DualPathGAN: Facial reenacted emotion synthesis

Jiahui Kong  |  Haibin Shen  |  Kejie Huang

Abstract
Facial reenactment has developed rapidly in recent years, but few methods have been built upon reenacted face in videos. Facial-reenacted emotion synthesis can make the process of facial reenactment more practical. A facial-reenacted emotion synthesis method is proposed that includes a dual-path generative adversarial network (GAN) for emotion synthesis and a residual-mask network to impose structural restrictions to preserve the mouth shape of the source person. To train the dual-path GAN more effectively, a learning strategy based on separated discriminators is proposed. The method is trained and tested on a very challenging imbalanced dataset to evaluate the ability to deal with complex practical scenarios. Compared with general emotion synthesis methods, the proposed method can generate more realistic facial emotion synthesised images or videos with higher quality while retaining the expression contents of the original videos. The DualPathGAN achieves a Fréchet inception distance (FID) score of 9.20, which is lower than the FID score of 11.37 achieved with state-of-the-art methods.

1 | INTRODUCTION

With the coordinated development of hardware and software, deep learning has been a huge success in many fields. Among them, facial emotion synthesis has always been an attractive topic. Facial emotion synthesis can edit the emotions of the characters in images. Therefore, any target person can be decorated with any target emotion, which benefits film and television special effects production, video game production, and video conference scenes. Moreover, the related technology can be used in general entertainment applications, which provides assistance for the current development of the entertainment industry and has important engineering significance. Meanwhile, research on facial emotion synthesis also enhances the machine's ability to understand human beings in the real world through the understanding of human facial structures, human facial contents, and the semantic information of human faces.

Early facial emotion synthesis methods relied on computer graphics to fit reference objects' emotional expressions through grid models or highly parameterised 3D models. But such methods lack flexibility and have high computational complexity. They usually require separate modelling for specific characters. In recent years, because of massive growth in the amount of available data, data-driven learning methods have become prevalent. Data-driven methods such as deep learning have shown great potential in facial analysis and synthesis fields. In this context, the facial emotion synthesis algorithm based on generative adversarial networks (GANs) has made great progress in flexibility and practicability.

The main goal of our paper is to design a facial emotion synthesis method that builds upon the reenacted face images. Facial reenactment has been a research hotspot in recent years. It transfers the action captured from the face of the source to the face of the target person, thereby achieving the purpose of reenactment. At present, very few algorithms consider implementing emotion synthesis on the face that has been reenacted. Most facial emotion synthesis methods only transfer the emotion to the target face, without considering the expression contents of the original person, such as the shape of the character's mouth. From the perspective of facial reenacted emotion synthesis task, the facial contents of the person in the synthesised image or videos should maintain the similarity with the semantic contents in the original reenacted image or videos, while accurately presenting the target emotion.

Until now, the application capabilities of the existing methods in actual scenes have been limited by data conditions, and it is difficult to cope with scenes such as poor lighting conditions, covered face, and extreme head poses of the characters. Existing methods can easily process images under...
good conditions, but have poor processing capabilities for the scenes with complex conditions, and sometimes even generates incorrect face images. Therefore, it is necessary to improve the stability of the facial emotion synthesis method under complex environmental conditions and enhance the engineering application ability.

To address the aforementioned issues, we propose a dual-path GAN-based facial-reenacted emotion synthesis method with better stability and higher video quality in various scenes while retaining the semantic contents of the source person. Our method is trained and tested on a very challenging while retaining the semantic contents of the source person. Our method is trained and tested on a very challenging while retaining the semantic contents of the source person.

Our proposed method can generate more realistic images than state-of-the-art methods and can better handle difficult samples. The main contributions of this article are as follows:

1. A dual-path GAN is designed to complete the task of facial-reenacted emotion synthesis. The network has two generation paths to deal with two independent generation tasks.
2. Simple tasks are used to stabilise the learning process of complex tasks, that is, handle long learning, resist overfitting, and deal with unbalanced data by means of internal parameter influence. In our paper, the simple task is the landmark-generation task, which only focusses on several lines in one image; the complex task is the image-generation task, which contains a large number of pixels in one image.
3. To deal with the possible gradient confusion problem in gradient back-propagation of the two generative paths, we propose a 'separated discriminators' strategy to supervise the two paths separately to improve the quality of the generated results, and use visualization to ensure that the mixed conditional information (landmark) is correct.
4. We propose a residual-mask network for image-quality optimization based on the attention mechanism. The network predicts an image mask and improves the final synthesised image quality. The final results of our facial emotion synthesis pipeline achieve a lower Fréchet inception distance (FID) of 9.20, which is 19% lower than that of DFT-Net [2].

The remainder of the paper is organised as follows. Section 2 introduces the background of the paper. In Section 3, the proposed dual-path GAN is described in detail. The residual-mask network and the entire pipeline is presented in Section 4. In section 5, the challenging dataset is introduced and the experimental results of our method are provided and discussed. Finally, Section 6 briefly concludes.

