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

Jun Shao
sh_jun@encs.concordia.ca
Tien D. Bui
bui@cse.concordia.ca

Computer Science and Software Engineering
Concordia University
Montréal, Québec, Canada

Abstract

Expression translation has received increasing attention from the computer vision community due to its wide applications in the real world. However, expression synthesis is hard because of 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.

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) [1, 12, 16, 28, 38] have opened a new door to the face manipulating technologies [5, 6, 20, 26, 36]. The advent of Condition GAN (cGAN) [22] and Cycle-GAN [39] made the attributes editing on the same subject possible without paired images belonging to the same subject. Many recent models [5, 26, 31] 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 [9]). 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 [5], Sanchez *et al.* [29] 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 tasks. 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 2).

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 [19, 35], 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 [20] 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 [26] 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.

2 RELATED WORKS

2.1 GAN

Generative Adversarial Networks (GANs) [12] were first proposed to generate images based on minimax game theory, then were improved by many other works [1, 13, 16]. Later, Mirza and Osindero [22] proposed a conditional GAN (cGAN) that embeds prior information into image generation. Cycle-GAN [39] 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 [5, 26, 31, 37]. Most of these works only adopt one generator for target features translation and then the reconstruction of the input image. Sanchez *et al.* [29] 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 2).

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 [3] proposed the first 3D Morphable model for 3D face reconstruction. Vlasic *et al.* [32] presented a multilinear model of 3D face meshes for expression translation. Cao *et al.* [4] introduced a method for facial image animation based on the 3D face mesh. Geng *et al.* [11] proposed a 3D-guided generative model for continuous expressions editing. paGAN [23] 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 expression synthesis leverages deep generative models. Many previous works [10, 27, 30] performed discrete or continuous facial expression by conditioning on facial landmarks. ExprGAN [6] can control the intensity of expression by conditioning on an embedding generated from expression labels. LEED [34] 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 [5] achieved multi-task translation among different domains with one model. But this model can only generate limited and discrete emotion expressions. Pumarola *et al.* [26] proposed GANimation with an attention mechanism to predict continuous expression translation by conditioning on AUs [9]. However, this model still generates some artifacts for expression translations with a large gap. Many other works [24, 31] leverage multi-level discriminators to extract expression-related features during model training.

Different from [10, 27, 30], our approach can perform continuous expression editing by conditioning on AUs code which can be extracted by Openface [2] conveniently. Unlike [5, 24, 25, 31], which only utilize one generator for both forward prediction and reconstruction, our method adopts two parallel generators to alleviate the interference mentioned by [29]. Besides using multi-level discriminators like [31], 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 [35]. Different from Cascade EF-GAN that adopts three local sub-networks to synthesize 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 method can simplify model training and reduce the computing memory.

3 PROPOSED METHOD

3.1 Problem Formulation

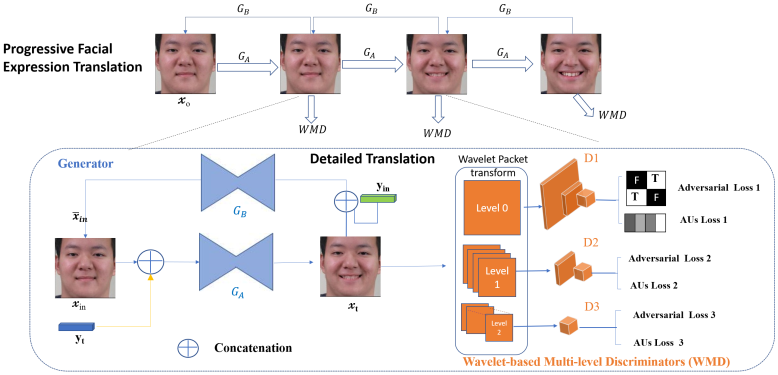

Figure 1: An overview of the WP2-GAN framework. The workflow of the progressive training is shown on the top, while the details of each step are shown in the zoom-in area. As two generators perform a similar task, we only show one stream of the translations. In each progressive step, one generator G_A 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.

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 [9] is adopted as AUs code, which can be extracted by

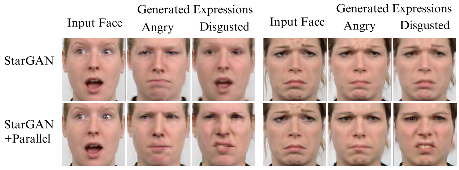

Figure 2: Comparison between StarGAN and modified StarGAN with parallel generators for expression translation. Each triplet contains the input face in the first column followed by outcome of the first translation (angry face) in the middle and then the result of second translation (disgusted face) based on the first outcome.

