Local and Global Perception Generative Adversarial Network for Facial Expression Synthesis

Yifan Xia, Wenbo Zheng, Yiming Wang, Hui Yu, *Senior Member, IEEE*, Junyu Dong and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*—Facial expression synthesis has gained increasing attention with the development of Generative Adversarial Networks (GANs). However, it is still very challenging to generate high-quality facial expressions since the overlapping and blur commonly appear in the generated facial images especially in the regions with rich facial features such as eye and mouth. Generally, existing methods mainly consider the face as a whole in facial expression synthesis without paying specific attention to the characteristics of facial expressions. In fact, according to the physiological and psychological research, the differences of facial expressions often appear in crucial regions such as eye and mouth. Motivated by this observation, a novel end-to-end facial expression synthesis method called Local and Global Perception Generative Adversarial Network (LGP-GAN) with a two-stage cascaded structure is proposed in this paper which is designed to extract and synthesize the details of the crucial facial regions. LGP-GAN can combine the generated results from the global network and local network into the corresponding facial expressions. In Stage I, LGP-GAN utilizes local networks to capture the local texture details of the crucial facial regions and generate local facial regions, which fully explores crucial facial region domain information in facial expressions. And then LGP-GAN uses a global network to learn the whole facial information in Stage II to generate the generate final facial expressions building upon local generated results from Stage I. We conduct qualitative and quantitative experiments on the commonly used public database to verify the effectiveness of the proposed method. Experimental results show the superiority of the proposed method over the state-of-the-art methods.**

***Index Terms*—Facial expression synthesis, Generative adversarial networks, Facial expression recognition, local facial region, facial mask.**

Manuscript received November 5, 2020; revised January 13, 2021; accepted April 11, 2021. (*Corresponding author: Hui Yu.*)

Y. Xia, Y. Wang and H. Yu are with the School of Creative Technologies, University of Portsmouth, Portsmouth, PO1 2DJ, UK (e-mail: Yifan.Xia@myport.ac.uk; yiming.wang@port.ac.uk; hui.yu@port.ac.uk).

W. Zheng is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China, as well as with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (zwb2017@stu.xjtu.edu.cn).

J. Dong is with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: dongjunyu@ouc.edu.cn).

F.-Y. Wang is with the Institute of Systems Engineering, Macau University of Science and Technology, Macau, 999078, China, and with the Qingdao Academy of Intelligent Industries, Qingdao, 266109, China, as well as with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: feiyue.wang@ia.ac.cn).

I. Introduction

FACIAL expression recognition has a wide range of applications in various domains such as healthcare, virtual reality (VR), augmented reality (AR), driver assistant systems and entertainment [24, 25, 29, 35, 39, 40]. The performance of most existing methods, especially deep learning-based methods, heavily relies on the quantity and quality of training data. However, collecting a large annotated facial expression database is often limited by the resources we can find. And it also usually requires professional expertise as well as time consuming and expensive. Therefore, how to effectively generate the required facial data with minimum costs has been an interesting research topic in recent years.

Facial expression synthesis is one of the most successful solutions to the problem of insufficient data. The most notable method for facial expression synthesis is Generative Adversarial Networks (GANs), which can effectively generate the new facial expression of an input facial image according to the target expression label. Whilst GANs have achieved impressive results on the task of facial expression synthesis, existing methods still suffer from some limitations. Most existing GAN-based methods are designed for general face synthesis tasks without considering the characteristics of facial expressions, which are not appropriate to synthesize facial expressions. In fact, according to psychology research such as Facial Action Coding System (FACS)[2, 3], it is obvious that the differences between facial expressions usually locate in some certain crucial regions such as eye and mouth. Moreover, studies have shown that the attention naturally focuses on specific facial regions when humans recognize and distinguish different facial expression[1]. For example, eyes play an important role for fear analysis while the mouth is vital for recognizing happiness. However, previous GAN-based methods mainly focus on the face as a whole but these local facial parts have been significantly overlooked for facial expression synthesis, which leads to overlapping and blur in the local facial regions of generated results. With this observation, we investigate and utilize features of local facial regions in the task of facial expression synthesis.

In this paper, we propose a novel end-to-end facial expression synthesis method called Local and Global Perception Generative Adversarial Network (LGP-GAN) by integrating local and global facial information to synthesize facial expression. LGP-GAN is a two-stage cascaded architecture that decomposes the process of facial expression synthesis. Stage I utilizes local networks to generate local facial regions. Stage

II uses a global network to generate final facial expressions using the output of the Stage I as input. The local network pays special attention to the most discriminative regions such as eye and mouth while the global network mainly focuses on the whole facial structure. In summary, the contributions of this paper are summarized as follows:

- Inspired by the observation that the differences between facial expressions mainly occur in some crucial facial regions, we propose a novel end-to-end facial expression method LGP-GAN with the local network capturing texture details of these crucial facial regions while the global network learning the general structure and profile of the face.
- The proposed LGP-GAN has a two-stage cascaded architecture that divides the facial expression synthesis process into local facial region generation and global facial image generation. It can fully utilize both the local and global facial information in the process of facial expression synthesis, which can synthesize facial expressions step by step.
- We have explored the role of crucial facial regions in facial expression synthesis. The qualitative and quantitative evaluation on the public database of facial expression shows that the proposed LGP-GAN has superiority over the state-of-the-art methods. It also shows the importance of crucial facial regions in facial expression synthesis, which can further improve the performance.

