IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 20, 2021, accepted April 4, 2021, date of publication April 9, 2021, date of current version April 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072057

Facial Landmarks and Expression Label Guided Photorealistic Facial Expression Synthesis

DEJIAN LI, WENQIAN QI, AND SHOUQIAN SUN
Zhejiang Key Laboratory of Design and Intelligence and Digital Creativity, School of Computer Science, Zhejiang University, Hangzhou 310027, China

Corresponding author: Shouqian Sun (ssq@zju.edu.cn)

This work was supported in part by the Provincial Key Research and Development Plan of Zhejiang Province under Grant 2019C03137, and in part by the Scientific Innovation 2030–Project of Artificial Intelligence in the Next Generation under Grant 2018AAA0100700.

ABSTRACT Facial expression manipulation plays an increasingly important role in the field of computer graphics and has been widely used in generating facial animations. However, it is still a very challenging task as it needs full understanding of the input face and very depending on the facial appearance. In this paper, we present an end-to-end generative adversarial network for facial expression synthesis. Given the facial landmarks and the expression label of a target image, our method automatically generates a corresponding expression facial image with the identity information and facial details well preserved. Both qualitative and quantitative experiments are conducted on the CK+ and Oulu-CASIA datasets. Experimental results show that our method has the compelling perceptual results even there exist large differences in facial shapes for unseen subjects.

INDEX TERMS Facial expression synthesis, generative adversarial networks.

I. INTRODUCTION

Transform the expression of a given face image to a target one while preserving the identity information has drawn much attention in the area of human computer interactions [1] and affective computing [2]. It has received considerable attention both in the form of industrial research communities and commercial products, and has been widely applied in face editing, facial animations, face data augmentation, etc. The past few decades has witnessed many facial expression synthesis methods, which can be roughly divided into two categories. The first category tries to synthesize facial expression via the traditional methods such as image reordering [3], [4], 2D expression mapping methods [5], [6] and 3D-based methods [7], [8], while the other mainly builds generative models based on multiple facial informations [9]–[11].

For the first category, most works are either focus on finding mappings between existing facial expression textures and target images or manipulating existing image patches directly. Specifically, Pighin *et al.* [12] create smooth transitions between different facial expressions by morphing between the corresponding textured 3D facial models. Yang *et al.* [7] propose to learn a 2D flow field which can warp the target face in a natural way. However, this kind of methods cannot generate unseen facial images, furthermore, the complex processes often requires plenty of computation resources.

The second category is often known as facial representation learning based methods. One of the most impressive representation model is generative adversarial network(GAN), which has recently obtained remarkable results for image synthesis [13], [14]. For instance, Shu *et al.* [15] propose a rendering-based disentangling network, which generates disentangled facial contributes related latent vectors for manipulating facial appearances. To control the intensity of facial expression, Ding *et al.* [11] present an expression controller module that achieves to learn a compact expression code for expression synthesis(from weak to strong). Therefore, different from computer graphics technique based method, GAN related methods tend to encode facial information into a latent space, which provides better flexibility in semantic-level image generation. However, there are still remains many challenges, such as introduces more prior knowledge into network, designs more sophisticated network structure to fine-grain control of the synthesized images, and so on.

In this paper, we propose an alternative approach to facial expression editing, which takes the advantage of GAN in synthesizing realistic images with prior domain information in facial expression. To be specific, the proposed approach

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

uses Convolutional Neural Network (CNN) based generator, while the facial landmark heatmaps and the expression labels are provided guidance to the generator network in the learning process. There are several losses are used to help training, one is the squared Euclidean loss in the image space, which helps stabilize the whole training process especially in preliminary stage, and the other is perceptual loss in the image feature space, which encourages the generated faces to keep as much detail information as possible. By combining these two losses with original adversarial GAN loss, our network is able to ensure that the resulting faces show desired effects of expression while the other facial contributes are well retained.

To summarize, our main contributions are:

- We introduce an end-to-end generative network for facial expression synthesis. Both landmark heatmaps and facial expression labels of the target faces are used as a controllable signal in the generating process, which provides a flexible way for learning and inference new facial images.
- The proposed method can generate photo-realistic facial expression images under any arbitrary pose and expression, which effectively alleviates the problem of data monotonicity in facial expression database.
- Experimental results on multiple facial expression databases validate the effectiveness and efficiency of our method. The results show that the proposed method cannot only has the compelling ability to synthesize the photo-realistic facial expression images, but also could promote facial expression recognition results on these databases.

The remainder of this paper is organized as follows. We give a brief survey on relevant works in Section 2. In Section 3 we describe the model and learning strategies of proposed method. The quantitative and qualitative experimental results are shown in Section 4. Finally, we conclude the paper in Section 5.

II. RELATED WORKS

This work is related to previous research on facial expression synthesis and Generative Adversarial Network.