2 | BACKGROUND

The method designed here is mainly based on GANs. Therefore, this section will introduce theoretical development related to GANs and the current development status of existing general facial emotion synthesis algorithms.

2.1 | Generative adversarial network

Recently, GAN is very popular in the image-generation field. [3] proposes a min-max game to train both a generator and a discriminator, forcing the generator to fool the discriminator and generate realistic results. Thereafter, to overcome the shortcoming of the original GAN, many methods have been proposed. DCGAN [4] replaces fully connected layers with convolution layers and make GAN more suitable for image-generation tasks. LSGAN [5], WGAN [6], WGAN-GP [7], and BEGAN [8] have modified the loss function to solve the unstable training and mode collapse problem. GAN is also widely used in image-to-image translation fields. The typical method, Pix2Pix [9] can handle paired data translation and have made a huge success. In addition, CycleGAN [10] handles unpaired data translation by introducing cycle-consistent loss. [11] proposes a few-shot unsupervised image-to-image translation that can learn from a small number of examples. StyleGAN [12] proposes a generator architecture that leads to an automatically learnt, unsupervised separation of high-level attributes.

2.2 | Facial emotion synthesis

Facial emotion or attributes synthesis has always been an active topic. Early methods edited the expression of the source image [13] using distortion and rendering. Some methods use highly parameterised 3-D Morphable Models and edit the parameters to obtain the face image with new emotion. Recently, GAN has shown great potential in facial emotion synthesis. [14] uses the target's landmarks as conditions and designs an expression synthesis generator and an expression removal generator to transfer emotion from any source label to another. ExprGAN [15] attempts to fuse noise with emotion labels and create a new presentation as the conditional information; thus, the generator can generate emotion with different intensities. StarGAN [16] uses unsupervised learning ideas borrowed from CycleGAN [10] and proposes an attributes mask to learning attributes synthesis from two datasets with different labels. AttGAN [17] and STGAN [18] design a network structure with skip-connection [19] to deal with the facial attributes synthesis task. Ganimation [20] uses action units as a conditional label and proposes an attention-based network. DFT-Net [2] adds another affine transformation operation to obtain better results. [21] introduces a guide mask to avoid background distortion. [22] uses continuous two-dimensional emotion labels to control automated expression editing. [23] achieved to generate images with high fidelity. [24] use pose information to guide the generation process. However, most methods ignore the existing facial content information of the characters in the original images and cannot cope with complex environmental conditions. The application capabilities of these methods in actual engineering scenarios are remain limited.
3 | DUAL-PATH GAN

To solve the issues mentioned above, a dual-path GAN is proposed for facial-reenacted emotion synthesis. This section will introduce the dual-path GAN in detail, including the dual-path generator in Section 3.1 and the separated discriminators in Section 3.2.

3.1 | Dual-path generator

The structure of the dual-path generator is shown in Figure 1a. The dual-path generator first uses a facial landmark detection algorithm to detect the landmarks of the original input face image and draws the corresponding landmark image in lines. Then the image encoder $E_m$ encodes the face image $x$ to obtain the image features. The landmark encoder $E_l$ receives the target label $c_t$ and facial landmark image $y$ and encodes the condition information to guide the generation procedure of the image path. After that, the image feature and condition information are processed by the bottleneck layer $B$ to obtain the fused feature $f$. Finally, the image decoder $DE_m$ and the landmark decoder $DE_l$ decode the target emotional face image $x$ and the corresponding landmark image $y$ respectively from the fused feature $f$. The whole process can be described by the following formula:

$$f = B(E_l(y, c_t), E_m(x)),$$

$$\tilde{x} = DE_m(f),$$

$$\tilde{y} = DE_l(f).$$

In early attempts, we found the facial landmark quite suitable for guiding facial-reenacted emotion synthesis for the following reasons:

1. The facial landmark can fully describe the structure of a person's face.
2. The facial landmark can retain the face structure of the original person while synthesising the emotion. Therefore, it can help the synthesised face image preserve the pronunciation state of the characters’ mouth.
3. Facial landmark images can converge with a relatively small dataset compared with the learning process of face images.

Unlike the previous methods, the dual-path generator attempts to adaptively encode the effective condition variables to guide the image generation process. The network visually monitors the correctness and validity of the condition information by the synthesised facial landmark image $\tilde{y}$.

The landmark-generation path ought to be embedded in the main generator. We test an independent generator to solve the emotion synthesis of facial landmark images. In the observation of test samples, the network becomes overfitted, while the results are not satisfactory. But in the dual-path generator, the landmark generation path has to keep itself from confusing the information of the image generation path. The dual-path structure brings an extra task for the landmark-generation path. Thus the generator can continuously learn and avoid overfitting.

3.2 | Separated discriminators

As shown in Figure 1b, we use two different discriminators to separately supervise the two generation paths. The image discriminator $D_m$ is used to judge the fidelity and emotion category of $\{\tilde{x}, \tilde{y}\}$ which is the combination of the generated face image $\tilde{x}$ and the generated landmark image $\tilde{y}$. The landmark discriminator $D_l$ is used to judge the fidelity and emotion category of the generated facial landmark image $\tilde{y}$. The two discriminators have the same structure except for the number of input channels of the first layer. Both discriminators use the same structure as ACGAN [25].