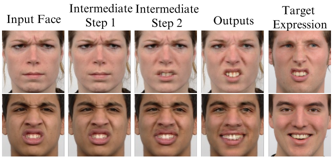

Figure 3: Display of progressive translation results by our WP2-GAN. The first column contains the input faces followed by two intermediate results and then the final outcomes. The last column shows images with target expressions. The progressive model provides a gradual transformation between expressions with a large gap.

OpenFace [2]. 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 1.

3.2 Parallel and Progressive Training Mechanism

Motivated by the discovery in [29], 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 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 [26].

Inspired by the impressive success of progressive training methods [19, 35] in geometric conversion, we design a novel progressive learning strategy for our task based on the parallel training mechanism. As shown in Figure 1, 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 [35], our progressive training is based on two parallel generators. 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 a similar idea of using parallel generators is [19]. 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 3.

Similar to [26, 35], 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:

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

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 Wavelet-based Multi-level Discriminators

Recently, wavelet-based multi-level discriminators have been successfully applied to face aging [20, 21]. Wavelet Packet Transform (WPT) can decompose an image into multi-level wavelet coefficients which contain both texture and geometric information [20]. 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 [20] 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 [13] 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.

3.4 Loss Functions

The loss functions used for our model include five items: (1) An adversarial loss $\mathcal{L}_{adv}$ used to distinguish fake images from real inputs. (2) An attention loss $\mathcal{L}_A$ to prevent the saturation of the attention mask. (3) A condition loss $\mathcal{L}_{cond}$ is adopted to guarantee the translation accuracy of expression. (4) A similarity loss $\mathcal{L}_{sim}$ to ensure two parallel generators proceed in the same direction. (5) Finally, a cycle-consistent loss $\mathcal{L}_{cyc}$ is utilized to preserve the identity-level consistency. The overall loss functions for G and D can be formulated as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{cond}\mathcal{L}_{cond} + \lambda_{sim}\mathcal{L}_{sim} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_A(\mathcal{L}_A(G, \mathbf{x}_{in}, \mathbf{y}_t) + \mathcal{L}_A(G, \mathbf{x}_t, \mathbf{y}_{in})), \tag{2}$$

where λ_{cond}, λ_{sim}, λ_{cyc} and λ_A are hyper-parameters for condition loss, similarity loss, cycle-consistent and attention loss, respectively. Due to the limit of paper length, please refers to supplementary materials for the detailed loss functions and structures of our neural networks.

4 EXPERIMENTS

4.1 Dataset

We train and test our model WP2-GAN on two public facial expression databases: RafD [17] and Compound Facial Expressions of Emotions Dataset (CFEED) [7]. 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 [2]. We also leverage Openface to extract the AUs code for every image.

4.2 Qualitative Experimental Results

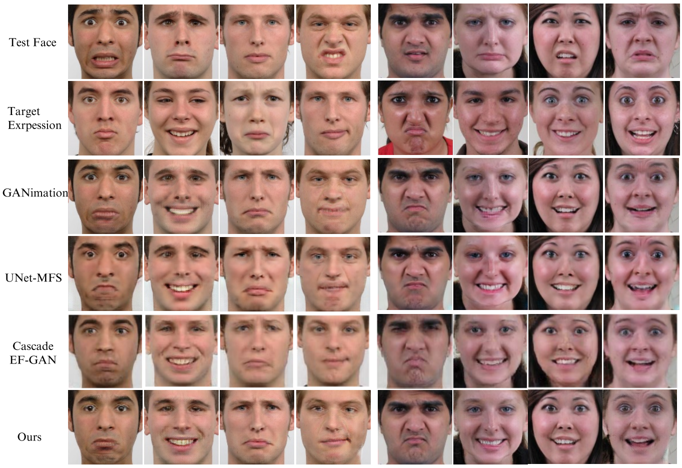

Figure 4: Qualitative comparison with previous works on RafD (left four columns) and CFEED (right four columns).

In this section, we test our approach on both RafD and CFEED and compare the results with three previous models: GANimation [25], UNet-MFS [18] and Cascade EF-GAN [35], all of which are conditioned on AUs code for continuous expression 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 [35] for Cascade EF-GAN due to the unavailability of the code.

As shown in Figure 4, 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.

4.3 Quantitative Experimental Results

We adopt a similar method of Cascade EF-GAN [35] and StarGAN [5] 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 [14]) on the filtered training set of each database, which only contains images with basic expression (i.e. angry, disgust, fearful, happy, sad, surprised or

	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

Table 1: Quantitative comparison among GANimation, Unet-MFS, Cascade EF-GAN and all variants of the proposed model.

