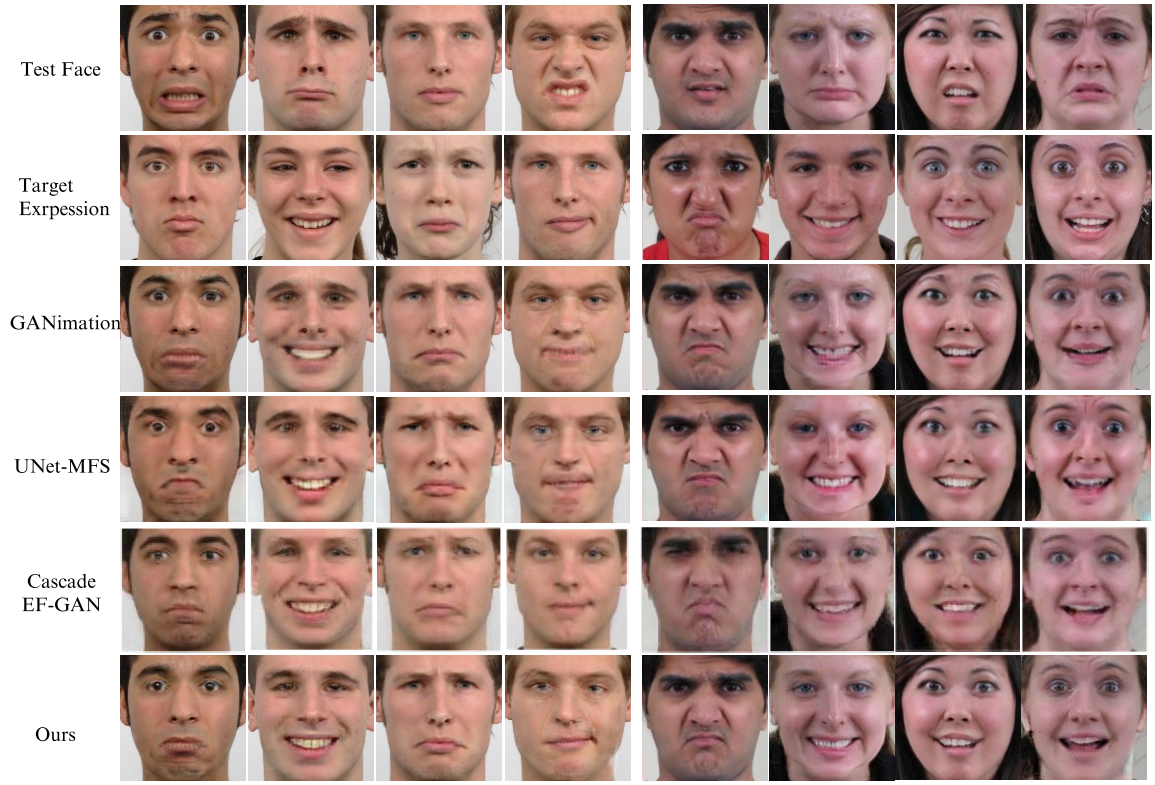


Figure 16: Qualitative comparison with previous works on RafD (left four columns) and CFEED (right four columns).

translation. We leverage the code issued publicly on Github and train GANimation and UNet-MFS with the same training set as described above. We obtain the results from [70] for Cascade EF-GAN due to the unavailability of the code.

As shown in Figure 16, GANimation and UNet-MFS generate results with obvious artifcats on test samples of both RafD and CFEED, especially on area of mouth. Although Cascade EF-GAN generates natural outcomes with much less artifacts, the results are a little blurring. In contrast, our method can vividly simulate the target expressions and generate photo-realistic images with high-fidelity, showing the superiority of our method for expression translation with obvious geometric deformation.

Table 10: Quantitative comparison among GANimation, Unet-MFS, Cascade EF-GAN and all variants of the proposed model.