During training, when $D_m$ receives $\{\tilde{x}, \tilde{y}\}$, it is necessary to separate the generated landmark image $\tilde{y}$ from the gradient backward graph. The different paths will be ensured to receive the pertinent gradient update information by this operation. In this way, the network can effectively utilise the advantages of the facial landmark image, which is easy to learn and converge quickly to help the learning process of the image-generation path. This training strategy improves the stability of the generated face image quality by internal parameter constraints and external explicit supervision.

3.3 | Training strategy

In order to train on the wild dataset, the training architecture of the dual-path GAN is shown in Figure 2, which is similar to StarGAN [16]. We use the cyclic consistent loss proposed in [10] to build the training cycle.

We first feed the source face image $x$, the source facial landmark image $y$, and the target emotion label $c_t$ to the dual-path generator $G$. The generator outputs the synthesised face image $\tilde{x}$ and landmark image $\tilde{y}$. The procedure can be simply formulated as

$$\tilde{x}, \tilde{y} = G(x, y, c_t).$$

Then the dual-path generator $G$ accepts the generated face image $\tilde{x}$, landmark image $\tilde{y}$, and original emotion label $c_o$ to reconstruct the input image $x$ and $y$. This operation can be formulated as

$$\tilde{x}_{rec}, \tilde{y}_{rec} = G(\tilde{x}, \tilde{y}, c_o),$$

where $x_{rec}$ indicates the reconstructed image and $y_{rec}$ represents the reconstructed facial landmark image.
The reconstruction loss is calculated between the inputs and the reconstructed results to prevent identity loss and confirm the correctness of emotion synthesis.

### 3.4 Loss function

#### 3.4.1 Adversarial loss

We use the loss proposed by WGAN-GP [7] to train our network. The adversarial loss consists of the image discriminator $D_m$ and landmark discriminator $D_l$. As defined in Equation (4), the dual-path generator $G$ translates the input face image $x$ and corresponding facial landmark image $y$ to the synthesised face image $\hat{x}$ and landmark image $\hat{y}$ with target emotion. The adversarial loss of generator $L_{adv}^G$ and of discriminator $L_{adv}^D$ are formulated as follows:

$$L_{adv}^D = -\mathbb{E}[D_l(\hat{y})] + D_m(\{x, y\}) + D_m(\{\hat{x}, \hat{y}\}),$$

$$L_{adv}^G = -\mathbb{E}[D_l(\hat{y})] + D_m(\{\hat{x}, \hat{y}\}).$$

We refer to the terms $D_l(\hat{y})$ and $D_m(\{x, y\})$ as the probability distributions over the input data given by $D_l$ and $D_m$.

The terms $D_l(\hat{y})$ and $D_m(\{\hat{x}, \hat{y}\})$ are the probability distributions over the synthesised results.

#### 3.4.2 Classification loss

Our goal is to generate images that can be classified into the target emotion category. The structures of both discriminators have an auxiliary output layer to estimate the emotion categories. To achieve this condition, we impose classification loss to optimise the network. The classification loss of real data $L_{cls}^r$ is defined as

$$L_{cls}^r = \mathbb{E}[-\log(D_l(c_o | y)) - \log(D_m(c_o | \{x, y\}))].$$

The term $D_m(c_o | \{x, y\})$ represents a conditional probability distribution of $\{x, y\}$ over the original emotion label $c_o$ given by the image discriminator $D_m$. The term $D_l(c_o | y)$ given by the landmark discriminator $D_l$ represents the conditional distribution of $y$ over the original emotion label $c_o$. The discriminators $D_l$ and $D_m$ learn to classify the real data to its corresponding original emotion label $c_o$ by minimising this objective. We also need to calculate the classification loss of the synthesised data $L_{cls}^f$, with the former defined as

$$L_{cls}^f = \mathbb{E}[-\log(D_l(c_o | \hat{y})) - \log(D_m(c_o | \{\hat{x}, \hat{y}\}))].$$
The dual-path generator aims to minimise this objective to generate results that can be classified as the target emotion label $c_t$.

3.4.3 | Reconstruction loss

We finally apply a cyclic consistent loss [10] as a reconstruction loss to maintain the person’s identity. We use the generated face image $\tilde{x}$, the generated landmark image $\tilde{y}$, and original emotion label $c_t$ to reconstruct the source image by dual-path generator $G$. The procedure is formulated in Equations (4) and (5). Then we apply the L1 loss to evaluate the reconstructed face image $x_{rec}$ and reconstructed landmark image $y_{rec}$. The reconstruction loss $L_{rec}$ is

$$L_{rec} = E[\|x - x_{rec}\|_1 + \|y - y_{rec}\|_1]. \quad (10)$$

3.4.4 | Total loss

The total losses in our approach are defined as follows:

$$L^{D}_{dp} = L^{D}_{adv} + \lambda_{dcl}L^{cl}_{cls}, \quad (11)$$

$$L^{G}_{dp} = L^{G}_{adv} + \lambda_{gcl}L^{cl}_{cls} + \lambda_{rec}L_{rec}. \quad (12)$$

$L^{D}_{dp}$ is the loss function of the discriminators, and $L^{G}_{dp}$ is the loss function of the dual-path generator. $\lambda_{dcl}$, $\lambda_{gcl}$, and $\lambda_{rec}$ are the adjustable weights of the corresponding terms.