The structure of this paper is as follows. Section II introduces the related work about Generative Adversarial Networks and facial expression synthesis. Section III describes the methodology about our proposed LGP-GAN in detail. Section IV provides the qualitative and quantitative experimental results of the proposed method on the commonly used public database. Finally, we conclude this paper and give future works in Section V.

II. Related Work

In recent years, the technologies of machine learning and deep learning has achieved impressive performance in various fields such as computer vision and pattern recognition [26, 28, 30, 31, 36, 38, 42, 43]. One of the deep learning methods, Generative Adversarial Networks (GANs) [27] designed according to the game theory have attracted increasing attention recently. The typical GAN consists of a generator and a discriminator. The generator generates fake images making the generated images as realistic as possible. On the other hand, the discriminator learns to distinguish the authenticity of the generated images. Both generator and discriminator are simultaneously trained in the same framework, which is so-called adversarial learning. After interactive confrontations, the generator can finally generate indistinguishable images resembling the real images. Based on classical GANs, many variants of GANs have been developed to further improve performance such as CGAN [4] and WGAN [5]. So far, GANs have become one of the most notable generative models have been applied to various computer vision tasks such as image-to-image translation [7, 32, 34], image synthesis [21, 23, 33] and image inpainting [22]. In particular, recent advances in GANs have achieved remarkable results in the facial attribute editing task which is one of the most successful applications of GANs. StarGAN [6] and AttGAN [8] are two representative methods in facial attribute editing. StarGAN [6] takes the facial image and the target facial attribute as input to complete the facial attribute editing task only using one generator and one discriminator. AttGAN [8] is a encoder-decoder architecture, which is very similar to StarGAN in terms of facial attribute editing but uses the latent representation to represent the facial attribute.

Facial expression synthesis can be treated as a subproblem of facial attribute editing which has been a hot research topic in various research fields, especially for facial analysis. Previous works just treat this task as a general image-to-image translation in which the facial expressions are considered as special facial attributes. Some existing methods [6, 8] handle the task of facial expression synthesis by modifying relevant facial attributes such as smiling, mouth open and mouth closed. For instance, StarGAN [6] uses a single generator to synthesize new facial expression of the input facial image according to given desired labels of facial expressions. However, these methods are not appropriate to generalize the facial expression synthesis task due to various facial deformations of facial expressions. In fact, facial expression synthesis is a more complex task compared with facial attribute editing, which usually has large transformations in facial regions. Recent advances in GANs specially designed for facial expression synthesis have shown impressive results and drawn prevalent attention. Zhou et al. [9] proposed a conditional difference adversarial autoencoder (CDAAE) to generate facial images with desired emotion states. Chen et al. [13] proposed Double Encoder Conditional GAN (DECGAN) for facial expression synthesis. There are also some works that utilize geometry information to guide the facial expression generation such as GC-GAN [11] and G2GAN [10]. Ding et al. [12] proposed ExprGAN which edited the facial expressions according to the controllable expression intensity. Pumarola et al. [20] designed a GAN network structure based on StarGAN generating new facial expressions according to the given Action Units (AU) labels, which achieved impressive results on facial expression synthesis.

However, existing GAN-based methods mainly consider the face as a whole without paying special attention to local facial regions which leads to overlapping and blur in the local facial regions of generated results. In fact, researches in physiology and psychology [1–3] have shown that the differences between facial expressions often appear in crucial regions such as eye and mouth instead of evenly appearing in the whole face. Moreover, the importance of local facial regions has been attracted increasing attention from various research fields especially facial analysis in healthcare. For example, Liu et al. [36] proposed a hierarchy convolutional neural network for facial paralysis evaluation which can extract facial paralysis features from local facial regions and reduce the impact of redundant information. Inspired by this, the main aim of this work is to consider and utilize the crucial facial regions in the task of facial expression synthesis. In this