A. FACIAL EXPRESSION SYNTHESIS

In general, facial expression synthesis can be roughly divided into the traditional based and the deep learning based two kinds of approaches. Traditional approaches include flow based, 2D expression mapping based and 3D-based methods. Specifically, flow based methods use an 2D expression flow map to transfer facial expression [7]. The 2D flow map is generated from a 3D flow projection, which is computed from the pair of aligned 3D face shapes. Similarly, 2D expression mapping methods [16] try to transfer expressions between different facial expression images, which computes the feature difference vectors and uses them for expression interpolation. For instance, Liu *et al.* [5] present an expression ratio image (ERI) technique which captures the illumination and the geometric changes of one person's expression to enhance the facial expression mapping. 3D-based methods focus on estimate 3D shape from photographs and synthesizes facial expression images between different 3D faces. Blanz *et al.* [17] propose a photo-realistic animation system which estimates 3D shape, pose and the other rendering parameters from single images in video. However, the traditional based methods can only generate facial expression images between given images instead of unseen ones. Moreover, this kind of methods have very complex processes and often requires plenty of computation resources. Recently, deep learning-based methods have been proposed, which can be divide into deep belief network, variational autoencoders, and generative adversarial networks three subclasses. In the first subclass, Susskind *et al.* [9] study a deep belief network to generate facial expressions for a given identity and facial action unit labels. Reed *et al.* [18] present a higher-order Boltzmann machine which incorporates multiplicative interactions to model the distinct factors of variation such as facial expressions and identity. In the second subclass, Yeh *et al.* [19] present a fully automatic approach to edit faces with the variational auto-encoders, which learns to encode the facial appearance flow. Kaneko *et al.* [20] propose a conditional attribute controller for facial expression manipulation, which uses a filter module to control the facial attributes. In the third subclass, Ding *et al.* [11] propose a expression generative adversarial network, which controls the type of the expression and its intensity at the same time. Song *et al.* [21] present a geometry guided generative adversarial network, which combines facial geometry information to guide the facial expression synthesis.

B. GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial networks have exhibited a remarkable capability in image generation. The core idea is to train paired discriminator and generator networks in a minmax two-player game, where the goal of the discriminator is to classify whether the input images are from the real source or from the generator, while the goal of the generator is to generate "fake" images which data distribution is very close to the "real" images. To be specific, there are several types of GAN. The traditional GAN [22] gets a random noise as input and then outputs an photo-realistic, and Mirza and Osindero [23] propose an extension of the GAN which allows to control the generated image through some extra information such as class labels. Zhu *et al.* [24] propose a cycle GAN for learning to translate an image from a source domain X to a target domain Y, which does not need the pairs data and eases the GAN model training. On the other hand, Larsen *et al.* [25] combines variational auto-encode [26] with GAN, which adopts the discriminator of GAN as the latent representation regularizer and alleviate the blurry results. Our approach also adopts an encode-decode structure, but there are two main differences: First, a facial landmark transfer module and expression transfer network is ingenious designed, so a face with different facial landmarks and expressions can be well synthesized.

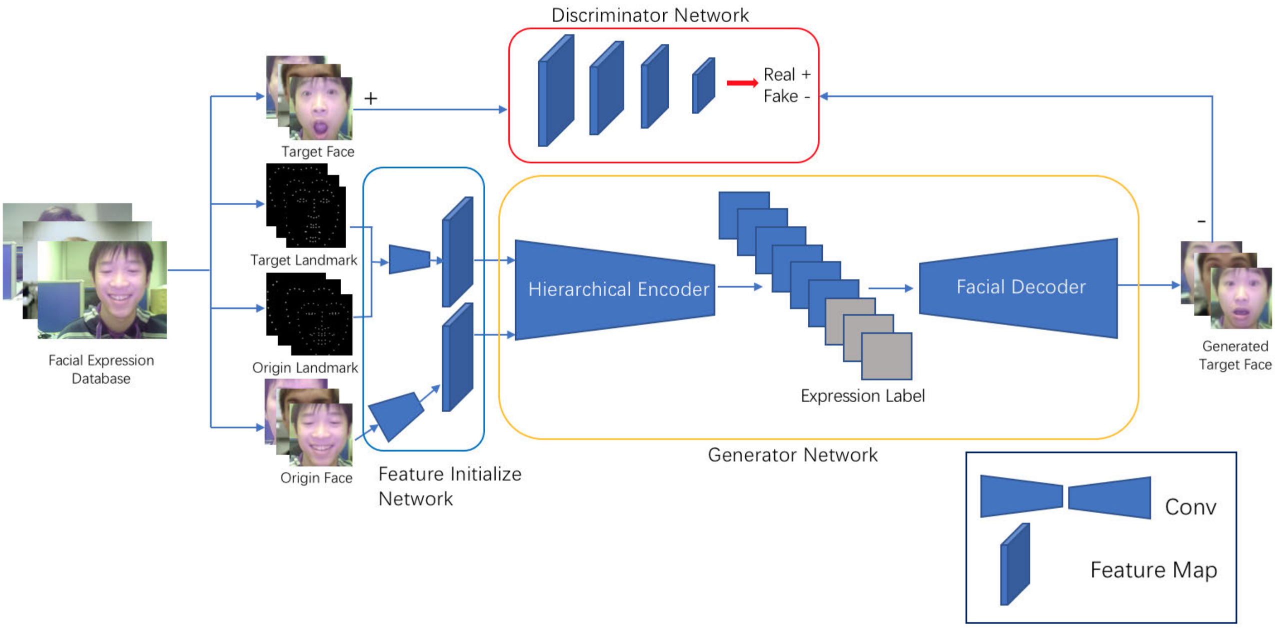

FIGURE 1. Overview of the proposed method, which includes three components: facial image and geometry feature initialize network, image encoder-decoder generator network, and image discriminator network. Specially, conv is short for the convolutional operation and the figure is best viewed in color on screen.