3.5 | Implementation detail

In our experiment, we use Adam [26] as the optimiser with a learning rate of 0.0001 for the dual-path GAN. We also use Instance normalisation [27] in generator and batch normalisation [28] in discriminator. The weights $\lambda_{dcl}$ and $\lambda_{rec}$ are set to 1 and 10 respectively. The model is trained on the training set for 300,000 iterations with a batch size of 32. Limited to the GPU resource, the input and output image size is 128 $\times$ 128. Our model is trained on two GTX1080Ti GPUs for about 2 days.

4 | RESIDUAL-MASK NETWORK

Because the results of the dual-path GAN still have excessive image noise and severe colour shift, a residual-mask network is proposed to further optimise the quality of the results generated by the dual-path generator.

4.1 | Network architecture

We design a residual-mask network $R$ based on the attention mechanism to further improve the quality of the synthesised face images. The network’s basic architecture is shown in Figure 3. The input of the network is a residual image $\Delta x$ obtained by subtracting the synthesised face image $\tilde{x}$ generated by the dual-path generator from the original image $x$. The residual-mask network $R$ infers the image mask $m$ from the residual information. This process can be formulated as

$$m = R(x - \tilde{x}). \quad (13)$$

Then the final synthesised face image $x_{final}$ is calculated by

$$x_{final} = mx + (1 - m)\tilde{x}. \quad (14)$$

Embedding the attention mechanism into the dual-path generator could generally improve the quality of the image. However, the bottleneck layer of the dual-path generator needs to mix the information of different tasks when processing multiple tasks, and the landmark may not contribute to the prediction of the image mask. These phenomena occur because the dual-path generator itself must already process two information channels that have strong relevance and mutual restriction characteristics. The image is relevant to the facial landmark, but there is no visual relevance between facial landmarks and masks. Therefore, the landmark information may not contribute to mask prediction. Instead, it will bring a greater process burden.

Because embedding the mechanism into the dual-path generator does not give the most desirable results, we take a different perspective and design an easy-to-use individual model for applying attention mechanism. Using a separate model to apply the attentional mechanism can effectively improve image quality and avoid putting too much burden on the network. Furthermore, a simple, easy-to-train model can be quickly and easily applied to other scenarios.

We assume that the difference $\Delta x$ between the generated image with target emotion $\tilde{x}$ and original image $x$ is already known, and then this difference is extracted for processing and converted into the required image mask $m$ while suppressing the noise information. With the help of the image mask $m$, the unchanged pixels of the original image $x$ are retained to improve the quality of the final generated image $x_{final}$.

4.2 | Training strategy

The architecture during training is also shown in Figure 3. We train this network by the typical GAN training strategy. During training, the residual-mask network $R$ tries to minimise the distance between the distribution of the final synthesised image $x_{final}$ while the discriminator attempts to distinguish the generated images. In our observation, the network will converge after about 2000 iterations on the training set.
4.3 | Loss function

4.3.1 | Adversarial loss

In this chapter, the adversarial loss of residual-mask network $L_{\text{adv-r}}^R$ and the one of the discriminator $L_{\text{adv-r}}^D$ are defined as

\[
L_{\text{adv-r}}^R = -\mathbb{E}[D(x_{\text{final}})],
\]
\[
L_{\text{adv-r}}^D = -\mathbb{E}[D(x)] + \mathbb{E}[D(x_{\text{final}})].
\]

4.3.2 | Content loss

We use the content loss [29] to keep the content consistency between the final calculated $x_{\text{final}}$ and the image $\tilde{x}$ generated by dual-path generator. The loss function $L_{\text{cnt}}$ is formulated as follows:

\[
L_{\text{cnt}} = (\text{vgg}^{19}_{\text{relu}_4}(x_{\text{final}}) - \text{vgg}^{19}_{\text{relu}_4}(\tilde{x}))^2.
\]

\text{vgg}^{19}_{\text{relu}_4}(\cdot) represents the feature vector of the relu4_1 layer extracted by a pre-trained VGG19 net.

4.3.3 | Regularisation loss

The image mask is output from a sigmoid layer. To avoid saturation of the final image mask $m$, a regularisation loss is added to maximise the value of the image mask $m$. The regularisation loss $L_{\text{reg}}$ is formulated as

\[
L_{\text{reg}} = -\mathbb{E}[|m|].
\]

4.3.4 | Smoothing loss

To obtain an image mask with sufficient smoothness, the total variation loss [30] is used to present the smoothing loss. The form of the smoothing loss $L_{\text{tv}}$ is

\[
L_{\text{tv}} = \sum (|\tilde{x}_{i,j} - \tilde{x}_{i,j-1}| + |\tilde{x}_{i,j} - \tilde{x}_{i+1,j}|).
\]

where $\tilde{x}_{i,j}$ indicates the pixel value at coordinates $(i, j)$.

