Cascade EF-GAN: Progressive Facial Expression Editing with Local Focuses

Rongliang Wu[1], Gongjie Zhang[1], Shijian Lu*[1], and Tao Chen[2]

[1]Nanyang Technological University
[2]Fudan University
ronglian001@e.ntu.edu.sg, {gongjiezhang,shijian.lu}@ntu.edu.sg, eetchen@fudan.edu.cn

Abstract

Recent advances in Generative Adversarial Nets (GANs) have shown remarkable improvements for facial expression editing. However, current methods are still prone to generate artifacts and blurs around expression-intensive regions, and often introduce undesired overlapping artifacts while handling large-gap expression transformations such as transformation from furious to laughing. To address these limitations, we propose Cascade Expression Focal GAN (Cascade EF-GAN), a novel network that performs progressive facial expression editing with local expression focuses. The introduction of the local focuses enables the Cascade EF-GAN to better preserve identity-related features and details around eyes, noses and mouths, which further helps reduce artifacts and blurs within the generated facial images. In addition, an innovative cascade transformation strategy is designed by dividing a large facial expression transformation into multiple small ones in cascade, which helps suppress overlapping artifacts and produce more realistic editing while dealing with large-gap expression transformations. Extensive experiments over two publicly available facial expression datasets show that our proposed Cascade EF-GAN achieves superior performance for facial expression editing.

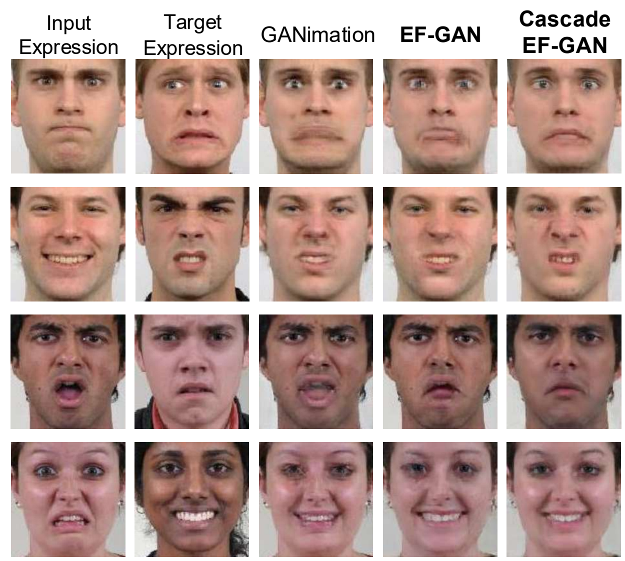

Figure 1. Illustration of the proposed facial expression editing method: The introduction of local focuses (in **EF-GAN**) helps better preserve details and reduce blurs and artifacts. The proposed progressive facial expression editing (in **Cascade EF-GAN**) further helps remove overlapping artifacts and generates more realistic expression images.

1. Introduction

Facial expression opens a window to people's internal emotions and conveys subtle intentions [22] and there exists many research works on automatic facial expression recognition [44, 36, 45, 49, 45]. In this day and age of digital media, facial expression editing, which transforms the expression of a given facial image to a target expression without losing identity properties, can potentially be applied in different areas such as photography technologies, movie industry, entertainment, etc. It has been attracting increasing attention from both academia and industry.

Inspired by the recent success of Generative Adversarial Nets (GANs) [10], several research works [30, 33, 7, 29, 6] have been reported and achieved very impressive facial expression editing results. On the other hand, existing methods are still facing a number of constraints. First, existing methods tend to generate incoherent artifacts and/or blurs, especially around those expression-rich regions such as eyes, noses and mouths. Second, existing methods tend to produce overlapping artifacts when the source facial ex-

*Corresponding author. This work is supported by Data Science & Artificial Intelligence Research Centre, NTU Singapore.

pression has a large gap with the target facial expression, such as transformation from furious to laughing[1].

The task of facial expression editing needs to maintain person identity. As humans, a natural way to identify facial images is to pay special attention to eyes, noses and mouths, largely because these regions contain rich identity-related information [1, 12]. On the other hand, almost all GANs-based facial expression editing methods [30, 33, 7, 29, 6] simply process the input facial image as a whole without paying special attention to local identity-related features, which could be one major reason why most existing methods generate incoherent artifacts and blurs around eyes, noses and mouths.

In addition, to the best of our knowledge, all existing GANs-based facial expression editing methods [30, 33, 7, 29, 6] perform a single-step transformation to the target expression. On the other hand, the single-step transformation often produces overlapping artifacts (around regions with large facial expression changes) while dealing with large-gap transformations due to the limitation of network capacity. Since facial expression changes are continuous by nature, a large-gap transformation should be better accomplished if the network decomposes it into a number of small transformations.