Second, several auxiliary losses are well incorporated with adversarial loss to help generate photo-realistic images.

III. APPROACH

The proposed model is illustrated in Fig. 1. Our method consists of three components, the first component is facial image and geometry feature initialize network, which transfers the image into a latent image space, the second component is generator network, which can be further divided into hierarchical encoder and facial decoder two modules, and the last component is discriminator network, which aims to distinguish between the target image and the generated image. Next, we will explain each component and network losses in detail.

A. SYSTEM ARCHITECTURE

As shown in Fig. 1, the input image I is first transfers into a latent image space, yielding a initialize state $E_{ini}(I)$. Then we feed both the initialized state of facial image $E_{ini}(I_{face})$ and the facial landmarks $E_{ini}(I_{geo})$ to the generator network. Specifically, the generator network includes hierarchical encoder and facial decoder two modules, hierarchical encoder is composed of several independent geometric information migration unit blocks (see in Fig. 2). Each migration unit block has two paths for information update, one for facial image features and the other for facial landmark heatmap features. Table 1 lists the architecture of these two paths in a migration unit block, we omit the last three feature map update step operations in the table for better understanding.

The main purpose of migration unit block is to transfer the target's geometric information into the input image. In order to achieve this, we first calculate a target landmark mask M_t, which represents the importance of each face location:

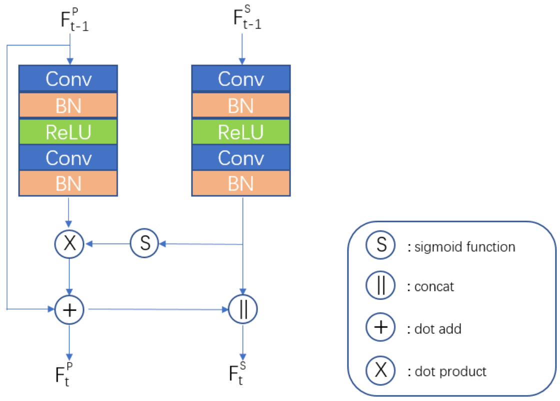

FIGURE 2. Sample of one migration unit block.

$$M_t = \sigma(conv_s(F^s_{t-1})) \tag{1}$$

where σ normalizes the input value between 0 and 1, $conv_s$ is the convolution operation, F^s_{t-1} is the facial landmark information in the prior migration unit block.

Then we multiply input facial image information and M_t together, which indicates the current position in the image features should be retained or suppressed. At the same time, the residual module(shown in Fig. 3) is designed and added, which cannot only retain the initial image coding information, but also solve the problems of gradient diffusion and gradient explosion when the network is going deep:

$$F^P_t = M_t \odot conv_p(F^P_{t-1}) + F^P_{t-1} \tag{2}$$

TABLE 1. A detailed description of the architecture of the proposed migration unit block.

Layer Type		Kernel	Stride	Pad
1_1	ReflectionPad2d1	1	-	-
1_2	Conv1	3	1	-
1_3	BatchNorm1	-	-	-
1_4	ReLU1	-	-	-
1_5	Dropout	-	-	-
1_6	ReflectionPad2d2	1	-	-
1_7	Conv2	3	1	-
1_8	BatchNorm2	-	-	-
2_1	ReflectionPad2d21	1	-	-
2_2	Conv1	3	1	-
2_3	BatchNorm1	-	-	-
2_4	ReLU1	-	-	-
2_5	Dropout	-	-	-
2_6	ReflectionPad2d2	1	-	-
2_7	Conv2	3	1	-

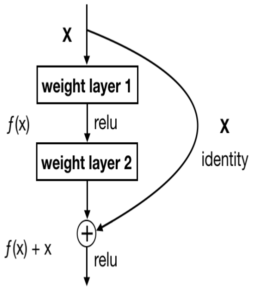

FIGURE 3. Sample of residual block module, which consists of two weight layers and one add operation.

where $conv_p$ is the convolution operation, $\odot$ is the dot product operation, F_{t-1}^P and F_t^P is the input and output of the facial image features.

After update the facial image features, the facial landmark information should be updated simultaneously. Since the facial landmark information is used to pay attention to the position of facial image features in the migration unit block. Therefore, the facial landmark information update should cover the updated facial image features:

$$F_t^S = conv_s(F_{t-1}^S) \,||\, F_t^P \tag{3}$$

where $||$ is the feature map concat operation, F_{t-1}^S and F_t^S is the input and output of the facial landmark information.