	RafD			CFEED		
Method	Accuracy↑	FID↓	SSIM↑	Accuracy↑	FID↓	SSIM↑
GANimation	85.36%	45.34	0.6646	77.46%	25.83	0.6507
UNet-MFS	88.36%	56.44	**0.6905**	84.39%	28.48	**0.6769**
Cascade EF-GAN	89.38%	42.36	–	85.81%	27.15	–
Ours (WP2-GAN)	**89.47%**	**41.74**	0.6818	**87.97%**	**24.91**	0.6659
Parallel-GAN	87.31%	46.74	0.6590	76.67%	29.55	0.6467
P2-GAN	89.00%	43.44	0.6770	85.46%	23.75	0.6579
WP-GAN	87.71%	46.65	0.6753	85.52%	25.72	0.6619

3.4.4 Quantitative Experimental Results

We adopt a similar method of Cascade EF-GAN [70] and StarGAN [7] to evaluate the expression translation accuracy of our model. Particularly, we train different models on the training sets of RafD and CFEED and test them on the unseen test sets. We then train an expression classifier (Resnet-18 [26]) on the filtered training set of each database, which only contains images with basic expression (i.e. angry, disgust, fearful, happy, sad, surprised or neutral). We obtain two classifiers with a test accuracy of 100% on RafD and 88.67% on CFEED, respectively. Finally, we evaluate the performance of our models for basic expression translation by classifying the generated images with the classifier. Higher expression recognition accuracy represents higher expression translation accuracy of models.

The quantitative comparison among GANimation, UNet-MFS, Cascade EF-GAN and variants of our method is displayed in Table 10. The results of Cascade EF-GAN are from [70]. We can observe that our approach obtains the highest expression translation accuracy compared with three previous models on both RafD and CFEED. The proposed model was trained progressively thus can overcome the drawback of limited training data and significantly exceed the baseline GANimation in terms of expression translation accuracy by 4.11% on RafD and 10.51% on CFEED, respectively. Our method also outperforms UNet-MFS and Cascade EF-GAN by 3.58%/2.16% on CFEED and slightly on RafD, showing the superiority of our method in expression translation accuracy.

We further evaluate the image quality in terms of and Fréchet Inception Distance (FID) [27] and structural similarity (SSIM) index [67]. A lower FID score and a higher SSIM normally represent a higher image quality.

As shown in Table 10, our method achieves the lowest FID scores on two databases

compared with three baselines, even outperforms the latest state-of-the-art model (Cascade EF-GAN) by 0.62 on RafD and 2.24 on CFEED, demonstrating the advantage of the proposed model. Our method also exceeds two baselines in terms of SSIM on two databases. Although UNet-MFS achieves slightly higher SSIM scores than our method, we can infer that our model can predict expressions with higher quality considering the qualitative results shown in Figure 16 as well as FID scores in Table 10,

Higher expression translation accuracy and image quality achieved by our model demonstrate the superiority of our approach in expression translation with a large gap.

3.4.5 Ablation Study

In this section, we study the contributions of each component of our proposed model and compare the expression translation effects on both RafD and CFEED among variants of the proposed model. The baseline we compare in this section is GANimation. Parallel-GAN means the model with two generators and is trained in a parallel method, while P2-GAN denotes the model trained in a parallel and progressive way. Compared to P2-GAN, we introduce the wavelet-based multi-level discriminators to our final model. Compared to WP2-GAN, WP-GAN only has one generator. The single generator works as the reconstruction model for itself in each progressive step of training.

We can observe in Figure 17 that both the baseline and Parallel-GAN fail to produce natural expressions but generating some artifacts in areas near mouth and eyes. The introduction of progressive training enables the model P2-GAN to generate vivider results but still with some artifacts. In contrast, our proposed model can generate more realistic images with high-fidelity such as much clearer teeth. In our approach, the utilization of wavelet-based multi-level discriminators can further help the model to capture expression-related features, thus generating photo-realistic images.