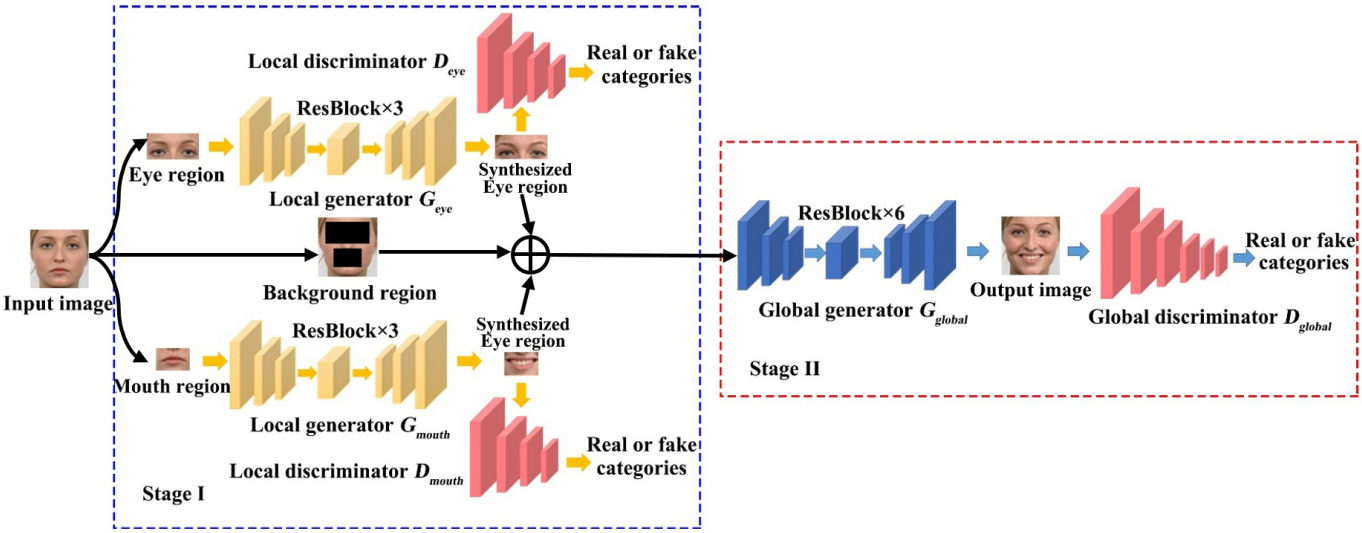


Fig. 1. The overview of the LGP-GAN for facial expression synthesis. Our method is a end-to-end model with a two-stage cascaded structure consisting of two local networks and one global network: In Stage I, the two local generators generate the crucial regions of eyes and mouth, which learn the transformation of different local facial regions between different facial expressions; In Stage II, the global network is used to the global network is used to perceive and supplement the global facial information outside of the crucial regions, which can further refine the generated results.

paper, we propose a Local and Global Perception Generative Adversarial Network (LGP-GAN) with a two-stage cascaded architecture. This kind of two-stage scheme has been proposed in previous works such as image segmentation, for instance, Zhao et al. [41] proposed a two-stage network for pancreas segmentation, which first determines the candidate regions and then refine the segmentation on these regions. Different from [41], we utilize two-stage scheme to fully utilize local and global facial information for facial expression synthesis in this paper. The proposed LGP-GAN divides the facial expression synthesis process into two parts: local facial region generation and global facial image generation, which can synthesize facial expressions step by step.

III. Method

In this section, we describe the details of the proposed LGP-GAN for facial expression synthesis. The proposed LGP-GAN is a two-stage cascaded architecture integrating local and global facial information to synthesize facial expression step by step. It divides the facial expression synthesis process into two parts: local facial region generation and global facial image generation. The proposed LGP-GAN aims at generating a target expression of a given facial image while retaining identity properties, which can be used to solve the problem of insufficient training data in facial expression recognition and applied in various fields such as healthcare and entertainment. Specifically, we denote the input RGB facial image as I_x with arbitrary facial expression. The facial expression image is characterized by an n-dimensional vector $f = [f^{(1)}, f^{(2)}, ..., f^{(n)}]^T$, where each attribute $f^{(i)}$ is a binary value (0 or 1) indicating the category of facial expressions, such as happy, sad, or surprise. Our objective is to transform the input image I_x with facial expression f_x to an output image I_y which has the target facial expression f_y. More details about the proposed LGP-GAN are described in the following parts.

A. Network Architecture

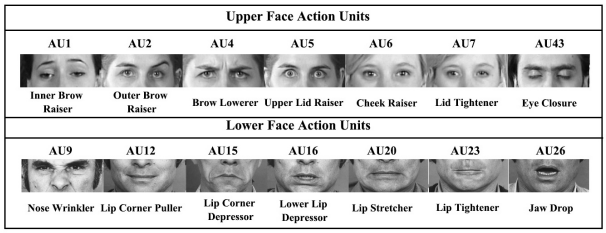


Fig. 2. Examples of facial action units [18].

TABLE I
List of AUs involved in six basic facial expressions.

Facial Expression	Action Units
Happy	AU6+AU12
Sad	AU1+AU4+AU15
Surprise	AU1+AU2+AU5+AU26
Disgust	AU9+AU15+AU16
Anger	AU4+AU5+AU7+AU23
Fear	AU1+AU2+AU4+AU5+AU7+AU20+AU26

As illustrated in Fig. 1, the proposed LGP-GAN has a cascaded structure which includes two local networks and one global network. The input includes a facial image I_x and a target facial expression label f_y. It outputs an image I_y that displays the face from the input image with the emotion represented by f_x. The local networks capture the local texture details corresponding to the local facial regions of the eye and mouth while the global network is used to learn the whole facial information. The rationale of this design is that the differences different facial expressions often occur in crucial regions. Existing methods simply take the whole facial image as input to synthesize facial expressions without paying special

Fig. 3. Crucial region samples of eight facial expressions on RaFD database [14]. It is obvious that the differences between facial expressions usually locate in the crucial regions of eyes and mouth. For example, eyes play an important role for fearful expressions while the mouth is vital for recognizing happy.

attention to local facial regions, which leads to overlapping and blur around local facial regions such as eyes and mouths. One main reason is that local facial regions have fewer pixels than other regions and thus the network is prone to fit the regions outside of local facial regions during network training. In this paper, the design of cascaded structure including two local networks and one global network can help the network better learn features in different local regions of a facial image. The proposed cascaded structure can divide the facial expression synthesis into local facial region generation and global facial image generation and thus synthesizes facial expressions step by step.

1) Preprocessing: The Facial Action Coding System (FACS) [2, 3] is a well-known taxonomy of human facial expressions. FACS defines 44 Action Units (AUs) (see Fig. 2) which represent the basic motions of single or groups of muscles. According to FACS, each specific facial expression can be expressed by a set of muscle contractions of AUs. The AU examples for six basic facial expressions is shown in Table 1. In Table 1, most AUs appear in the eye and mouth region of six basic facial expressions. For example, happy can be expressed with a combination of cheek raise (AU6) and lip corner puller (AU12). Therefore, the eye and mouth are the crucial regions of facial expressions, which have a major contribution to facial expressions. Inspired by this, we design two local networks to capture the local texture details of the crucial facial regions.