We assume that the face image is located on a high-dimensional manifold, traversing along the specific direction can achieve the smooth and transition of emotions while keeping the identity of the character unchanged. Therefore, we concat the expression label information and the facial feature maps to smooth the final generation procedure. To be specific, the expression label is first represented by one-hot encoding then in order to align the original pixel values, it is further adjusted into $[-1, 1]$, where -1 corresponds to 0 in the original image encoding. The pictorial example of transform process is visualized in Fig. 4. After expression label concatenation, the facial decoder is applied to generate the corresponding facial expression image within the given facial landmark and expression label information. Table 2 shows the architecture of the proposed decoder.

TABLE 2. A detailed description of the architecture of the proposed decoder. The output shape is described as (height, width, channels).

Layer Type		Kernel	Output	Stride	Pad
1	ConvTranspose2d1	3	64×64×128	2	1
2	BatchNorm1	-	-	-	-
3	ReLU1	-	-	-	-
4	ConvTranspose2d2	3	128×128×64	2	1
5	BatchNorm2	-	-	-	-
6	ReLU2	-	-	-	-
7	ReflectionPad2d	3	134×134×64	-	-
8	Conv	7	128×128×3	1	0
9	Tanh	-	-	-	-

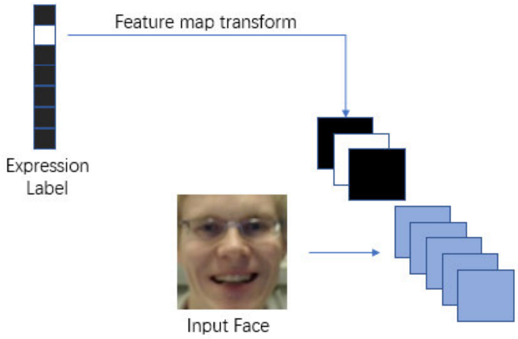

FIGURE 4. Illustration of one-hot expression label convert into feature maps.

In general, the discriminator D usually outputs a scalar value, which represents the probability value of the input image is real or generated. Inspired by the PatchGAN [27] model, ordinary identification of the whole image is not suitable for the image field requiring high-definition details and high-resolution. The discriminator in PatchGAN outputs a $N \times N$ matrix, each value can be traced back to a small certain area in the original image, in other words, the influence of the region or position on the final output can be obtained, so that the model can pay more attention to the local details of the image to a certain content, which improves the control of the detail area (see in Fig. 5).

Table 3 gives a detailed of our discriminator network structure settings. Specifically, our method adopts two discriminators, namely image discriminator and geometry discriminator. These two discriminators have the same network structure, and the main difference between the two is that the input discriminative image pairs are different. Image discriminator is used to identify whether the target image generated by reference based on the original image is true, so its input is the original image and the target image pair. While the geometric shape discriminator is used to determine whether the target image generated by the target face feature image is true or not, so its input is the target image and the target face feature image pair.

B. TRAINING LOSSES

1) ADVERSARIAL LOSS

Adversarial learning in GAN is through a minmax game between generator G and discriminator D, which gives

TABLE 3. A detailed description of the architecture of the proposed discriminator D. The output shape is described as (height, width, channels).

	Layer Type	Kernel	Output	Stride	Pad
1	Conv1	4	64×64×64	2	1
2	LeakyReLU1	-	-	-	-
3	Conv2	4	32×32×128	2	1
4	BatchNorm2	-	-	-	-
5	LeakyReLU2	-	-	-	-
6	Conv3	4	16×16×256	2	1
7	BatchNorm3	-	-	-	-
8	LeakyReLU3	-	-	-	-
9	Conv4	4	15×15×512	1	1
10	BatchNorm4	-	-	-	-
11	LeakyReLU4	-	-	-	-
12	Conv5	4	14×14×1	1	1
13	BatchNorm5	-	-	-	-
14	Sigmoid	-	-	-	-

generated images indistinguishable from real ones:

$$\min_G \max_D \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))] \quad (4)$$

where the generate images are labeled as 0 while the real images are labeled as 1.

2) RECONSTRUCTION LOSS

Since the generated images cannot be guaranteed to have fine texture details, and the whole adversarial training process is quite sensitive to all hyper-parameters. Therefore, we utilize a small weight of L1 reconstruction loss to stable the optimizer and to make sure the whole network is in a state of convergence at the beginning of training:

$$\mathcal{L}_{\text{pixel}} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\left(\widehat{P}_t\right)_{i,j} - (P_t)_{i,j}\right| \quad (5)$$

where H and W is the height and width of the image, and i and j is the position in the image space, P_t and $\widehat{P}_t$ is the real target image and the generated target image, respectively.