Although WP-GAN with a single generator can generate natural facial expression with less artifacts, the outcomes of WP2-GAN are better matched with the target expressions. For example, WP2-GAN produces more obvious contemptuous expression than WP-GAN on the first sample of RafD and much clearer teeth on the second

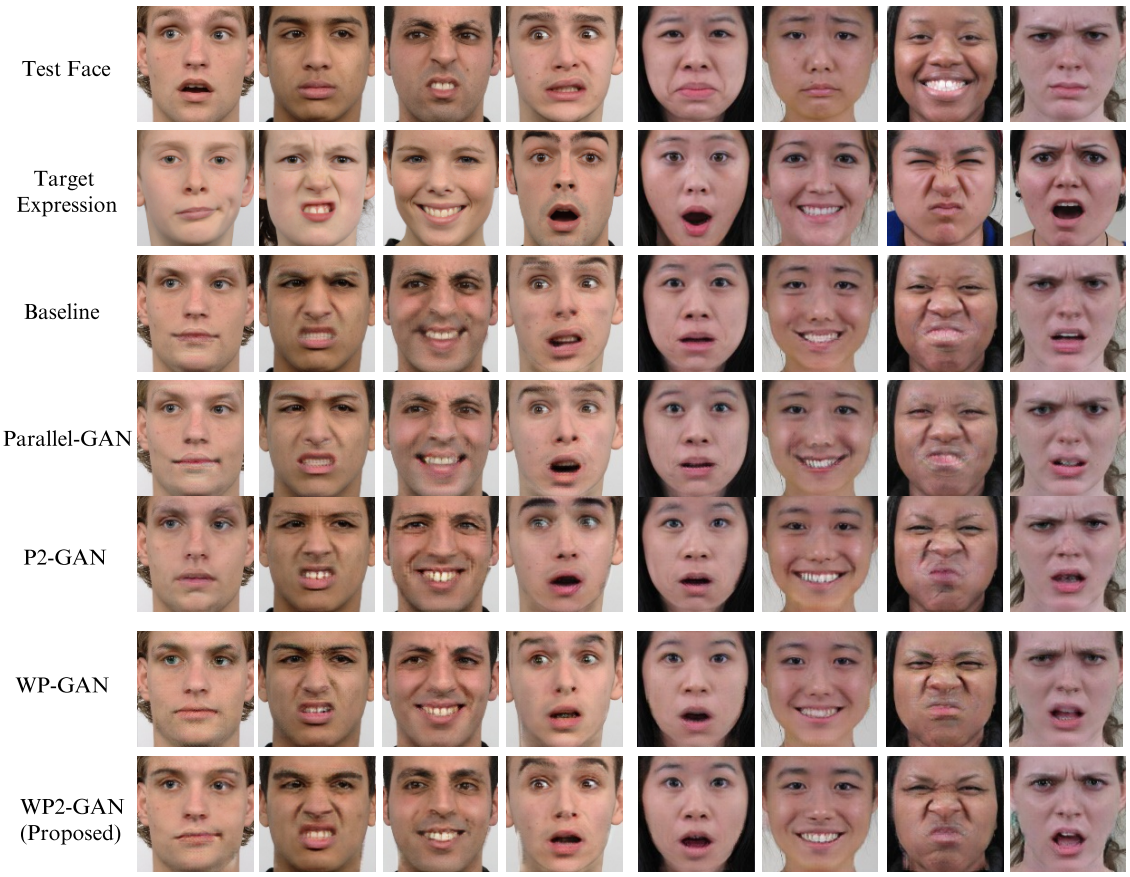


Figure 17: Comparison of expression translation between the proposal and variants of the proposed model on both RafD (left four columns) and CFEED (right four columns).

samples of two databases. This further demonstrates the contribution of two parallel generators adopted in our approach.