4.3.5 | Total loss

We refer to the term $L_{\text{rm}}^D$ as the loss function of the discriminator $D$ and the term $L_{\text{rm}}^R$ as the loss function of the residual-mask network $R$. These two terms are as follows:

\[
L_{\text{rm}}^D = L_{\text{adv-r}}^D,
\]
\[
L_{\text{rm}}^R = L_{\text{adv-r}}^R + \lambda_{\text{cnt}} L_{\text{cnt}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{tv}} L_{\text{tv}}.
\]

$\lambda_{\text{cnt}}, \lambda_{\text{reg}},$ and $\lambda_{\text{tv}}$ are adjustable weights of the corresponding loss terms.

4.4 | Implementation detail

During training, the learning rate of the network is set to 0.0001. Each time the discriminator and the generator are trained. The network is trained for 10,000 iterations. However, from the observation results of the test data, the network converges around 2000 iterations. In the loss function, $\lambda_{\text{cnt}} = 10, \lambda_{\text{reg}} = 1, \text{ and } \lambda_{\text{tv}} = 0.00001, \lambda_{\text{reg}}$ and $\lambda_{\text{tv}}$ are much more sensitive than $\lambda_{\text{cnt}}$ in our model. These two parameters control the final quality of our generated images. If $\lambda_{\text{reg}}$ and $\lambda_{\text{tv}}$ are set to a large number, we may obtain a blank mask while a small one does not help improve the image quality at all. $\lambda_{\text{reg}}$ and $\lambda_{\text{tv}}$ can affect each other, and if $\lambda_{\text{reg}}$ is too large, $\lambda_{\text{tv}}$ will fail, and vice versa. We must choose suitable values for them through experiments. The model was trained on one GTX1080Ti for about 2.5 h.

4.5 | Emotion synthesis pipeline

The entire pipeline of our facial reenacted emotion synthesis method is shown in Figure 4. Firstly, an input image $x$ is feed into the dual-path generator. The generator outputs a synthesised image with the target emotion $\tilde{x}$. Secondly, the residual-mask network predicts an image mask $m$ from a residual image $\Delta x$ of the synthesised image $\tilde{x}$ and the input $x$. Finally, the final generated image $x_{\text{final}}$ is calculated by the image mask $m$, the synthesised image $\tilde{x}$, and the input image $x$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{The residual-mask network structure and training architecture. The training strategy is the same as that of the original GAN. A discriminator is used to determine the fidelity of the result $x_{\text{final}}$. GAN, generative adversarial network.}
\end{figure}
The test pipeline of facial emotion synthesis. The input will be processed by the dual-path generator and residual-mask network in order, and the final synthesised result will be obtained by computing with the original image.

Visual comparison of our method with Ganimation [20], StarGAN [16] and DFT-Net [2]. The results of Ganimation have more unnatural colour patches; StarGAN has more artificial traces and less sharpness; DFT-Net has more unnatural distortions; the facial emotional image synthesised by ours has a less exaggerated emotion expression, and fewer artificial traces on the human face, which makes the image more visually natural. Please zoom in for more details. GAN, generative adversarial network.

## EXPERIMENT

In this section, we will introduce the challenging database, and compare our method with some state-of-the-art methods that we train on the wild dataset. For a fair comparison, we retrain all methods on the same training set. In addition, we evaluate the results generated from a wild video frame database. Some methods may be modified to fit specific tasks.

### 5.1 Database

#### 5.1.1 EmoVoxCeleb

EmoVoxCeleb [1] database is used in the simulation because we want to build a method that is powerful enough to generalise on wild data and handle complex situations. This database is built upon the VoxCeleb1 [31] database which consists of 153,352 video clips of 1,252 celebrities cut from open source videos.

The challenge of this database is that the images have different resolutions, and the background are very complex—for example, low light conditions, extreme head poses, covered faces etc. The face emotion synthesis methods only focussing on the frontal face with good-quality images are not competent to handle complicated situations mentioned above.

Moreover, the emotion label is extraordinarily unbalanced. In the training, we randomly select 106,765 video frames. Our training set contains around 20,000 images of emotion categories ‘neutral’, ‘happy’, ‘surprised’, ‘sad’, and ‘angry’. But we are limited by the number of ‘disgusted’ and ‘fearful’ images and have only 3920 and 3608 images for these two classes, respectively.

#### 5.1.2 VoxCeleb2

The VoxCeleb2 [32] database, an extended version of VoxCeleb1, is also used to test our model on both single images and consecutive video frames. VoxCeleb2 offers 1,092,009 video clips, which is a much bigger database than VoxCeleb1. From this database, 30 video clips of different people are randomly selected and decomposed into 4615 frames to form a test set.
5.2 | Visual comparison

5.2.1 | Normal sample

The visual comparison of the synthesised results on the normal samples is shown in Figure 5. The normal samples refer to those that have uncovered faces and nearly frontal head poses. Compared with other methods, our results are more realistic because of fewer artificial traces on the face visually.