3) PERCEPTION LOSS

Usually, in order to make the generated images look more uniform, pixel by pixel similarity measurement method is introduced, however, this setting will loss a lot of detail information. This is because the generator has great uncertainty when reconstructing the image features in the hidden space, and most of the original detail location information is difficult to be preserved in the space after encoding. In other words, the generated image will look blur by directly expanding and reconstructing all possible positions. In order to make up the high-frequency information in the image, we use convolution neural network to constrain the similarity of the two images in the feature space:

$$\mathcal{L}_{\text{Per}} = \frac{1}{W_l \times H_l \times C_l}\sum_{i=1}^{W_l}\sum_{j=1}^{H_l}\sum_{k=1}^{C_l}\left|\varphi_l\left(\widehat{P}_t\right) - \varphi_l\left(P_t\right)\right| \quad (6)$$

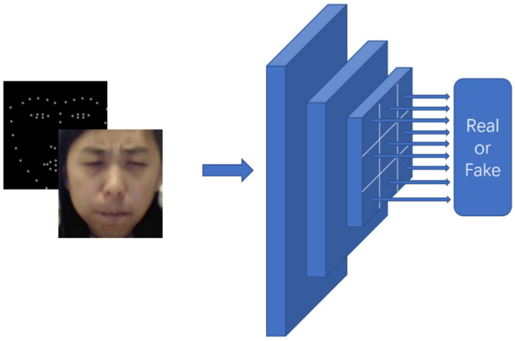

FIGURE 5. A sample of patchGAN based discriminator structure.

where φ_l is the output value of standard forward of layer l, W, H and C is the height, width and depth of feature maps.

4) OVERALL LOSS

The final training loss function is a weighted sum of all the losses, which can be defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{GAN}} + \lambda_2\mathcal{L}_{\text{Per}} + \lambda_3\mathcal{L}_{\text{pixel}} \quad (7)$$

where the hyperparameters λ_1, λ_2, and λ_3 are used to balance the three terms.

IV. EXPERIMENTS

We evaluate our method on CK+ [28] and Oulu-CASIA [29] two facial expression datasets, both qualitative and quantitative analysis are conducted to show the effectiveness of our method. Qualitative analysis focuses on the visual effect of face image generated by model, while quantitative analysis focuses on the quality evaluation of face image generated by model and the help to improve the effect of expression classification.

A. DATASETS

1) CK+

The extend Cohn-Kanade facial expression database (CK+) contains 123 different subjects and total 593 image sequences with seven basic expressions, i.e., angry, contempt, disgust, fear, happy, sad and surprise. For each expression sequence of each subject, these images record the process of one neutral expression to a specific expression and then back to neutral expression, and the last frame of each image sequence has a label for action units. In the experiment, we only use the start three and the peak three frames in the image sequences.

2) OULU-CASIA

The Oulu-CASIA NIR&VIS facial expression database (Oulu-CASIA) includes videos with the six typical expressions (happiness, sadness, surprise, anger, fear, disgust) from 80 subjects captured with two imaging systems, NIR (Near Infrared) and VIS (Visible light), under three different illumination conditions: normal indoor illumination, weak illumination (only computer display is on) and dark illumination (all lights are off).In the experiment, we only choose the

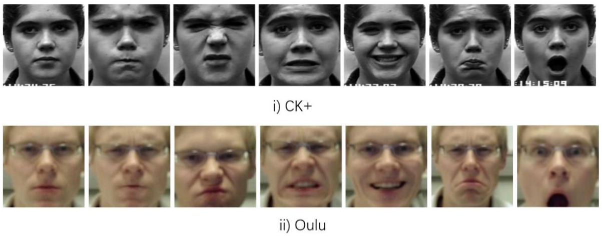

FIGURE 6. Samples of processed facial images in CK+ [28] and Oulu-CASIA [29] datasets.

image sequences with strong illumination captured by a VIS camera, and for each expression image in the sequence.

For both two datasets, we crop and detect the facial landmarks by the open-source toolkit DLib [30]. All images are rescaled to 128 × 128 and the pixel values of the input images are normalized to a range of [−1,1]. As for the training and validation sets, we randomly divide the datasets based on the subject's identity to ensure no overlap, and we keep only 10% of total data as the validation set. The preprocessed sample images in each dataset are shown in Fig. 6.

B. IMPLEMENTATION DETAILS

In the experiment, we use the Adam optimizer with learning rate of 0.0002, batchsize is set to 16, and all variables in the network are initialized with normal distribution in a value of 0.02. The hyperparameters λ_1, λ_2, and λ_3 are grid searched from the value options {1, 5, 10}, and the final decisions are set to 5, 10, 10. As for percepture loss we use the ImageNet [31] pretrained model VGG19 [32] and choose the layer $l = conv_{1\_2}$. The network was trained on an NVIDIA GeForce GTX TITAN Xp GPU in the PyTorch environment.

C. ANALYSIS

1) OULU-CASIA DATASET

In the Oulu-CASIA dataset, we conduct two aspect of analysis: one is the effect of image generation on the current data set and the effect compared with the existing methods, and the other is to increase the influence of different number of generated face images on expression recognition results.