We also perform the quantitative comparison between the variants and our proposed model. Table 10 shows that our proposed model achieves the best performance among its variants. Specially, the proposed model outperforms WP-GAN on RafD and CFEED by 1.76%/2.45% in expression translation accuracy and 4.91/0.81 in FID score, further illustrating the significance of parallel training in our model. However, compared with the baseline model, parallel training alone (Parallel-GAN) does not cause an obvious improvement of the translation accuracy but a decline of image quality. This could be caused by the loss of the constraint (using the same generator

for the forward and backward translation) imposed on the previous single generator. However, progressive training and wavelet-based multi-level discriminators equipped in our method impose extra constrains on the adversarial learning system, enforcing the model to proceed in a desired path.

3.4.6 Extensional Experiments

Our proposed model can be easily extended for continuous expression translation. Given the AUs code of a target expression, we can obtain the intermediate AUs code by a similar interpolation formula as that of the progressive training. Then, we use the intermediate AUs code as the target label of the progressive training. Our results for continuous expression translation are shown in Figure 18.

Figure 18: Continuous expression translation performed by our proposed model on both RafD (top) and CFEED (bottom). The first column contains the input images, followed by generated images with a continuous change of expression.

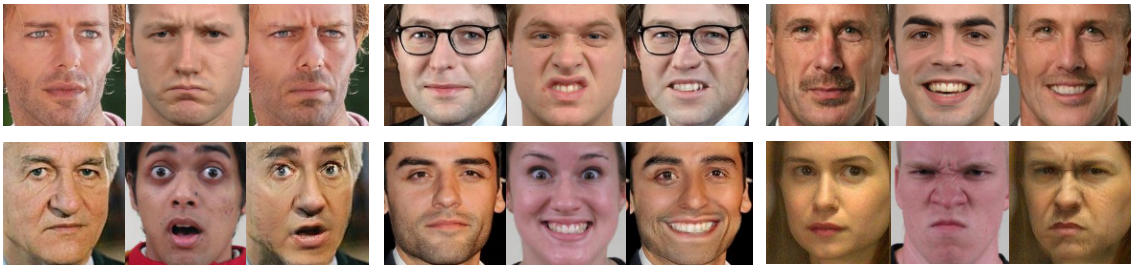


Figure 19: Sampled expression translation results by our proposed model on EmotioNet [17]. Each triplet contains the test face, the target expression and finally the synthesized image.

We also evaluate our method on images in the wild. We train the model on over 70,000 images from EmotiNet [17] then fine-tune the model on RafD and CFEED.

Figure 19 shows that our approach can be applied to images with different background in the wild.

We extensively test the impact of using immediate result instead of original input as resource to compute background information of the translation. Results show that this practice leads to the drop of both translation accuracy (84.30%) and image quality (FID: 31.02) on CFEED.

Finally, we explore the impact of steps in progressive training and train a model on CFEED adopting four steps. The results show that further increase of progressive training steps fails to bring an improvement to both translation accuracy and image quality (Accuracy: 87.23%, FID: 24.62, SSIM: 0.6753). Thus, we adopt three steps for our proposed method.

3.5 Conclusion

In this work, we consider facial expression editing as an image-to-image translation task and propose a novel wavelet-based multi-level generative network for progressive facial expression transformation. Our model consists of two generators that are trained in a parallel way to alleviate the interference caused by using the same generator for image reconstruction. Progressive training breaks the translation between large-gap expressions into several small steps, making the model robust to the synthesis of extreme expressions. Wavelet-based multi-level discriminators enforce the generators to generate high-quality images by extracting expression and identity-related facial features at multiple scales. Extensive experiments demonstrate the superiority of our approach for expression translation compared to the start-of-the-art models. Our method can synthesize photo-realistic images with vivid expression.

3.6 More Experimental Results

More results of expression translation by our proposed model on EmotioNet, RafD and CFEED are shown in Figure 20$\sim$ Figure 22.

Figure 20: Supplementary expression translation results by our proposed model on EmotioNet. In each triplet, the first column is the test face, followed by an image with the target expression and finally the synthesized image.

Figure 21: Supplementary results of expression translation by our proposed model on RafD (left) and CFEED (right). In each triplet, the first column is the test face, followed by an image with the target expression and finally the synthesized image.