It can be observed from the generated mouth area in Figure 5 that our results maintain the mouth shape of the source person. In contrast, the other methods destroy the person's pronunciation state in some emotion categories. Keeping the original facial contents leads to less exaggerated emotional expression. In facial reenacted emotion synthesis, the original person's pronunciation state is desired to be preserved, so it is acceptable to sacrifice some emotional intensity.

Comparing the results of different methods on the unbalanced data categories of ‘disgusted’ and ‘fearful’, our method can synthesise realistic face images, whereas other methods may not synthesise normal face images and may show significant image quality degradation. This indicates that our method can better deal with the training process on unbalanced data.

Figure 5 also shows the synthesised facial landmark images. It can be observed that the synthesised facial landmark image can accurately describe the facial structure in the target emotion category, and provides effective generation guidance for the face image-generation process.

5.2.2 | Hard sample

Figure 6 illustrates the synthesised results of the baseline methods and ours on hard samples. The hard samples in Figure 6 include two types, those with covered faces and those with extreme head poses. These two types are very common in wild data, and the previous methods cannot deal well with samples in such difficult scenes. In Figure 6, other methods are nearly ineffective in processing hard samples and cannot generate normal human faces, whereas our method can still output results with stable image quality.

![Figure 6](image_url) Comparison of synthesised results of different methods on hard samples. Ganimation performs poorly in scenes with covered faces and extreme head poses. Ganimation is almost impossible to synthesise normal faces; StarGAN synthesises the wrong facial contours and has more unnatural lines; DFT-Net has more serious distortion problems and some wrong artefacts; our method can keep the facial structure of the character clear, and generate more realistic results. Please zoom in for details. GAN, generative adversarial network.
5.3 | Ablation study

5.3.1 | Module ablation

Figure 7 shows the ablation study of our model. We created three other models besides ours, including a model that does not use the landmark path, a model without separated discriminators, and a model without a residual-mask network. We find that without the guidance of the landmark, the model fails to synthesise emotion while training on a large amount of data. The model also bears a certain degree of overfitting, and is unable to process the test image correctly. The model without separated discriminators performs very well in the effect of emotion synthesis, but there

![Image 1](image1.png)

**Figure 7** Ablation study of our proposed model. Without the landmark path, the method fails to synthesise emotions in most situations and has more artefacts. The method with only one discriminator has more strange lines and suffers from a colour shift problem. The difference between the result of single dual-path generator and the result of our whole method is not obvious. The quantitative comparison results are summarised in Table 1.

| Methods                          | PSNR | SSIM | FID   |
|----------------------------------|------|------|-------|
| W/O landmark path                | 26.75| 0.847| 26.01 |
| W/O separated discriminators     | 27.31| 0.894| 26.07 |
| W/O residual-mask network        | 27.64| 0.917| 16.87 |
| Ours                             | 32.56| 0.953| 9.20  |

Note: The bold values indicate that the value/score is the best when compared with other methods.

Abbreviations: FID, Fréchet inception distance; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index measure.

![Image 2](image2.png)

**Figure 8** The model trained on a small data with 3000 images per categories generate images with bad quality. Increasing the amount of data improves image quality but did not help that much. The model uses dropout but still has poor generalise ability. Our model with the landmark path deals well with both image quality and emotion categories.
are many artificial traces in the generated image that degrade the image quality, and in some cases, it fails to synthesise normal faces. Using the pure dual-path generator, which means without the residual-mask network, achieves a good balance between emotion synthesis and image quality but still has some colour shift and noise problems. The comparison results of image quality are tabulated in Table 1. The proposed two-path model generates the images with the best quality.

5.3.2 About landmark path

In our method, the Landmark path has multiple auxiliary functions. Through experiments, we have found that the landmark path can alleviate overfitting to a certain extent. We use a network without a landmark path trained on small and large amounts of data, to test the relationship between overfitting and training data size. We also train a network that uses dropout to alleviate overfitting. Figure 8 shows the comparison results between our proposed model and the above three models. It can be seen that increasing the amount of data does not effectively alleviate overfitting, but leads to very poor image quality on the test data. Dropout is not helpful either. Our method with a landmark path has better generalisability of test data.

In addition, our training data bear the problem of unbalanced categories. As illustrated in Figure 9, our proposed model can better handle categories with a small number of training data in the unbalanced dataset and generate more realistic images. The results generated by other methods have more artefacts and cannot generate correct face contours.

5.4 Talking state retention comparison

The primary purpose of our method is to make this network available in video editing after facial reenactment, which means it can retain transferred facial semantic information, such as the talking state of a person. The continuous pronunciation of characters is closely related to changes in mouth structure. We cut out the mouth of a person in a video and calculate the change in the structure of the mouth between two consecutive videos. We use the structural similarity index measure (SSIM) to represent this value and

![Table 2: Image quality evaluation comparison](image)

| Methods      | Ganimation | StarGAN | DFT-Net | Ours       |
|--------------|------------|---------|---------|------------|
| PSNR         | 28.38      | 28.73   | 30.22   | 32.56      |
| SSIM         | 0.916      | 0.909   | 0.934   | 0.953      |
| FID          | 16.51      | 23.17   | 11.37   | 9.20       |

Abbreviations: FID, Fréchet inception distance; GAN, generative adversarial network; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index measure.