Fig. 7 shows the comparison results of different facial expression synthesis method on Oulu-CASIA dataset. Each three lines in Fig. 7 are the results of ExprGAN [11], our method and the ground truth, we can see that our method retains many details of the initial input face, such as face identity. Furthermore, the synthesized expression looks very natural visually. In contrast, ExprGAN is unable to convert the original input face into a specific target expression face with fine details in some cases, and some of the generated faces are slightly blurred. Furthermore, we select an initial face and the corresponding initial landmark information and try to convert it to another person's landmark information, the results are shown in Fig. 8. Specifically, each three column image represents the input initial face, the target face and the final generated target face. Fig. 8 shows that even in some very challenging situations, such as the gender difference between the input face and the target face, our method still can accurately handle any switching between different expressions without being limited by the image itself. However, some results tend to ignore the glasses cases, this is mainly because of two reasons, one is the role of the L2 pixel loss is reduced to that of helping the generator preserve the face characteristics(e.g. identity, pose, expression), and the other is our training data is short of eye-glasses based samples, so there is not enough data to learn a good feature representation for this situation.

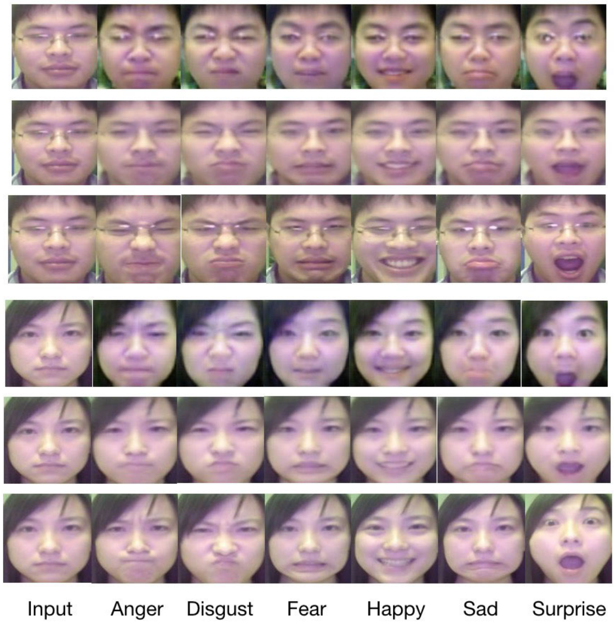

FIGURE 7. Comparison results of different facial expression synthesis method on Oulu-CASIA dataset.

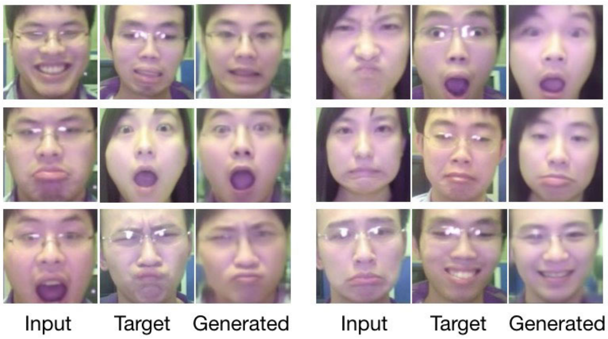

FIGURE 8. Examples of image landmark transfer.

It is popular use GAN as a data augmentation method to improve classification results. Table 4 shows the results of using our method to augment training images under the same expression classifier. That is, by given a facial input image, we can use our method to generate corresponding different poses and expressions new facial images. Table 4 shows that when the dataset are enlarge one times, the final recognition

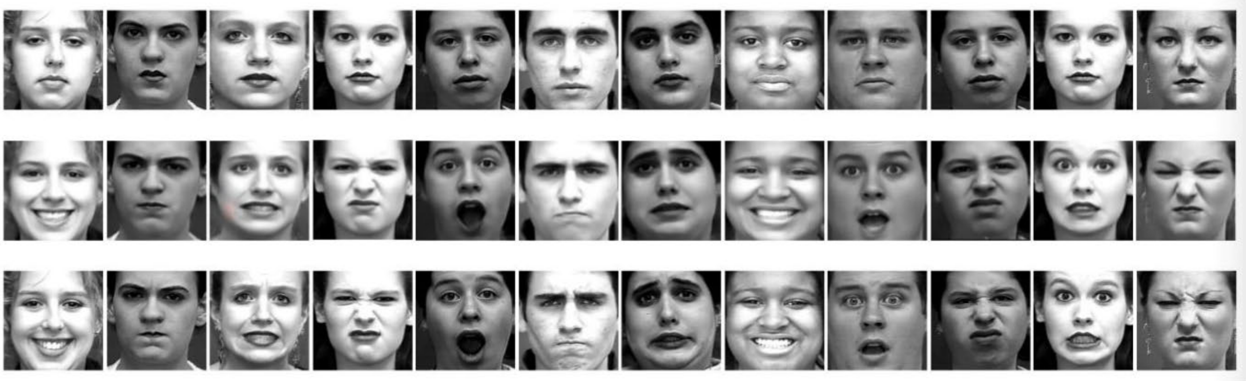

FIGURE 9. Facial expression generation results on CK+ dataset.