In this paper, we propose a novel *Cascade Expression Focal GAN (Cascade EF-GAN)* for progressive facial expression editing with local focuses. The Cascade EF-GAN consists of several identical EF-GAN modules in cascade that perform facial expression editing in a progressive manner. Specifically, an innovative cascade transformation strategy is designed which divides a large facial expression transformation into multiple small ones and performs facial expression transformation step-by-step progressively. Such progressive facial expression transformation helps suppress overlapping artifacts and achieve more robust and realistic expression editing while dealing with large-gap facial expression transformations. In addition, each EF-GAN module incorporates a number of pre-defined local focuses that capture identity-related features around eyes, noses and mouths, respectively. With the detailed identity-related features, the EF-GAN is capable of generating coherent facial expression images with much less artifacts. The results of our proposed Cascade EF-GAN are illustrated in Fig. 1.

The contributions of this work are threefold. First, we identify the importance of local focuses in facial expression editing, and propose a novel EF-GAN that captures identity-related features with several local focuses and mitigates the editing artifacts and blurs effectively. Second, we propose an innovative cascade design for progressive facial expression editing. The cascade design is robust and effective in suppressing overlapping artifacts while dealing with large-gap expression transformations. Third, extensive experiments show that the Cascade EF-GAN achieves superior facial expression editing quantitatively and qualitatively.

[1]Overlapping artifacts refer to the artifacts that original and target expressions are blended in the outputs.

2. Related Work

Generative Adversarial Nets: Generative Adversarial Nets (GANs) are powerful generative models that simultaneously train a generator to produce realistic fake images and a discriminator to distinguish between real and fake images. One active research topic is Conditional GANs [26] that include conditional information to control the generated images. In addition, CycleGAN [50] adopts cycle-consistency loss and achieves image-to-image translation with well preserved key attributes. GANs have demonstrated their powerful capabilities in different computer vision tasks such as natural image synthesis [3, 16], image style translation [15, 50, 17, 24, 37], super-resolution [20, 38, 31], image inpainting [41, 43, 28, 40], facial attributes editing [6, 46, 39, 32, 25, 5, 4], face image synthesis [18, 13, 42, 35, 47], etc. The GAN-generated images have also been applied to different computer vision tasks [47, 13, 48, 23]. Our Cascade EF-GAN is designed to perform facial expression editing, with conditional variables to control the target expressions and cycle-consistency to preserve identity information.

Facial Expression Editing: Facial expression editing is challenging as it requires high-level understanding of input facial images and prior knowledge about human expressions. Compared with general facial attributes editing which only considers appearance modification of specific facial regions [46, 21, 32], facial expression editing is a more challenging task as it often involves large geometrical changes and requires to modify multiple facial components simultaneously. Very impressive progress has been achieved with the prevalence of GANs in recent years. For example, G2-GAN [33] and GCGAN [30] adopt facial landmarks as geometrical priors to control the intensity of the generated facial expressions, where ground-truth images are essential for extracting the geometrical information. ExprGAN [7] introduces an expression controller to control the intensity of generated expressions, but it requires a pre-trained face recognizer for preserving the identity information. StarGAN [6] can translate images across domains with a single model and preserve identity features by minimizing a cycle loss, but it can only generate discrete expressions. GANimation [29] adopts Action Units as expression labels and can generate expressions in continuous domain. It also incorporates attention to better preserve the identity information. However, it tends to generate artifacts and blurs and cannot handle large-gap expression transformations well.

Instead of generating expressions on a whole face image as existing GAN-based methods, our proposed Cascade EF-GAN includes local focuses on eyes, nose and