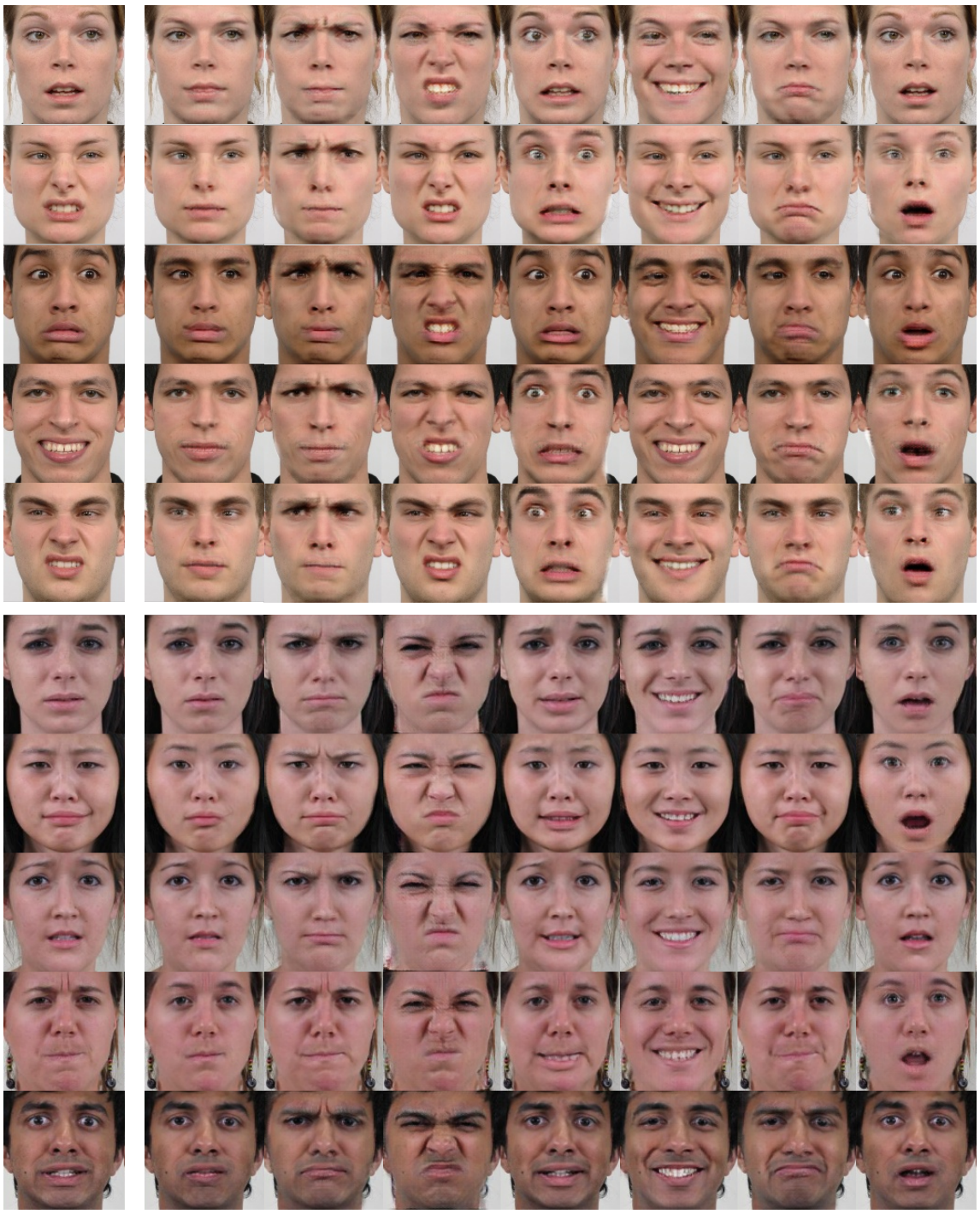


Figure 22: Supplementary results of seven basic expressions synthesized by our proposed model (Input, Neutral, Angry, Disgusted, Fearful, Happy, Sad, Surprised) on RafD (top) and CFEED (bottom).