TABLE 4. Results on different augmentation settings for Oulu-CASIA dataset.

Times	0	1	5	10	100
Acc	67.3	68.6	71.4	76.2	76.2

results increase slightly, from 67.3% to 68.6%. Furthermore, when the data expansion ratio is increased to ten times, the recognition accuracy will be significantly improved and becomes 76.2%. And when the dataset are enlarge 100 times, the overall perceptual performance of the face model reaches its peak.

2) CK+ DATASET

On the CK+ dataset, we do the image quality evaluation, two indicators are used to quantitatively evaluate the image quality, they are peak signal-to-noise ratio (PSNR) and structure similarity (SSIM). And our method is based on GAN, several evaluations on perceptual metrics like LPIPS [36], FID [37] are also widely used in image synthesis tasks [38], [39]. Here we chooses one more indicator FID which will be helpful to give a more objective evaluation.

$$PSNR(I_1, I_2) = 10lg\frac{P^2}{\frac{1}{wh}\sum_{x,y}(I_1(x,y) - I_2(x,y))^2} \tag{8}$$

where w and h represents the height and width of the image in pixels, respectively, P is the maximum possible of gray value of image pixel(for 8-bit digital image its value is 255).

$$SSIM(I_1, I_2) = \frac{1}{wh}\sum_{x,y}\frac{(2\mu_1\mu_2 + k_1^2P^2)(2\sigma_{12} + k_2^2P^2)}{(\mu_1^2 + \mu_2^2 + k_1^2P^2)(\sigma_1^2 + \sigma_2^2 + k_2^2P^2)} \tag{9}$$

where $\mu_1(x, y)$ and $\mu_2(x, y)$ represents the mean value of two images in the sliding window, $\sigma_1(x, y)$ and $\sigma_2(x, y)$ represents the variance of two images in the sliding window, $\sigma_{12}(x, y)$ represents the covariance of two images in the sliding window, k_1, k_2 and P are used to avoid the phenomenon that the final numerical result is unstable when the denominator is close to zero.

TABLE 5. PSNR, SSIM and FID results on CK+ dataset.

Method	PSNR	SSIM	FID
GAFP-GAN [33]	20.123	0.653	73.65
GC-GAN [34]	**27.665**	0.769	68.36
CDAAE [35]	26.973	0.765	81.48
our method	23.004	**0.775**	**67.69**

$$FID(g, r) = |\mu_g - \mu_r|^2 + T_r(\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{1/2}) \tag{10}$$

where g and h represents the generative images and the real images, respectively, μ_g and μ_r is the corresponding feature vector of g and r, Σ_g and Σ_r is the factor covariance matrix of feature vector, Tr is short for Trace, represents the trace of a matrix(the sum of the elements of the main diagonal). In the real calculation case, if the open root of the matrix is a complex number, then only the real part is taken.

Fig. 9 shows the effect of the proposed method on different facial expressions, the first row is the input image, the second row is the target image and the third row is the generated results. Facial expressions are transferred between two objects within the same identity, that is, the target key point information and the initial image are from the same person's face.

Table 5 shows the image quality results of the proposed method under three different indicators. Compared with CDAAE(use discrete expression label information), GCGAN(use facial geometry information) and GAFP-GAN(use face parsing information), our method achieves the highest score in SSIM indicator and FID score, but get slightly lower PSNR score. This is because the PSNR is calculated based on the error between the corresponding pixels, which may limit to capture perceptually relevant differences such as high texture details.

To further show the robust of our proposed method, face verification experiment is conducted, gallery images are non-transformed images and probe images are transformed results, and the results can be found in Table 6. Following

TABLE 6. Quantitatively results for identity-preserving.

Method	VGG-Face			Light CNN		
	Rank1	FAR1%	FAR0.1%	Rank1	FAR1%	FAR0.1%
original	100	75.54	48.30	100	79.86	59.57
C-GAN	71.33	52.50	19.97	75.21	48.45	36.91
ExprGAN	62.31	44.69	13.27	67.49	49.31	26.40
our method	82.13	58.45	21.64	84.62	63.86	44.17

TABLE 7. Results on different module settings.

Settings	FID score
w/o feature initialize network	76.85
w/o expression label smooth	69.48
all together	**67.69**

two commonly used face verification indicators TAR(True Accept Rate) and FAR(False Accept Rate). The Rank-1 identification rate, true accept rates at 1% and 0.1% false accept rates (TAR@FAR = 1%, TAR@FAR = 0.1%) are taken as evaluation metrics. We test two released face recognition models, namely the Light CNN [40] model and the VGG-Face [41] model respectively. From the table 5 we can see that our method is good at preserving identity property in favor of C-GAN method and ExprGAN method, but still have some gap between the original results, this is mainly because in some glasses wearing case, our method did not perform well.