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 1. The results of Cascade EF-GAN are from [35]. 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) [15] and structural similarity (SSIM) index [33]. A lower FID score and a higher SSIM normally represent a higher image quality. As shown in Table 1, 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 4 as well as FID scores in Table 1,

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

4.4 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 5 that both the baseline and Parallel-GAN fail to produce natural expressions but generating some artifacts in areas near mouth and eyes. The introduction

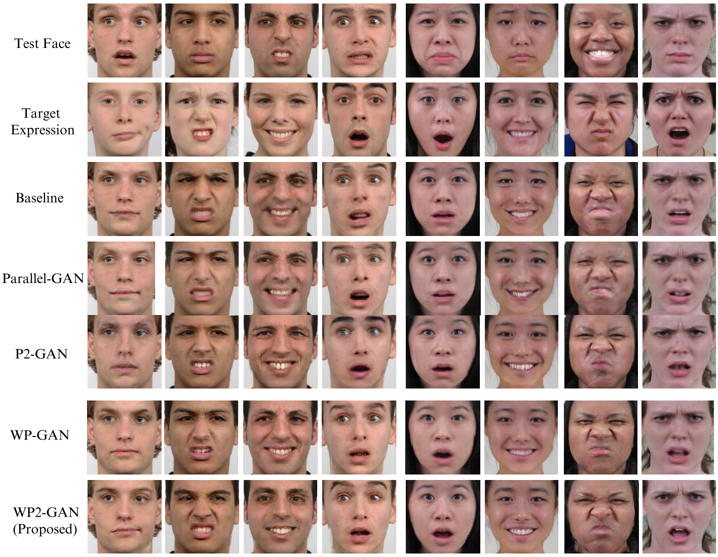

Figure 5: Comparison of expression translation between the proposal and variants of the proposed model on both RafD (left four columns) and CFEED (right four columns).

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 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 1 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.

4.5 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 6.

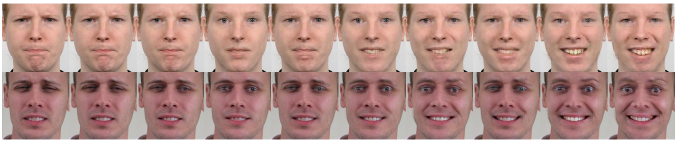

Figure 6: Continuous expression translation performed by our proposed model on both RafD (top) and CFEED (bottom). The first column contains the input images, followed by generated images with a continuous change of expression.

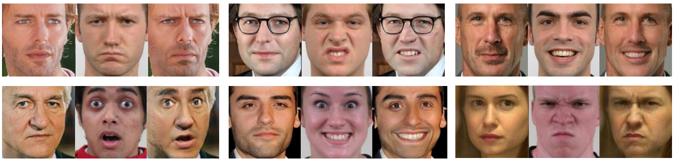

Figure 7: Sampled expression translation results by our proposed model on EmotioNet [8]. Each triplet contains the test face, the target expression and finally the synthesized image.

We also evaluate our method on images in the wild. We train the model on over 70,000 images from EmotiNet [8] then fine-tune the model on RafD and CFEED. Figure 7 shows that our approach can be applied to images with different background in the wild.

5 Conclusion

In this paper, 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.

Acknowledgment

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

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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*, pages 8789–8797, 2018.

[6] 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.

[7] 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. ISSN 0027-8424. doi: 10.1073/pnas.1322355111. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.1322355111`.

[8] 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*, pages 5562–5570, 2016.

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

[10] 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.

[11] 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*, pages 9821–9830, 2019.

[12] 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.

[13] 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.

[14] 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*, pages 770–778, 2016.

[15] 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.

[16] 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*, pages 5077–5086, 2017.

[17] 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. ISSN 0269-9931. doi: 10.1080/02699930903485076. URL http://www.tandfonline.com/doi/abs/10.1080/02699930903485076.

[18] 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*, pages 37–53. Springer, 2020.

[19] 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*, pages 363–379. Springer, 2020.

[20] 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*, pages 11877–11886, 2019.

[21] Yunfan Liu, Qi Li, Zhenan Sun, and Tieniu Tan. A3gan: An attribute-aware attentive generative adversarial network for face aging. *IEEE Transactions on Information Forensics and Security*, 16:2776–2790, 2021. doi: 10.1109/TIFS.2021.3065499.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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. doi: 10.1109/FG47880.2020.00015. URL https://doi.ieeecomputersociety.org/10.1109/FG47880.2020.00015.

[30] 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.

[31] 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.

[32] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3), 2005. ISSN 0730-0301. doi: 10.1145/1073204.1073209. URL https://doi.org/10.1145/1073204.1073209.

[33] 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.

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

[35] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5020–5029, 2020.

[36] 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*, pages 31–39, 2018.

[37] 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. ISSN 09252312. doi: 10.1016/j.neucom.2019.07.085. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231219310926.

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

[39] 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*, pages 2223–2232, 2017.