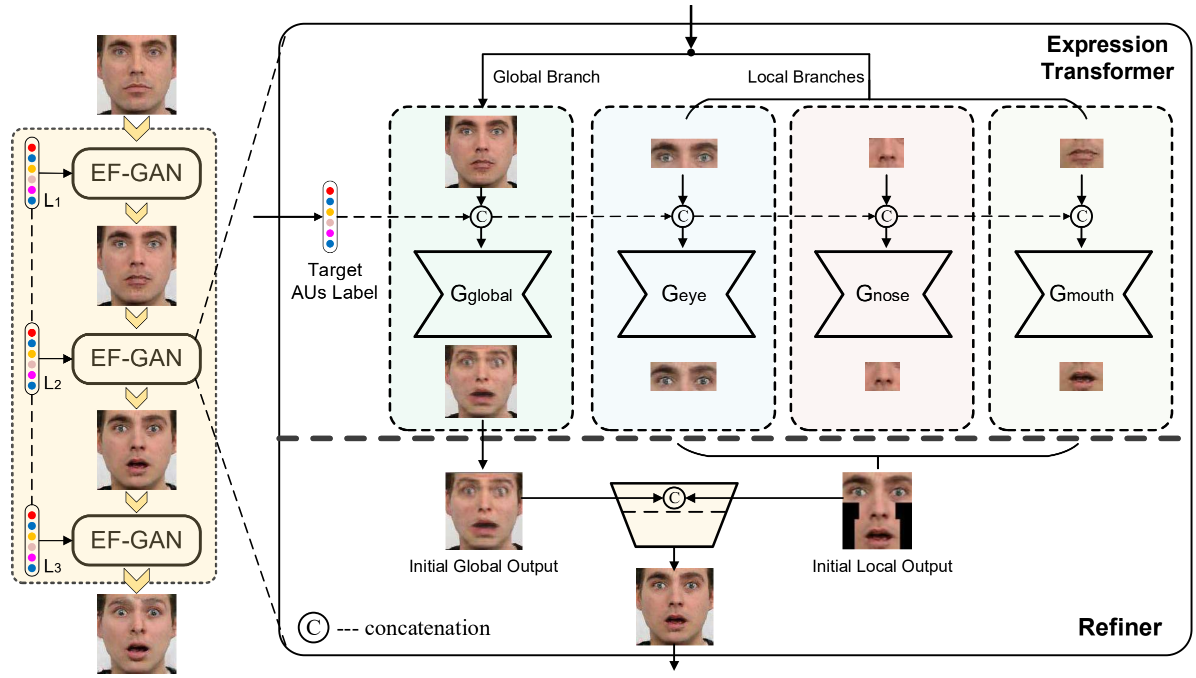

Figure 2. Illustration of our Cascade EF-GAN: The workflow is shown on the left, and the details of each EF-GAN are shown in the zoom-in area. The expression editing is decomposed into 3 steps and handled by each EF-GAN progressively. The EF-GAN consists of an Expression Transformer and a Refiner: The former generates the initial editing of the whole facial image and three local facial regions, and the latter fuses the initial editing and refines it to generate the expression image as the final output.

mouth regions that help suppress artifacts and preserve details clearly. In addition, the cascade strategy edits expressions in a progressive manner, which is able to suppress the overlapping artifacts effectively while dealing with transformations across very different expressions.

3. Proposed Methods

Fig. 2 shows the overall framework of our proposed Cascade EF-GAN. As Fig. 2 shows, the Cascade EF-GAN consists of multiple EF-GANs in cascade that performs expression editing in a progressive manner. Each EF-GAN shares the same architecture, which consists of an Expression Transformer and a Refiner. Specifically, several predefined local focuses branches are incorporated into each EF-GAN module for better preserving identity-related features and details around eyes, noses and mouths. More details are to be discussed in the ensuing subsections.

3.1. EF-GAN with Attention-Driven Local Focuses

The generative model within EF-GAN consists of an Expression Transformer that performs expression editing with local focuses, and a Refiner that fuses the outputs from the Expression Transformer and refines the final editing.

Expression Transformer. Fig. 2 shows the architecture of our Expression Transformer. Different from existing methods [30, 33, 7, 29, 6] that employ a single global branch to process the facial image, our Expression Transformer incorporates three additional local branches with pre-defined focuses on local regions around eyes, noses and mouths, respectively. The motivation is that convolutional kernels are shared across all spatial locations, but each facial region has distinct identity-related features. Simply processing a facial image as a whole with one set of convolutional kernels is thus not sufficient to capture identity-related details around each facial region. Inspired by [13, 47], our Expression Transformer tackles this challenge by processing facial images in both global and local branches, where the global branch captures global facial structures and local branches focus on more detailed facial features.

Specifically, the Expression Transformer takes a facial image and a target expression label as input. Similar to GANimation [29], we adopt the Facial Action Coding System (FACS) [9] that encodes expressions to Action Units (AUs) which can be extracted by using the open-source OpenFace [2]. We adopt the continuous AUs intensity as AUs labels to supervise the editing process. Given a source facial expression image, local focuses are first applied to the eyes, nose and mouth regions by cropping the corresponding local image patches. The landmarks for each local focus are also acquired by OpenFace [2]. The global facial image and its local patches are then fed to the corresponding

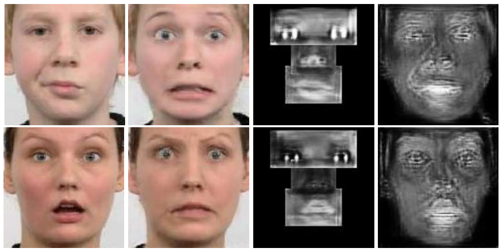

Figure 3. Illustration of local and global attention: global attention tends to capture the most salient expression changes, e.g. the mouth region; local attention captures fine expression changes at local regions, e.g. the eyes region. Note local attention responses are stitched for presentation purpose.