Table 7 gives the model of all two different module settings. Notably, our method benefits from the feature initialize module and the expression label smooth module, and by removing the feature initialize module the performance decrease largely, this mainly because manipulating on the original image is difficult and is hard for later encoder to find a proper latent space for the transformation.

V. CONCLUSION AND FUTURE WORK

In this paper, we present a facial landmarks and expression label guided photorealistic facial expression synthesis method. By corporating both the facial landmarks and expression labels from the facial image in the generator, we can generate facial images with arbitrary expressions and poses, which helps training a better expression recognition model. Experiments on Oulu-CASIA and CK+ two facial expression datasets demonstrate the effectiveness of our method.

Our future work will explore how to apply our method to a larger and more unconstrained facial expression dataset, and how to apply our model for real-time applications and achieve high recognition results.

REFERENCES

[1] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 116–134, Nov. 2007.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[3] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. 24th Annu. Conf. Comput. Graph. Interact. Techn.*, 1997, pp. 353–360.

[4] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz, "Exploring photobios," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, 2011.

[5] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, 2001, pp. 271–276.

[6] B.-J. Theobald, I. Matthews, M. Mangini, J. R. Spies, T. R. Brick, J. F. Cohn, and S. M. Boker, "Mapping and manipulating facial expression," *Lang. Speech*, vol. 52, nos. 2–3, pp. 369–386, Jun. 2009.

[7] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3D-aware face component transfer," in *Proc. ACM SIGGRAPH Papers*, 2011, pp. 1–10.

[8] T. Bolkart and S. Wuhrer, "A groupwise multilinear correspondence optimization for 3D faces," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3604–3612.

[9] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," in *Affective Computing, Emotion Modelling, Synthesis and Recognition*. 2008, pp. 421–440.

[10] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," 2014, *arXiv:1412.6583*. [Online]. Available: http://arxiv.org/abs/1412.6583

[11] H. Ding, K. Sricharan, and R. Chellappa, "ExprGAN: Facial expression editing with controllable expression intensity," 2017, *arXiv:1709.03842*. [Online]. Available: http://arxiv.org/abs/1709.03842

[12] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proc. ACM SIGGRAPH Courses*, 2005, p. 19.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[14] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2439–2448.

[15] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5541–5550.

[16] L. Williams, "Performance-driven facial animation," in *Proc. ACM SIGGRAPH Courses*, 2006, p. 16.

[17] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, 2003.

[18] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1431–1439.

[19] R. Yeh, Z. Liu, D. B Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," 2016, *arXiv:1611.09961*. [Online]. Available: http://arxiv.org/abs/1611.09961

[20] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6089–6098.

[21] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 627–635.

[22] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[25] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[26] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[28] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[29] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, ''Facial expression recognition from near-infrared videos,'' *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[30] V. Kazemi and J. Sullivan, ''One millisecond face alignment with an ensemble of regression trees,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[32] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] Z. Lu, T. Hu, L. Song, Z. Zhang, and R. He, ''Conditional expression synthesis with face parsing transformation,'' in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1083–1091.

[34] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, ''Geometry-contrastive GAN for facial expression transfer,'' 2018, *arXiv:1802.01822*. [Online]. Available: http://arxiv.org/abs/1802.01822

[35] Y. Zhou and B. E. Shi, ''Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder,'' in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 370–376.

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[37] 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,'' 2017, *arXiv:1706.08500*. [Online]. Available: http://arxiv.org/abs/1706.08500

[38] N. Wang, Y. Zhang, and L. Zhang, ''Dynamic selection network for image inpainting,'' *IEEE Trans. Image Process.*, vol. 30, pp. 1784–1798, 2021.

[39] J. Gu, Y. Shen, and B. Zhou, ''Image processing using multi-code GAN prior,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3012–3021.

[40] X. Wu, R. He, Z. Sun, and T. Tan, ''A light CNN for deep face representation with noisy labels,'' 2015, vol. 4, no. 8, *arXiv:1511.02683*. [Online]. Available: http://arxiv.org/abs/1511.02683

[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, ''Deep face recognition,'' in *Proc. BMVC*, 2015, pp. 41.1–41.12.

DEJIAN LI received the B.S. degree in computer science and technology from the Zhejiang University of Technology, Hangzhou, China, in 2015. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. His research interests include image processing, machine learning for facial expression recognition, and human–robot collaboration/human–robot interaction.

WENQIAN QI received the bachelor's degree from the School of Art and Design, Wuhan University of Technology, in 2020. She is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. Her research interests include machine/deep learning for emotional interaction, and human–robot interaction.

SHOUQIAN SUN received the B.S., M.S., and Ph.D. degrees from the School of Mechanical Engineering, Zhejiang University (ZJU), Hangzhou, China, in 1985, 1988, and 1991, respectively. From 1991 to 1993, he held a postdoctoral research position with the State Key Laboratory of CADCG, College of Computer Science and Technology, ZJU. He is currently a Research Professor with the College of Computer Science, ZJU. His research interests include big data, human–computer interface, human–robot interaction, and the ectoskeleton technology.