Chapter 4

Summary and Future work

In this chapter, we will give conclusions of our thesis work and ideas for future work.

4.1 Conclusions

In this thesis, we proposed the wavelet-based multi-level GANs for face aging and the WP2-GAN for progressive facial expression translation.

For face aging task, the Wavelet Packet Transform module and multi-level encoders are applied to the generator of GANs for the first time. Application of multi-level generator combined with wavelet transform decomposition can improve the identity verification confidence in face aging, and significantly reduce the time for model training by eliminating the use of an identity preserving module. Extensive experiments demonstrate the superiority and effectiveness of our method by synthesizing vivid aging effects and outperforming the existing state-of-the-art models in both face aging accuracy and identity verification confidence.

In the work related to expression translation, two parallel training generators were introduced to the task to alleviate the interference caused by using the same generator for image reconstruction. Progressive training breaks the translation between large-gap expressions into several small steps, making the model robust to the synthesis of extreme expressions. Wavelet-based multi-level discriminators enforce the generators to generate high-quality images by extracting expression and identity-related facial features at multiple scales. Extensive experiments demonstrate the superiority of our approach for expression translation compared to the start-of-the-art models.

4.2 Future Work

For the face aging model, according to the limits disclosed before, we can consider of adopting relative condition of age group and a self reconstruction loss to guarantee that the model can synthesize the same images as inputs when the target age group is the same as the original age group. To prevent the negligence of synthesizing gray hair for the aged group caused by the diversity of hair colors in the training set of CACD database, we can consider of adding hair colour information (black, blond, brown and gray) to the final condition label.

For the parallel training generators, more tasks such as face aging and head pose transformation can be deployed to validate the contribution of the parallel training mechanism.

Bibliography

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[6] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, Jun. 2015.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018.

[8] Xing Di and Vishal M Patel. Facial synthesis from visual attributes via sketch using multiscale generators. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1):55–67, 2019.

[9] Xing Di, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Multi-scale thermal to visible face verification via attribute guided synthesis. *arXiv preprint arXiv:2004.09502*, 2020.

[10] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[11] Shichuan Du, Yong Tao, and Aleix M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, Apr. 2014.

[12] Chi Nhan Duong, Khoa Luu, Gia Kha Quach, and Tien D Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5772–5780, 2016.

[13] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D Bui. Beyond principal components: Deep boltzmann machines for face modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4786–4794, 2015.

[14] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Deep appearance models: A deep boltzmann machine approach for face modeling. *International Journal of Computer Vision*, 127(5):437–455, May 2019.

[15] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10013–10022, 2019.

[16] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T. Hoang Ngan Le, Marios Savvides, and Tien D. Bui. Learning from longitudinal face demonstration-where

tractable deep modeling meets inverse reinforcement learning. *International Journal of Computer Vision*, 127(6-7):957–971, Jun. 2019.

[17] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016.

[18] Face++. Face++ research toolkit., 2020. (accessed on 2020).

[19] Han Fang, Weihong Deng, Yaoyao Zhong, and Jiani Hu. Triple-gan: Progressive face aging with triple translation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Workshops*, pages 804–805, 2020.

[20] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

[21] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, Nov. 2010.

[22] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[23] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9821–9830, 2019.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[28] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5077–5086, 2017.

[29] Zhizhong Huang, Shouzhen Chen, Junping Zhang, and Hongming Shan. Pfa-gan: Progressive face aging with generative adversarial network. *IEEE Transactions on Information Forensics and Security*, 2020.

[30] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3341, 2014.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and Emotion*, 24(8):1377–1388, Dec. 2010.

[33] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[34] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570. PMLR, 2015.