To utilize these crucial facial regions in our proposed LGP-GAN, the first step is to locate appropriate eye and mouth regions since too small or large regions can lead to less discriminative information or more redundant information from facial regions. The simple and effective method is based on the facial landmarks since the facial images can be easily detected and aligned by the recently advanced face detection and alignment methods. We first detect the facial landmarks, and then use these landmarks to obtain two local facial regions (eyes and mouth) as shown in Fig. 3. The advantage of this step is that it can focus on the most important and salient regions and thus reduces the interferences from other unrepresentative facial regions. Therefore, the whole face is divided into two parts based on the local facial regions. For a given facial image I_x, we can obtain the eye region I_{eye}, mouth region I_{mouth} and the background region I_{bg} outside of the crucial regions.

TABLE II
GENERATOR NETWORK ARCHITECTURE (N: THE NUMBER OF OUTPUT CHANNELS, K: KERNEL SIZE, S: STRIDE SIZE, P: PADDING SIZE).

	Layers	Layer Information
Global generator	Conv+IN+ReLU	N64, K7×7, S1, P3
	Conv+IN+ReLU	N128, K4×4, S2, P1
	Conv+IN+ReLU	N256, K4×4, S2, P1
	6×Residual Block	N256, K3×3, S1, P1
	DeConv	N128, K4×4, S2, P1
	DeConv	N64, K4×4, S2, P1
	Output	
	Color image: Conv+Tanh	N3, K7×7, S1, P3
	Attention mask: Conv+Sigmoid	N1, K7×7, S1, P3
Local generator	Conv+IN+ReLU	N64, K7×7, S1, P3
	Conv+IN+ReLU	N128, K4×4, S2, P1
	Conv+IN+ReLU	N256, K4×4, S2, P1
	3×Residual Block	N256, K3×3, S1, P1
	DeConv	N128, K4×4, S2, P1
	DeConv	N64, K4×4, S2, P1
	Output	
	Color image: Conv+Tanh	N3, K7×7, S1, P3
	Attention mask: Conv+Sigmoid	N1, K7×7, S1, P3

2) Generator: As shown in Fig. 1, our generator G has a two-stage cascaded architecture which consists of three basic generators: two local generators $G_{local} = \{G_{eye}, G_{mouth}\}$ and one global generator G_{global}. All three generators have similar architectures, but they are assigned with different learning tasks. We leverage two local generators to capture the deformation of crucial facial regions and one global generator to lean the whole facial texture. We borrow the idea of the basic architecture of the generators from [6] which has proven

to be successful for the task of the image-to-image translation, facial attribute editing and facial expression synthesis. The architecture of this kind of generators consists of a set of convolutional layers for downsampling, residual blocks and deconvolution layers for upsampling. The parameters of layers in each generator such as filter size, kernel size and stride are set referring to [6]. Table 2 illustrates the details of the generator network architecture including the local generator and global generator. The input channel of each generator is $3 + n$ defined by the channel of the input RGB image and the dimension of labels such as the number of categories of facial expressions. In order to focus on the facial areas that need to accommodate the changes in the process of facial expression synthesis, we also utilize attention mechanisms in each generator as in [20]. As shown in Table 2, we split the output layer of each generator into two parallel parts, one to generate the color mask C and the other to generate the attention mask M. The attention mask M is a weight matrix calculated adaptively by the network when generating the facial expression. To utilize the attention mask M, the final output of the generator I_{output} can be represented as:

$$I_{output} = (1 - M) \times C + M \times I_{input} \tag{1}$$

where $\times$ represents element-wise multiplication; I_{input} is the input image.

In Stage I, the crucial regions of eyes I_{eye} and mouth I_{mouth} are concatenated with the target facial expression label f_y as input to two local networks G_{eye} and G_{mouth}. The local generators G_{local} generate local facial regions $I_{local} = \{\hat{I}_{eye}, \hat{I}_{mouth}\}$, which learns the transformation of different local facial regions between different facial expressions; e.g., transforming mouth open to mouth closed. As shown in Table 2, each local generator consists of two downsampling convolutional layers, three bottleneck residual blocks, and two Deconvolutional layers. All layers are followed by Instance Normalisation (IN) [19] and ReLU activation except the output layer, where the Tanh function and Sigmoid function are used instead in the output layer of the color image and mask respectively.

In Stage II, the global network is used to perceive and supplement the global facial information outside of the crucial regions, which can further refine the generated results. For the input of the global network, after padding the generated local facial regions $I_{local} = \{\hat{I}_{eye}, \hat{I}_{mouth}\}$, we first blend generated local facial regions I_{local} and background region I_{bg} outside of the crucial regions to an aggregated result and then concatenate the target facial expression label f_y to it. As shown in Table 2, the global generator has a similar structure with the local generators but utilizing six bottleneck residual blocks. The output of the global network G_{global} is the final high-quality facial image I_y with target facial expression f_y.