Figure 4. Illustration of intermediate and final editing by our Cascade EF-GAN: the progressive editing helps suppress overlapping artifacts and produce more realistic editing while dealing with large-gap expression changes.

branches of the Expression Transformer for expression editing. Note all branches share similar network architectures without sharing weights.

We also incorporate attention in global and local branches for better details capture and artifacts suppression. The use of visual attention has been investigated in GANimation [29], where attention was designed to guide the network to focus on transforming expression-related regions. On the other hand, applying attention in a single global image often introduces vague attention responses as illustrated in column 4 of Fig. 3. This is because the global attention tends to focus on the most salient changes, e.g. the mouth regions in Fig. 3, whereas fine changes around the eyes and nose are not well attended. The exclusive attention in the aforementioned local branches helps to achieve sharper responses at local regions as shown in column 3.

Specifically, each branch outputs color feature maps M_C and attention map M_A. With the original input image I_{in}, the initial output of each branch is generated by:

$$\mathcal{I}_{init} = M_A \otimes M_C + (1 - M_A) \otimes I_{in}, \quad (1)$$

where $\otimes$ denotes element-wise multiplication. This strategy eases the learning process greatly as the network does not need to output the initial results directly under the constraint of facial images statistics.

The Expression Transformer just generates initial expression editing as shown in Fig. 2. Specifically, the global branch captures global facial structures and features but generates blurs and artifacts around local regions due to the miss of local details. The local branches preserve local details better but they miss the big picture of the whole facial expression. The outputs of the two types of branches are therefore sent to the Refiner for fusion and further improvement.

Refiner. The Refiner is responsible for fusing the outputs of different branches of the Expression Transformer and generating the final expression editing. As Fig. 2 shows, the outputs of the three local branches are first stitched into a single image according to their respective locations within a facial image. The stitched image is then concatenated with the output of the global branch and fed to the Refiner to generate the final expression editing.

3.2. Cascade Facial Expression Transformation

Cascade Framework. Given an input facial image, the aforementioned EF-GAN is able to generate high-fidelity expression editing in most cases. On the other hand, our study shows that EF-GAN tends to produce overlapping artifacts around the regions with large expression changes while dealing with large-gap expression transformations. We refer large-gap expression transformations to those transformations that involve great appearance and geometrical modifications for editing the expression, such as transformation from furious to laughing. To address this constraint, we propose Cascade EF-GAN that performs expression editing in a progressive manner. Specifically, the Cascade EF-GAN decomposes a large-gap expression transformation into multiple small ones and performs large-gap expression transformations in cascade. It allows better preservation of facial structures and identity-related features as well as robust handling of large-gap facial transformations.

As Fig. 2 shows, the cascade expression editing is achieved by cascading multiple EF-GANs together, where the expression image from the previous EF-GAN is fed to the ensuing one as input for further editing. We empirically use 3 EF-GANs and Fig. 4 shows intermediate and final expression editing by the proposed Cascade EF-GAN. As Fig. 4 shows, the challenging large-gap expression editing is accomplished progressively in multiple steps, leading to realistic facial images of target expressions smoothly.

Intermediate Supervision: Another issue in implementing the progressive editing is how to include supervision information into each intermediate step. With the AU labels of input expression and the target expression, the straightforward approach is to generate intermediate AUs by linear interpolation. However, such interpolated AUs may not reside on the manifold of natural AUs and lead to weird synthesis. We address this issue by training an Interpolator to produce the intermediate AUs. Specifically, we first generate pseudo intermediate targets by linear interpolation and extract the residuals between the pseudo targets and the original AUs labels of input expression. The original AUs labels and residuals are then fed to the Interpolator to produce the intermediate AUs for providing supervision for the intermediate expression. Besides, a discriminator is trained to maximize the Wasserstein distance between the real and interpolated AUs while the Interpolator is trained to minimize the distance between them, so that the interpolated ones cannot be distinguished from real samples. Note all EF-GANs use the same AUs Interpolator.

3.3. Learning the Model

Loss Function The loss function for training the proposed EF-GAN contains five terms: 1) the adversarial loss for improving the photo-realism of the synthesized facial expression images to make them indistinguishable from real samples; 2) the conditional expression loss to ensure generated facial expression images to align with the provided target AUs labels; 3) the content loss for preserving the identity information and consistency of the image content. 4) the attention loss to encourage the attentive module to produce sparse attention map and pay attention to the regions that really need modification. 5) the interpolation loss to constrain the interpolated AUs label has desired sematical meaning and resides on the manifold of natural AUs. The overall objective function is expressed as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cond} + \lambda_2 \mathcal{L}_{cont} + \lambda_3 \mathcal{L}_{attn} + \lambda_4 \mathcal{L}_{interp} \tag{2}$$

where λ_1, λ_2, λ_3 and λ_4 are the hyper-parameters. In Cascade EF-GAN, the total loss is the sum of the loss of each EF-GAN with equal weights. Due to the limit of the paper length, please refer to the supplementary materials for the detailed loss and network architecture.