![Figure 10: The comparison of mouth structure similarity in consecutive frames](image)

The comparison of mouth structure similarity in consecutive frames. Our method has the same trends as the source video frames and also achieves the closest SSIM value, which indicates that our results have mouth change similar to that of the source person and maintain the original speech state. SSIM, structural similarity index measure.
has better coherence. The real video frame, with similar change trends and values. We can observe that our curve is the closest representation of the original video and the generated videos. The SSIM between consecutive frames is only computed around the mouth area to ensure that this experiment focuses only on the talking state. We can observe that our curve is the closest representation of the real video frame, with similar change trends and values. We also achieve a lower fluctuation, which means that our result has better coherence.

In Figures 5 and 6, we have already shown the ability of our method to synthesize emotion and good quality images. In Figure 10, we randomly select one video from the test data and compare the mouth change curve of the first 50 frames in the original video and the generated videos. The SSIM between consecutive frames is only computed around the mouth area to ensure that this experiment focuses only on the talking state. We can observe that our curve is the closest representation of the real video frame, with similar change trends and values. We also achieve a lower fluctuation, which means that our result has better coherence.

### 5.5 Generation quality evaluation

#### 5.5.1 Image quality

Our method can synthesize images with higher quality and stability. To evaluate the quality of the generated images, we use general evaluation indicators PSNR, SSIM [33], and FID [34] to compare the quality of the images generated by different methods. The images with better quality will have smaller FID, higher PSNR, and higher SSIM. The calculation results of the indicators are listed in Table 2. Among the four methods, our method achieves the highest PSNR and SSIM. We also achieve the lowest FID of 9.20 among the four methods. In a word, our method generates images with the best quality.

### 5.6 Amazon Mechanical Turk perceptual evaluation

#### 5.6.1 Image evaluation

The human eye can better evaluate face images' visual quality. This experiment uses Amazon Mechanical Turk (AMT) to manually evaluate and compare the generated face quality of different methods. Firstly, the synthesized images with seven emotion categories are generated by ours, StarGAN, Ganimation, and DFT-Net. We then randomly select images from the generated data for evaluation. The four different results from the same single source image are concatenated side by side to make up the picture for displaying on the questionnaire page. According to the AMT website's rules, we can only upload limited data. We successfully upload 224 images for the AMT evaluation task. The volunteers are asked to choose an image with the best quality from among the four displayed. To obtain more feedback, each sample is evaluated by three randomly selected volunteers. The results are summarized in Table 4. The sum of the four statistical results is 100%. In Table 4, it can be concluded that the synthesized images using our method have a higher proportion of good quality.
quality faces, and are more realistic and natural than other methods.

5.6.2 | Video evaluation

We use various methods to process the consecutive video frames and combine the synthesised frames into video files. The visual coherence and naturalness of the videos are evaluated by AMT. In this experiment, we use all 30 video clips in the test set for evaluation. We process each frame in videos to obtain synthesised video clips with seven emotion categories. Because 1 of the 30 test files is broken, we finally upload 198 synthesised video files for comparison. The specific layout of the questionnaire is the same as in the image AMT evaluation task. Each comparison questionnaire invited three randomly selected volunteers to do the evaluation task. The feedback results are summarised in Table 5. The sum of the four statistical results is 100%. Because fewer artificial traces can change between frames when processing consecutive frames, our synthesised videos are more coherent and natural visually. Table 5 indicates that our method can generate more stable face images and has a better ability to process video data.

6 | CONCLUSION

We have proposed a method for facial-reenacted emotion synthesis, including a dual-path generated adversarial network and a residual-mask network. Compared with existing facial emotion synthesis methods, our method is able to synthesise face images with higher quality while maintaining the pronunciation state of the original person and cope with more complicated scenes. Our method also shows an excellent ability to process video data. In future work, there is the potential to introduce time-series information to achieve facial emotion synthesis of video data.

ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China under Grant 2018Y FB0904902.

ORCID

Jiahui Kong https://orcid.org/0000-0002-5171-7154
Kejie Huang https://orcid.org/0000-0003-3722-9979