3) Discriminator: As shown in Fig. 1, the discriminator D for the proposed method contains two local discriminators $D_{local} = \{D_{eye}, D_{mouth}\}$ and one global discriminator D_{global}. The local discriminator examines different local regions to evaluate the quality of local details, while the global discriminator examines the final output of the whole facial information to judge the holistic facial features. All discriminators have similar structures which are adapted from the PatchGAN architecture of [7]. Table 3 shows the details of the discriminator network architecture including the local discriminator and global discriminator. As shown in Table 3, the basic architecture of these three discriminators has no normalization and contains several convolutional layers for downsampling followed by Leaky ReLU activations with a slope of 0.01. The number of downsampling convolutional layers of the local discriminator D_{local} is set to 4 based on the size of the local regions while that of the global discriminator D_{global} is set to 6. We also select the parameters of layers in each discriminator such as filter size, kernel size and stride according to [7]. In this paper, we not only utilize each discriminator to distinguish whether the synthesized facial expression is a real or not but also use the auxiliary classifier of each discriminator to predict its category of facial expression.

TABLE III
DISCRIMINATOR NETWORK ARCHITECTURE (N: THE NUMBER OF OUTPUT CHANNELS, K: KERNEL SIZE, S: STRIDE SIZE, P: PADDING SIZE, h AND w: THE HEIGHT AND WIDTH OF THE INPUT IMAGE, n_d: THE NUMBER OF CATEGORIES OF FACIAL EXPRESSIONS).

	Layers	Layer Information
Global discriminator	Conv+Leaky ReLU	N64, K4×4, S2, P1
	Conv+Leaky ReLU	N128, K4×4, S2, P1
	Conv+Leaky ReLU	N256, K4×4, S2, P1
	Conv+Leaky ReLU	N512, K4×4, S2, P1
	Conv+Leaky ReLU	N1024, K4×4, S2, P1
	Conv+Leaky ReLU	N2048, K4×4, S2, P1
	Output	
	Real or fake: Conv	N1, K3×3, S1, P1
	Classifier: Conv	N(n_d), K$\frac{h}{64} \times \frac{w}{64}$, S1, P0
Local discriminator	Conv+Leaky ReLU	N64, K4×4, S2, P1
	Conv+Leaky ReLU	N128, K4×4, S2, P1
	Conv+Leaky ReLU	N256, K4×4, S2, P1
	Conv+Leaky ReLU	N512, K4×4, S2, P1
	Output	
	Real or fake: Conv	N1, K3×3, S1, P1
	Classifier: Conv	N(n_d), K$\frac{h}{16} \times \frac{w}{16}$, S1, P0

B. Loss Function

A weighted sum of four losses is imposed to supervise the proposed LGP-GAN in an end-to-end manner, including adversarial loss, facial expression classification Loss, reconstruction loss and attention loss. Given a facial image I_x with the facial expression f_x, the input of the network is $I_{input} = \{I_x, I_{eye}, I_{mouth}\}$ and the target facial expression is f_y. We introduce a two-stage cascaded generator $G = \{G_{local}, G_{global}\}$ in the proposed LGP-GAN. The output of two local generator networks $G_{local} = \{G_{eye}, G_{mouth}\}$ and one global generator network G_{global} are $I_{local} = \{\hat{I}_{eye}, \hat{I}_{mouth}\}$ and I_y respectively. We define $I_{output} = \{\hat{I}_{eye}, \hat{I}_{mouth}, I_y\}$ to represent the output of three subnetworks in the following section. And the output attention mask of three subnetworks is defined as $M = \{M_{eye}, M_{mouth}, M_y\}$. In order to generate the realistic facial image, we use a a composite discriminator $D = \{D_{local}, D_{global}\}$ in LGP-GAN to classify the output as real/fake and its category. $D_{local} = \{D_{eye}, D_{mouth}\}$ and

D_{global} are used to examine the outputs of the local generator $I_{local} = \{\hat{I}_{eye}, \hat{I}_{mouth}\}$ and global generator I_y.

Adversarial Loss: In the proposed method, we utilize an adversarial loss from WGAN-GP [5] to make the generated facial images indistinguishable from the real images. The adversarial loss can be written as:

$$\mathcal{L}_{adv} = \sum_{D_i \in D, G_i \in G} \{\mathbb{E}_{I_i}[logD_i(I_i)] - E_{I_i, f_y}[logD_i(G_i(I_i, f_y))] - \lambda_{gp}\mathbb{E}_{\tilde{I}}[\left\|\nabla_{\tilde{I}} D_i(\tilde{I})\right\|_2 - 1]^2\} \quad (2)$$

where $I_i \in I_{input}$, $\tilde{I}$ is the random interpolation distribution between the input image and generated image. And λ_{gp} is a penalty coefficient which was set as 10 in the experiment.

Facial Expression Classification Loss: The goal of facial expression synthesis is to translate an input image I_x with arbitrary facial expression f_x into an output image I_y with the target facial expression f_y. To achieve this aim, we add an auxiliary classifier on top of the discriminator and impose the Categorical Cross Entropy (CCE) as classification loss function during the training process. The objective can be divided into two parts: a classification loss of real images used to optimize the parameters of the discriminator and a classification loss of generated images used to optimize the parameters of the generator. The facial expression classification loss can be formulated as:

$$\mathcal{L}_{cls}^{D} = \sum_{D_i \in D} \mathbb{E}_{I_i \in I_{input}}[-logD_i(f_x|I_i)] \quad (3)$$

$$\mathcal{L}_{cls}^{G} = \sum_{D_i \in D} \mathbb{E}_{I_i \in I_{output}}[-logD_i(f_y|I_i)] \quad (4)$$