Training Scheme: It is difficult to obtain good expression editing if we directly cascade multiple EF-GAN modules and train them from scratch. We conjecture that this is largely due to the noisy facial images from the early-stage EF-GAN modules. Taking such noisy facial images as input, the later stages of the Cascade EF-GAN can be easily affected and produce degraded editing. In addition, the undesired editing will be accumulated, which makes network parameters difficult to optimize.

We design a simple yet effective scheme to address this issue. Specifically, we first train a single EF-GAN to perform a single-step facial expression transformation. Then, we use the weights of the well-trained EF-GAN to initialize all following EF-GAN in the cascade and fine-tune all network parameters end-to-end. With this training scheme, each EF-GAN module in the cascade will have good initialization, thus the intermediate facial expression images become useful to enable later stages to learn meaningful expression transformation information.

4. Experiments

4.1. Datasets

The Cascade EF-GAN is evaluated over Radboud Faces Dataset (RaFD) [19] and Compound Facial Expressions of Emotions Dataset (CFEED) [8]. RaFD consists of 8,040 expression images collected from different angles. We use the facial images captured by the 90° camera, leading to 1,608 facial expression images. CFEED contains 5,060 compound expression images collected from 230 participants. We randomly sample 90% for training and the rest are used for testing.

In our experiments, we crop the images into 128×128 patches with faces in the center. The sizes of the three local patches (i.e. patches of eyes, nose and mouth) are fixed to 40×92, 40×48 and 40×60, respectively. The center of each patch is the average position of the corresponding key points over all training samples.

4.2. Qualitative Experimental Results

The proposed Cascade EF-GAN is evaluated over two publicly available facial expression datasets described in previous section. Fig. 5 shows qualitative experimental results, where images in columns 1-5 are from the RaFD dataset and images in columns 6-10 are from the CFEED dataset. Each column includes an expression editing task, along with facial expression editing by state-of-the-art methods – StarGAN [6] and GANimation [29] as well as our proposed Cascade EF-GAN.

As Fig. 5 shows, state-of-the-art methods are prone to generate blurs and artifacts and even corrupted facial expressions around eyes, nose and mouth regions. Our Cascade EF-GAN instead generates more realistic facial expressions with much less blurs and artifacts, and its generated images are also much clearer and sharper. The better synthesis is largely attributed to the inclusion of attention-driven local focuses that helps to better preserve identity-related features and details in the corresponding facial regions. In addition, state-of-the-art methods tend to produce overlapping artifacts while handling large-gap expression transformations. Our Cascade EF-GAN instead suppresses such overlapping artifacts effectively, largely due to our cascade design that performs human-like progressive expression transformation rather than a single-step one. More results are provided in the supplementary materials.