REFERENCES

1. Albanie, S., et al.: Emotion recognition in speech using cross-modal transfer in the wild. In: Proceedings of the 26th ACM international conference on Multimedia (MM ’18), pp. 292–301. (2018). https://doi.org/10.1145/3240506.3240578
2. Wang, J., et al.: Df-net: disentanglement of face deformation and texture synthesis for expression editing. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3881–3885. IEEE (2019). https://doi.org/10.1109/ICIP2019.8803416
3. Goodfellow, I., et al.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pp. 2672–2680. (2014). https://dl.acm.org/doi/10.5555/2969033.2969125
4. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR). preprint arXiv:1511.06434 (2016). https://arxiv.org/abs/1511.06434v2
5. Mao, X., et al.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2794–2802. (2017). https://doi.org/10.1109/ICCV.2017.304
6. Martin, A., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML’17), pp. 214–223. (2017). https://dl.acm.org/doi/abs/10.5555/3305381.3305404
7. Gulrajani, I., et al.: Improved training of wasserstein gans. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17), pp. 5769–5779. (2017). https://dl.acm.org/doi/10.5555/3292222.3295327
8. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary Equilibrium Generative Adversarial Networks. arXiv.org e-Print Archive. preprint arXiv:1703.10717 (2017). https://arxiv.org/abs/1703.10717
9. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. (2017). https://doi.org/10.1109/CVPR.2017.632
10. Zhu, J.-Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2242-2251. (2017). https://doi.org/10.1109/ICCV.2017.244
11. Liu, M., et al.: Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10550-10559. (2019). https://doi.org/10.1109/ICCV.2019.01065
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396-4405. (2019). https://doi.org/10.1109/CVPR.2019.00453
13. Liu, Z., Shan, Y., Zhang, Z.: Expressive expression mapping with ratio images. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 271–276. ACM (2001). https://doi.org/10.1145/383259.383289
14. Song, L., et al.: Geometry guided adversarial facial expression synthesis. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 627–635. ACM (2018). https://doi.org/10.1145/3240508.3240612
15. Ding, H., Kumar, S., Chellappa, R.: ExpGAN: facial expression editing with controllable expression intensity. In: Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence (2018), oral. https://arxiv.org/abs/1709.03842
16. Choi, Y., et al.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8789–8797. (2018). https://doi.org/10.1109/CVPR.2018.00916
17. He, Z., et al.: ArtGAN: Facial Attribute Editing by Only Changing What You Want. IEEE Transactions on Image Processing, 28(11), 5464–5478. (2019). http://dx.doi.org/10.1109/tip.2019.2916751
18. Liu, M., et al.: Stgan: a unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3668-3677. (2019). https://doi.org/10.1109/CVPR.2019.00379
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28
20. Albert, P., et al.: GANimation: One-Shot Anatomically Consistent Facial Animation. Int. J. Comput. Vis. 128(3), 698–713. (2020). http://dx.doi.org/10.1007/s11263-019-01210-3
21. Yang, H.S., Moon, Y.S.: Face attribute editing using attgan and guide mask. In: Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), and Communication (ICEIC), pp. 1–3. IEEE (2019). https://doi.org/10.23919/ELINFO.COM.2019.8706471
22. Lindt, A., et al.: Facial expression editing with continuous emotion labels. In: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1–8. (2019). https://doi.org/10.1109/FG.2019.8756558
23. Fu, C., et al.: High-Fidelity Face Manipulation With Extreme Poses and Expressions. In: IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2218–2231. (2021). http://dx.doi.org/10.1109/TIFS.2021.3050065
24. Hu, Y., et al.: Pose-guided photorealistic face rotation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8398–8406. (2018). https://doi.org/10.1109/CVPR.2018.00876
25. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, vol. 70, pp. 2642–2651. JMLR. org (2017). https://dl.acm.org/doi/10.5555/3305890.3305954
26. Kingma, D.P., Jimmy, B.: Adam: a method for stochastic optimization. In: The 3rd International Conference for Learning Representations, San Diego. (2015). http://arxiv.org/abs/1412.6980
27. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalisation: The Missing Ingredient for Fast Stylisation. arXiv.org e-Print Archive. preprint arXiv:1607.08022 (2017). https://arxiv.org/abs/1607.08022
28. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, pp. 448–456. (2015). https://dl.acm.org/doi/10.5555/3045118.3045167
29. Gatys, L.A., Ecker, A.S., Bethge, M.A.: Neural Algorithm of Artistic Style. J. Vis. (2015), 16(12), 326(2016). http://dx.doi.org/10.1167/16.12.326
30. Rudin, L.L., Stanley, O., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. Nonlinear Phenom. 60(1-4), 259–268. (1992). http://dx.doi.org/10.1016/0167-2789(92)90242-f
31. Nagrani, A., et al.: Voxceleb: Large-scale speaker verification in the wild. Comput. Speech Lang 60, 101027 (2020). http://dx.doi.org/10.1016/j.csl.2019.101027
32. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep Speaker Recognition. In: Proc. Interspeech 2018. preprint arXiv:1806.05622 pp. 1086–1090. (2018). http://dx.doi.org/10.21437/Interspeech.2018-1929
33. Borji, A.: Pros and cons of GAN evaluation measures. Comput. Vis. Image Understand. 179, 41–65. (2019). http://dx.doi.org/10.1016/j.cviu.2018.10.009
34. Martin, H., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17), pp. 6629–6640. (2017). https://dl.acm.org/doi/10.5555/3295222.3295408

How to cite this article: Kong J, Shen H, Huang K. DualPathGAN: Facial reenacted emotion synthesis. IET Comput. Vis. 2021;1–13. https://doi.org/10.1049/cvi2.12047