Reconstruction Loss: To maintain the identity information that the faces in both input and output images are from the same person, we utilize reconstruction loss including a cycle reconstruction loss and a self reconstruction loss to add extra constraints. Inspired by [7], we use L1 loss for the reconstruction since it produces less blur compared with L2 loss. During the process of facial expression synthesis, we minimize the L1 difference between the original image and its final reconstruction which is the final output image of the whole network. The cycle reconstruction loss and self reconstruction loss can be formulated as:

$$\mathcal{L}_{cycle} = \mathbb{E}_{I_x}[\|G(G(I_x, f_y), f_x) - I_x)\|_1] \quad (5)$$

$$\mathcal{L}_{self} = \mathbb{E}_{I_x}[\|G(I_x, f_x) - I_x)\|_1] \quad (6)$$

Attention Loss: The purpose of the attention mask is to make the network focus on the facial areas where large changes occur in the process of facial expression synthesis. However, the attention mask is very easy to saturate to 1 during training which significantly reduces the effect of the generator. Therefore, we borrow the idea of attention loss from [20] to avoid this situation using L2 loss. The attention loss can be represented as:

$$\mathcal{L}_{att} = \mathbb{E}_{I_x}[\|M_{eye}\|_2 + \|M_{mouth}\|_2 + \|M_y\|_2] \quad (7)$$

Full Loss: The full loss function for G and D are expressed, respectively, as

$$\mathcal{L}_D = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cls}^{D} \quad (8)$$

$$\mathcal{L}_G = -\mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cls}^{G} + \lambda_2 \mathcal{L}_{cycle} + \lambda_3 \mathcal{L}_{self} + \lambda_4 \mathcal{L}_{att} \quad (9)$$

where λ_1, λ_2, λ_3 and λ_4 are the hyper-parameters that represent the weight of each loss function. During network training, referring to [6, 8, 20] and the network performance, we set the value $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 10$, $\lambda_4 = 0.1$ in our experiment.

IV. Experiment

This section first introduces the database and evaluation metrics used in our experiment and then describes the implementation details of our experiment. Finally, this section presents the experimental results of our proposed method and compares with existing methods.

A. Database and Evaluation Metrics

The Radboud Faces Dataset (RaFD) [14] is used as our training data. The RaFD consists of 4,824 images with a size of 681×1024 collected from 67 participants under the laboratory settings. The facial images of each participant are captured by cameras from three different angles. We only use the frontal facial images which are about 1,608 in total in order to ensure the eye and mouth region are visible. And we randomly select 90% facial images for the training set and the remaining 10% facial images for the test set. The facial images in this database are labelled by eight discrete categories of facial expressions including angry, contemptuous, disgusted, fearful, happy, neutral, sad and surprised.

We use two different evaluation metrics in our experiment for quantitative evaluation of the facial expression synthesis task. Inception Score (IS) [15] and Fréchet Inception Distance (FID) [16] are widely used to evaluate the quality of the synthesized images. The IS scores measure the image quality through the probability outputs of the pre-trained Inception network which calculates the KL divergence between the conditional distribution and marginal distribution. The FID scores utilize an pre-trained Inception network to extract the final average pooling features of the real images and the synthesized images and measure their similarity.

B. Implementation Details

During the stage of preprocessing, all the input facial images were first aligned and cropped to the size of 128×128 according to the facial landmarks detected by Dlib [37]which is an open-source library capable of detecting 68 landmarks of the face. And then we obtained the corresponding eye and mouth region of each facial image according to the detected facial landmarks. In our experiments, the sizes of the eye region and mouth region are 48×96 and 48×64, respectively. The proposed LGP-GAN was trained in the PyTorch framework through an NVIDIA GeForce GTX 1080 GPU with 8 GB memory on a desktop PC. We used Adam optimizer [17] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ to train the model. The batch size was

Fig. 4. Qualitative results of facial expression synthesis on RaFD database (Input, Angry, Contemptuous, Disgusted, Fearful, Happy, Neutral, Sad, Surprised).

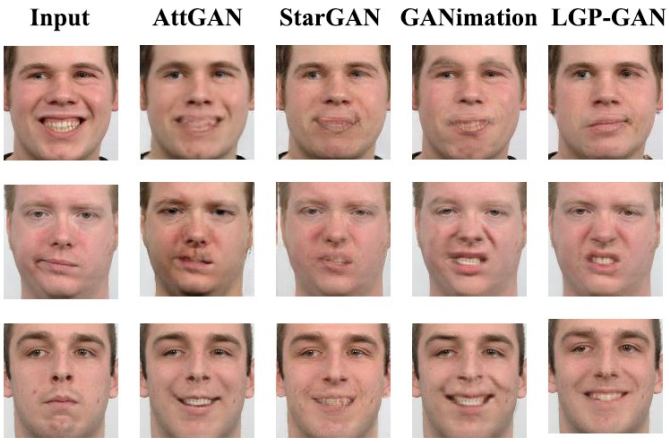


Fig. 5. Qualitative comparison of facial expression synthesis on RaFD database (target facial expression from top to bottom: contemptuous, disgusted and happy).

set as 16. For the learning rate, we initially set the learning rate as 0.0001 for the first 100k iterations and decayed linearly to 0 for the next 100k iterations. We trained the network updating the discriminator five times for each generator update.