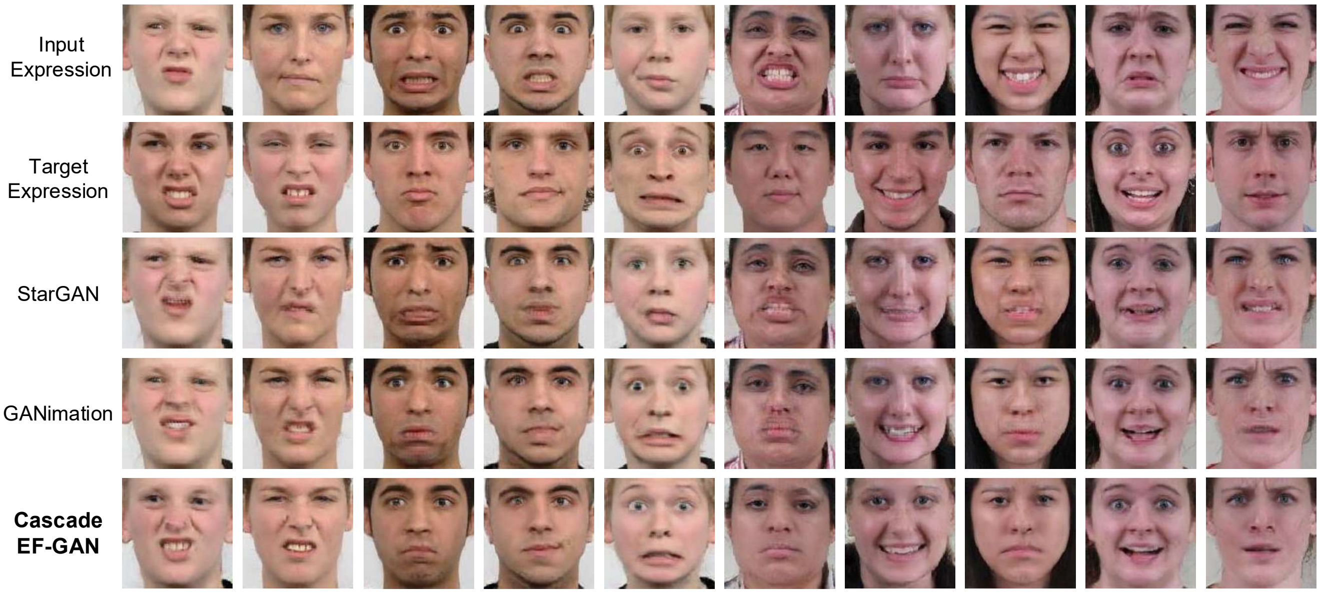

Figure 5. Expression editing by Cascade EF-GAN and state-of-the-art methods: Columns 1-5 show the editing of RaFD images, and the rest shows that of CFEED. Our Cascade EF-GAN produces more realistic editing with better details and less artifacts.

Table 1. Quantitative comparison with state-of-the-art on RaFD and CFEED datasets with facial expression recognition accuracy.

Dataset	Method	R	G	R + G
RaFD	StarGAN [6]	92.21	82.37	88.48
	GANimation [29]		84.36	92.31
	Ours		**89.38**	**93.67**
CFEED	StarGAN [6]	88.23	77.80	81.87
	GANimation [29]		79.46	84.42
	Ours		**85.81**	**89.25**

Table 2. Quantitative comparison with state-of-the-art on RaFD and CFEED datasets with PSNR (higher is better) and FID (lower is better).

	RaFD		CFEED	
	PSNR↑	FID↓	PSNR↑	FID↓
StarGAN [6]	19.82	62.51	20.11	42.39
GANimation [29]	22.06	45.55	20.43	29.07
Cascade EF-GAN	**23.07**	**42.36**	**21.34**	**27.15**

4.3. Quantitative Experimental Results

Expression Classification Accuracy: We follow the evaluation method of StarGAN [6] and ExprGAN [7] for quantitative evaluations. Specifically, we first train different expression editing models on the training set and perform expression editing on the same, unseen testing set. Then the generated images are evaluated in different expression recognition tasks. A higher classification accuracy indicates more accurate and realistic expression editing.

Two classification tasks are designed to evaluate the quality of the generated images: 1) train an expression classifier by using the original training images and apply the classifier to classify the expression images that are generated by different editing methods; 2) train classifiers by combining the natural and generated expression images to classify the original test set images. The first task evaluates whether the generated images lie in the manifold of natural expressions, and the second evaluates whether the generated images help train better classifiers.

Table 1 shows the expression classification accuracy on RaFD and CFEED (only seven primary expressions are evaluated for CFEED). Specifically, **R** means to train a classifier with original training set images then apply it to recognize the expression of testing set images. **G** means to use the same classifier (as in **R**) to recognize the expression of the generated images. **R + G** means to train classifiers by combining real and the generated images of different methods then apply them to recognize the expression of testing set images. As the table shows, our cascade EF-GAN achieves the highest accuracy in the first task, with 89.38% on RaFD and 85.81% on CFEED, showing its superiority in generating more realistic expression images. Additionally, it can help to train more accurate expression classifiers, where the accuracy is improved by 1.46% and 1.02% on RaFD and CFEED when our generated images are combined with real images in classifier training. As a comparison, StarGAN [6]

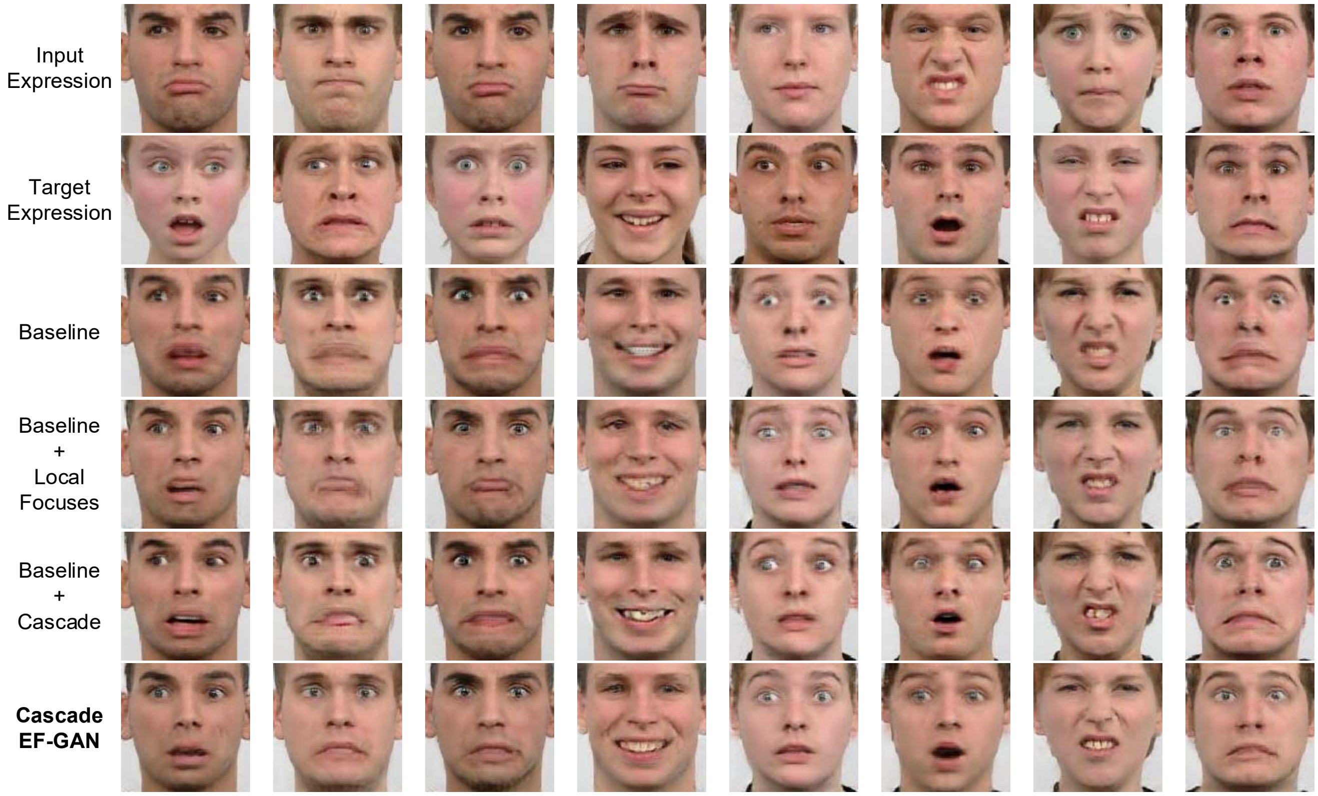

Figure 6. Ablation study of our Cascade EF-GAN over RaFD images: The inclusion of local focuses and cascade strategy helps to produce sharper expression images with better preserved details and less artifacts. Zoom-in may be needed for more details.

Table 3. Ablation study of classification accuracy on RaFD.

Method	G	R + G
Baseline	84.61	92.37
Baseline + Local Focuses	86.59	92.50
Baseline + Cascade	87.94	92.77
Cascade EF-GAN	**89.38**	**93.67**

and GANimation [29] generated images tend to degrade the classification, probably due to the artifacts and blurs within their generated images.

PSNR and FID: We also evaluate the quality of the generated images with peak signal-to-noise ratio (PSNR) [14] and Fréchet Inception Distance (FID) [11] metrics. The PSNR is computed over synthesized expressions and corresponding expressions of the same identity while the FID scores are calculated between the final average pooling features of a pretrained inception model [34] of real faces and the synthesized faces. As shown in Table 2, our proposed Cascsde EF-GAN outperforms the state-of-the-art method by 1.01/ 3.19 under the measurement of PSNR and FID on RaFD dataset, and 0.91/ 1.92 on CFEED, respectively.

4.4. Ablation Study

We perform ablation studies over the RaFD dataset to study the contributions of our proposed local focuses and cascade designs. Several models are trained including: 1) **Baseline** where only global attention is adopted as GANimation [29]; 2) **Baseline + Local Focuses** (i.e. EF-GAN) that includes the local focuses branches into the **Baseline**; 3) **Baseline + Cascade** that includes the progressive editing (with 3 EF-GAN modules) into the **Baseline**; and 4) **Cascade EF-GAN** that includes both progressive editing and local focuses as illustrated in Fig. 2.

Fig. 6 shows qualitative results. Each column represents an expression editing task and the corresponding editing by the aforementioned models. It is obvious that the **Baseline** tends to lose details around eyes and mouths, resulting in blurs, artifacts and even corruptions there. The generated expressions are not well aligned with target expressions either for a number of sample images. The **Baseline + Local Focuses** reduces artifacts and corruptions greatly, and generates clearer and sharper expression images. The inclusion