[35] Jianshu Li, Shengtao Xiao, Fang Zhao, Jian Zhao, Jianan Li, Jiashi Feng, Shuicheng Yan, and Terence Sim. Integrated face analytics networks through

cross-dataset hybrid training. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1531–1539, 2017.

[36] Jianshu Li, Pan Zhou, Yunpeng Chen, Jian Zhao, Sujoy Roy, Yan Shuicheng, Jiashi Feng, and Terence Sim. Task relation networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 932–940. IEEE, 2019.

[37] Peipei Li, Yibo Hu, Qi Li, Ran He, and Zhenan Sun. Global and local consistent age generative adversarial networks. pages 1073–1078. IEEE, Aug. 2018.

[38] Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Hierarchical face aging through disentangled latent characteristics. In *European Conference on Computer Vision (ECCV)*, pages 86–101. Springer, 2020.

[39] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7245–7254, 2020.

[40] Jun Ling, Han Xue, Li Song, Shuhui Yang, Rong Xie, and Xiao Gu. Toward fine-grained facial expression manipulation. In *European Conference on Computer Vision (ECCV)*, pages 37–53. Springer, 2020.

[41] Wallace Lira, Johannes Merz, Daniel Ritchie, Daniel Cohen-Or, and Hao Zhang. Ganhopper: Multi-hop gan for unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, pages 363–379. Springer, 2020.

[42] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *PProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 773–782, 2018.

[43] Si Liu, Yao Sun, Defa Zhu, Renda Bao, Wei Wang, Xiangbo Shu, and Shuicheng Yan. Face aging with contextual generative adversarial nets. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 82–90, 2017.

[44] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11877–11886, 2019.

[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[46] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[47] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[48] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision (ECCV)*, pages 739–755. Springer, 2020.

[49] Minho Park, Hak Gu Kim, and Yong Man Ro. Photo-realistic facial emotion synthesis using multi-level critic networks with multi-level generative model. In *International Conference on Multimedia Modeling*, pages 3–15. Springer, 2019.

[50] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[51] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[52] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, pages 1–16, 2019.

[53] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018.

[54] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[55] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[56] E. Sanchez and M. Valstar. A recurrent cycle consistency loss for progressive face-to-face synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 53–60. IEEE Computer Society, may 2020.

[57] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6):123–151, 2005.

[58] Jun Shao and Tien D. Bui. Wavelet-based multi-level gan for face aging. In *Submitted to the journal Computer Vision and Image Understanding*, 2021.

[59] Jun Shao and Tien D. Bui. Wp2-gan: Wavelet-based multi-level gan for progressive facial expression translation with parallel generators. In *British Machine Vision Conference*, 2021.

[60] Jingkuan Song, Jingqiu Zhang, Lianli Gao, Xianglong Liu, and Heng Tao Shen. Dual conditional gans for face aging and rejuvenation. In *International Joint Conferences on Artificial Intelligence*, pages 899–905, 2018.

[61] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635, 2018.

[62] Yunlian Sun, Jinhui Tang, Zhenan Sun, and Massimo Tistarelli. Facial age and expression synthesis using ordinal ranking adversarial networks. *IEEE Transactions on Information Forensics and Security*, 15:2960–2972, 2020.

[63] Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):385–401, 2009.

[64] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[65] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*, pages 24–es. 2006.

[66] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2378–2386, 2016.

[67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[68] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7947, 2018.

[69] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *European Conference on Computer Vision (ECCV)*, pages 781–798. Springer, 2020.

[70] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5021–5030, 2020.

[71] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31–39, 2018.

[72] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning continuous face age progression: A pyramid of gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[73] Jiangfeng Zeng, Xiao Ma, and Ke Zhou. Photo-realistic face age progression/regression using a single generative adversarial network. *Neurocomputing*, 366:295–304, Nov. 2019.

[74] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[75] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5818, 2017.

[76] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Fang Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9251–9258, 2019.

[77] Jian Zhao, Shuicheng Yan, and Jiashi Feng. Towards age-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[78] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[79] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.