C. Qualitative Evaluations

For qualitative evaluations, eight categories of facial expressions (Angry, Contemptuous, Disgusted, Fearful, Happy, Neutral, Sad and Surprised) in the RaFD database were used for facial expression synthesis to evaluate our proposed LGP-GAN. Fig. 4 illustrates the generated results of each category of facial expressions using our proposed LGP-GAN. As Fig. 4 shows, our proposed LGP-GAN can generate realistic facial images with eight categories of facial expressions and their local regions such as eyes and mouths are clear and sharp. The further evaluation is to compare the performance of the proposed LGP-GAN with the current state-of-the-art approaches including StarGAN [6], AttGAN [8] and GANimation [20]. StarGAN and AttGAN are two representative methods treating the task of facial expression synthesis as a subproblem of facial attribute editing. GANimation is another representative method that designs a special GAN network architecture and achieves impressive results on facial expression synthesis recently. Fig. 5 shows the qualitative comparison of facial expression synthesis using different methods on the RaFD database. In Fig. 5, the input facial images of each column are from the RaFD database. Each column represents the process of synthesizing

a target facial expression using different approaches including AttGAN, StarGAN, GANimation and our proposed LGP-GAN. As illustrated in Fig. 5, we can conclude observations as follows. Firstly, among these state-of-the-art methods, GANimation has better performance than AttGAN and StarGAN in terms of facial expression synthesis. For instance, AttGAN and StarGAN fail to transform happy into contemptuous in the first row of Fig. 5, which retains some details of happy especially in the local region such as the mouth. The experimental results demonstrate that the methods for facial attribute editing are not generalizable enough for generating the facial expression and facial expression synthesis is a more complex task because of large transformations and deformations in facial regions. Secondly, as shown in Fig. 5, all state-of-the-art methods are prone to generating blur and overlapping around local regions when large facial deformation occurs such as transforming from mouth open to mouth closed. The reason is that these methods only consider the general facial information without paying special attention to local facial regions. Finally, it is obvious that the proposed LGP-GAN method has shown clear superior overall quality over the existing state-of-the-art methods in terms of quality, clarity, and coherence of target facial expressions. In particular, the proposed LGP-GAN has fewer blurs and overlapping in generated facial expressions especially in local facial regions such as the eye and mouth region. The main reason is that the proposed LGP-GAN can capture texture details of crucial facial regions and learn the whole facial information simultaneously. In addition, the proposed LGP-GAN has a two-stage cascaded architecture that divides the facial expression synthesis into local facial region generation and global facial image generation, which enables to learn the facial details step by step.

D. Quantitative Evaluations

TABLE IV
COMPARISON OF OUR METHOD WITH OTHER METHODS ON RAFD DATABASE USING IS AND FID.

Method	$IS \uparrow$	$FID \downarrow$
StarGAN [6]	1.29	14.73
AttGAN [8]	1.26	18.18
GANimation [20]	1.30	12.20
LGP-GAN	**1.31**	**11.88**

For quantitative evaluations, we evaluate the quality of the generated facial expressions using IS [15] and FID [16] metrics. The better quality of the synthesized image is, the smaller of the FID and the larger of IS are.We also compared the proposed method with state-of-the-art methods on the RaFD database. Table 4 illustrates the comparison results using IS and FID metrics on the RaFD database. For IS metric, our proposed LGP-GAN slightly outperforms other state-of-the-art methods within a max margin of 0.05 as shown in Table 4. It is because that IS just evaluated the synthetic results using the output of pre-trained Inception network on ImageNet instead of comparing the synthetic distribution with the real distribution, which is not well appropriate to facial image database such as RaFD. But IS is still one of the most widely used metrics for evaluating the performance of GANs. For FID metric, as shown in Table 4, our proposed LGP-GAN has the lowest FID, which indicates a distinct lead over other state-of-the-art methods. Compared with IS, FID considers both real samples and synthetic samples which is one of the most robust metrics for evaluating the performance of GANs. In terms of IS and FID metric, GANimation and the proposed LGP-GAN outperform StarGAN and AttGAN. It shows that the approaches for facial attribute editing is not well suitable for the facial expression synthesis task. In addition, the proposed LGP-GAN achieves better performance than GANimation, which shows the importance of crucial facial regions in facial expression synthesis and can further improve the performance.

V. CONCLUSION

In this paper, we propose a novel end-to-end facial expression synthesis method called Local and Global Perception Generative Adversarial Network (LGP-GAN). The proposed LGP-GAN has a two-stage cascaded architecture including two local networks and a global network. The proposed LGP-GAN can fully utilize local facial information and global facial information in the process of facial expression synthesis. The quantitative and qualitative evaluations on the public database show that the proposed LGP-GAN has superior performance compared with state-of-the-art approaches on the task of facial expression synthesis.

Although the proposed LGP-GAN can work well on the lab-collected datasets, it cannot perform perfectly sometimes for the facial expressions from natural and un-controlled conditions. Moreover, the facial images used in this paper are labelled by eight discrete categories of facial expressions. Thus, the proposed method can only generate discrete facial expressions. However, not all facial expressions can be represented by discrete categories which ignore the intensity levels of facial expressions. In the future, we will explore how to adapt LGP-GAN to unconstrained facial expression datasets for expression synthesis considering different intensity levels of facial expressions.

REFERENCES

[1] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PloS one*, vol. 12, no. 5, p. e0177239, 2017.

[2] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press*, 1978.

[3] P. Ekman, "Facial action coding system," *A Human Face*, 2002.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. 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*, 2018, pp. 8789–8797.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[8] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[9] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2017, pp. 370–376.

[10] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 627–635.

[11] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive gan for facial expression transfer," *arXiv preprint arXiv:1802.01822*, 2018.

[12] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," *arXiv preprint arXiv:1709.03842*, 2017.