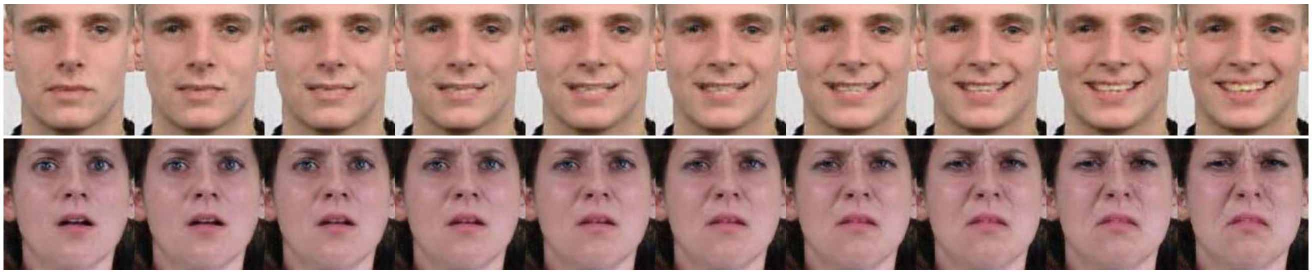

Figure 7. Continuous expression editing by Cascade EF-GAN: For input images in Column 1, the rest gives continuous editing.

Figure 8. Facial expression editing by the Cascade EF-GAN on wild images: In each triplet, the first column is input facial image, the second column is the image with desired expression and the last column is the synthesized result.

of the cascade strategy in **Baseline + Cascade** helps better maintain the identity features and face structure, and the generated expressions also align better with the target expressions. This is largely because the cascade design mitigates the complexity of large-gap changes by decomposing them into smaller steps. Finally, the **Cascade EF-GAN** which includes both the cascade design and local focuses is able to generate clean and sharp facial expressions that are well aligned with both target expressions and input identities, clearly better than all other models. This shows that the proposed local focuses and the cascade editing strategy are complimentary to each other.

We also conduct quantitative experiments to evaluate each proposed component in the Cascade EF-GAN. Table 3 shows experimental results. The quantitative experimental results further verify the effectiveness of the proposed local focuses and progressive transformation strategy.

4.5. Discussion

Continuous Expression Editing: Our Cascade EF-GAN can be easily adapted to generate continuous expressions. Given the source and target AUs, intermediate AUs of different stages can be derived with the Interpolator. Continuous expressions at intermediate stages can thus be derived with the intermediate AUs and the source images. Fig. 7 shows the continuous editing by Cascade EF-GAN.

Facial Expression Editing on Wild Images: Editing expression on wild images is much more challenging as the images are captured with complex background and uncontrolled lighting. Our Cascade EF-GAN can be adapted to handle wild images well as illustrated in Fig. 8, where the Cascade EF-GAN is first pre-trained on RaFD and CFEED images and then fine-tuned with wild expressive images from AffectNet [27]. As Fig. 8 shows, Cascade EF-GAN can transform the expressions successfully while maintaining the expression-unrelated information unchanged.

5. Conclusion

This paper presents a novel *Cascade Expression Focal GAN (Cascade EF-GAN)* for realistic facial expression editing. EF-GAN is designed by incorporating three local focuses on eyes, noses and mouths to obtain better preserved identity-related features and details. Such identity-related features reduce model's identity uncertainty, resulting in clearer and sharper facial expression images. In addition, the proposed Cascade EF-GAN performs expression editing in a progressive manner, decomposing large-gap expression changes into multiple small ones. It is therefore more robust in realistic transformation of large-gap facial expressions. Extensive experiments over two publicly available facial expression datasets show that the proposed Cascade EF-GAN achieves superior expression editing as compared with state-of-the-art techniques. We expect that Cascade EF-GAN will inspire new insights and attract more interests for better facial expression editing in the near future.

References

[1] Robert R Althoff and Neal J Cohen. Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):997, 1999.

[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] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Ying-Cong Chen, Xiaohui Shen, Zhe Lin, Xin Lu, I Pao, Jiaya Jia, et al. Semantic component decomposition for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9859–9867, 2019.

[5] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2416, 2019.

[6] 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.

[7] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

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

[10] 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.

[11] 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.

[12] Janet Hui-wen Hsiao and Garrison Cottrell. Two fixations suffice in face recognition. *Psychological science*, 19(10):998–1006, 2008.

[13] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448, 2017.

[14] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.

[18] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[19] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[21] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[22] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[23] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[24] Xiaodan Liang, Hao Zhang, and Eric P Xing. Generative semantic manipulation with contrasting gan. *arXiv preprint arXiv:1708.00315*, 2017.

[25] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 282–297, 2018.

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

[27] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature

learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[29] 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.

[30] 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.

[31] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[32] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4030–4038, 2017.

[33] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 627–635. ACM, 2018.

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[35] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[36] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, 2020.

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[38] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, pages 3891–3897, 2018.

[39] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 168–184, 2018.

[40] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17, 2018.

[41] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[42] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3990–3999, 2017.

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[44] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[45] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018.

[46] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–432, 2018.

[47] Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jianshu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, et al. 3d-aided deep pose-invariant face recognition. In *IJCAI*, volume 2.

[48] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems*, pages 66–76, 2017.

[49] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016.

[50] 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.