[13] M. Chen, C. Li, K. Li, H. Zhang, and X. He, "Double encoder conditional gan for facial expression synthesis," in *2018 37th Chinese Control Conference (CCC)*. IEEE, 2018, pp. 9286–9291.

[14] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE transactions on affective computing*, 2017.

[19] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[20] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 818–833.

[21] Z. Wang, G. Healy, A. F. Smeaton, and T. E. Ward, "Use of neural signals to evaluate the quality of generative adversarial network performance in facial image generation," *Cognitive Computation*, vol. 12, no. 1, pp. 13–24, 2020.

[22] M. Chen, Z. Liu, L. Ye, and Y. Wang, "Attentional coarse-and-fine generative adversarial networks for image inpainting," *Neurocomputing*, 2020.

[23] Y. Chen, S. Xia, J. Zhao, M. Jian, Y. Zhou, Q. Niu, R. Yao, and D. Zhu, "Person image synthesis through siamese generative adversarial network," *Neurocomputing*, 2020.

[24] X. Sun and M. Lv, "Facial expression recognition based on a hybrid model combining deep and shallow features," *Cognitive Computation*, vol. 11, no. 4, pp. 587–597, 2019.

[25] Y. Guo, Y. Xia, J. Wang, H. Yu, and R.-C. Chen, "Real-time facial affective computing on mobile devices," *Sensors*, vol. 20, no. 3, p. 870, 2020.

[26] Y. Xia, H. Yu, and F.-Y. Wang, "Accurate and robust eye center localization via fully convolutional networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 5, pp. 1127–1138, 2019.

[27] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.

[28] L. Chen, X. Hu, W. Tian, H. Wang, D. Cao, and F.-Y. Wang, "Parallel planning: a new motion planning framework for autonomous driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 236–246, 2019.

[29] Z. Lian, Y. Li, J.-H. Tao, J. Huang, and M.-Y. Niu, "Expression analysis based on face regions in read-world conditions," *International Journal of Automation and Computing*, vol. 17, no. 1, pp. 96–107, 2020.

[30] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. Masero, and A. Hussain, "Deep learning based single image super-resolution: a survey," *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 413–426, 2019.

[31] B. Hu and J. Wang, "Deep learning based hand gesture recognition and uav flight controls," *International Journal of Automation and Computing*, vol. 17, no. 1, pp. 17–29, 2020.

[32] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE transactions on circuits and systems for video technology*, 2019.

[33] M. Yuan and Y. Peng, "Bridge-gan: Interpretable representation learning for text-to-image synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[34] Y. Pang, J. Xie, and X. Li, "Visual haze removal by a unified generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3211–3221, 2019.

[35] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, "Realistic facial expression reconstruction for vr hmd users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, 2019.

[36] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham,

"Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2325–2332, 2020.

[37] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[38] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review." *Neurocomputing*, 2021.

[39] Y. Wang, X. Dong, G. Li, J. Dong, and H. Yu, "Cascade regression-based face frontalization for dynamic facial expression analysis," *Cognitive Computation*, pp. 1–14, 2021.

[40] S. Zhang, H. Yu, T. Wang, J. Dong, and T. D. Pham, "Linearly augmented real-time 4d expressional face capture," *Information Sciences*, vol. 545, pp. 331–343, 2021.

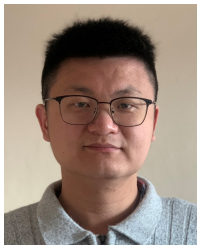
[41] N. Zhao, N. Tong, D. Ruan, and K. Sheng, "Fully automated pancreas segmentation with two-stage 3d convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 201–209.

[42] W. Liu, Z. Wang, N. Zeng, Y. Yuan, F. E. Alsaadi, and X. Liu, "A novel randomised particle swarm optimizer," *International Journal of Machine Learning and Cybernetics*, pp. 1–12, 2020.

[43] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, and X. Liu, "A dynamic neighborhood-based switching particle swarm optimization algorithm," *IEEE Transactions on Cybernetics*, 2020.

Yifan Xia received the M.Sc. degree from Ocean University of China in 2017. He is currently pursuing the Ph.D. degree at the School of Creative Technologies, the University of Portsmouth, UK. His research interests include computer vision, especially eye gaze estimation and facial expression recognition.

Wenbo Zheng received his bachelor degree in software engineering from Wuhan University of Technology, Wuhan, China, in 2017. He is currently a Ph.D. candidate in the School of Software Engineering, Xi'an Jiaotong University, as well as the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision and machine learning.



Yiming Wang received his Bachelor of Science (B.Sc) and Master of Science (Msc), both in the field of computer science and technology, from Zhengzhou University, China, in 2011 and 2014 respectively. He obtained his Doctor of Philosophy (PhD) in the field of computer vision from University of Portsmouth, UK in 2018. His research interests include human machine interaction, medical image processing and machine learning.

Hui Yu is a Professor with the University of Portsmouth, UK. His research interests include methods and practical development in visual computing, machine learning and AI with the applications focusing on human-machine interaction, virtual/augmented reality and robotics as well as 4D facial expression generation, perception and analysis. He serves as an Associate Editor for IEEE Transactions on Human-Machine Systems and Neurocomputing journal.

Junyu Dong received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in 2003. He joined the Ocean University of China in 2004, where he is currently a Professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision, and underwater vision.

Fei-Yue Wang (S'87-M'89-SM'94-F'03) received his Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined The University of Arizona in 1990 and became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center at the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems. His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation.